Investigating the Influence of Changes in Emotional Expressions on Identity Recognition of Unfamiliar Faces and Voices

Hanjian Xu

Integrated Program in Neuroscience
McGill University, Montreal, Canada
November 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract ..... *i*

Résumé ..... *iii*

Acknowledgments ..... *v*

Contributions of original knowledge ..... *vi*

Contributions of authors ..... *vii*

List of abbreviations ..... *viii*

List of figures ..... *ix*

List of tables ..... *x*

Chapter 1. General Introduction *1*

1.1 How do we recognize faces? **2**

1.1.1 Face processing models 2

1.1.1.1 Functional face processing model 2

1.1.1.2 Neurological face processing model 5

1.1.2 Familiar and unfamiliar faces 5

1.1.3 Mental representations of face 6

1.2 How do we recognize voices? **7**

1.2.1 Voice processing model 7

1.2.2 Mental representation of voice 9

1.3 Flexibility of faces and voices **9**

1.3.1 Variance in faces 10

1.3.2 Variance in voices 11

1.4 Emotional expression - common variance in face and voice **12**

1.4.1 Recognition of emotional faces and voices 13

1.4.2 Identity recognition with varied emotional expressions 13

1.4.3 Emotion-specific influence on face memory 14

1.4.4 What's special about emotional expression? 15

1.5 Overview of the current work **16**

Chapter 2. Influence of emotional prosody, content and repetition on memory recognition of speaker identity *18*

2.1 Abstract **19**

2.2 Introduction **20**

2.3 Experiment 1 ..... **23**

2.3.1 Methods 24

2.3.1.1 Participants 24

2.3.1.2 Stimuli 24

2.3.1.3 Speaker Selection 25

2.3.1.4 Acoustic Features Analysis 25

2.3.1.5 Procedure 27

2.3.1.6 Data Analysis ... 28
2.3.2 Results ... 29
2.3.2.1 Encoding ... 29
2.3.2.2 Recognition ... 30
2.3.3 Discussion ... 33
2.4 Experiment 2 ... 36
2.4.1 Methods ... 36
2.4.1.1 Participants ... 36
2.4.1.2 Stimuli ... 36
2.4.1.3 Procedure ... 37
2.4.1.4 Data Analysis ... 37
2.4.2 Results ... 38
2.4.2.1 Encoding ... 38
2.4.2.2 Recognition ... 39
2.4.3 Discussion ... 40
2.5 General Discussion ... 42
2.6 Limitations ... 45
2.7 Conclusion ... 46
2.8 Acknowledgements ... 46
2.9 Conflict of interest ... 47
2.10 References ... 48
2.11 Supplementary Materials ... 58
Connecting Chapters 2 to 3 ... 60
Chapter 3. Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition ... 61
3.1 Abstract ... 62
3.2 Introduction ... 63
3.3 General Methods ... 66
3.3.1 Participants ... 66
3.3.2 Stimuli ... 67
3.3.3 Procedure ... 68
3.3.4 Dependent Measures ... 69
3.3.5 Data Analysis ... 70
3.3.5.1 Recognition Accuracy ... 70
3.3.5.2 Drift Diffusion Models (DDMs) ... 70
3.3.5.3 Stimulus-based physical feature analysis ... 71
3.4 Main Study ... 72
3.4.1 Methods ... 72
3.4.1.1 Participants ... 72
3.4.1.2 Procedure ... 72
3.4.1.3 Data Analysis ... 73
3.4.2 Results ... 75
3.4.2.1 Encoding: Implicit Memory (Priming) ... 75
3.4.2.2 Recognition: Identity Memory ... 75
3.4.3 Summary ... 81

3.5 Follow-up Study 1: High vs. low arousal emotions 82
3.5.1 I: Fearful vs. Sad 82
3.5.1.1 Methods 82
3.5.1.2 Results 82
3.5.2 II: Fearful vs. Neutral 83
3.5.2.1 Methods 83
3.5.2.2 Results 83
3.6 Follow-up Study 2: High-arousal *Multi* vs. *Uni* 84
3.6.1 Methods 85
3.6.2 Results 85
3.7 General Discussion 86
3.7.1 Implicit and explicit identity memory 86
3.7.2 High arousal interferes with emotion-independent identity memory 87
3.7.3 Multiple exemplar memory advantage compared to repeated high-arousal expressions 89
3.8 Limitations and Future Directions 90
3.9 Conclusion 91
3.10 References 93
3.11 Supplementary Information 103
Connecting Chapters 3 to 4 112

Chapter 4. Cross-emotional-expression recognition in unfamiliar faces and voices – an fMRI study 113
4.1 Abstract 114
4.2 Introduction 115
4.3 Methods 118
4.3.1 Participants 118
4.3.2 Stimuli 118
4.3.2.1 Face 118
4.3.2.2 Speech 119
4.3.3 Experimental protocol 119
4.3.4 Behavioral analysis 120
4.3.5 FMRI acquisition and preprocessing 122
4.3.6 Univariate analysis and ROI definition 122
4.3.7 FMRI single trial analysis 123
4.4 Results 123
4.4.1 Faces 123
4.4.1.1 Behavioral results 123
4.4.1.2 fMRI results 125
4.4.2 Voices 128
4.4.2.1 Behavioral results 128
4.4.2.2 fMRI results 129
4.4.3 Individual differences – an Exploratory Analysis 130
4.5 Discussion 131
4.5.1 Face representation built upon multiple exemplar exposures 131
4.5.2 Neural correlates of familiarized facial identities 132
4.5.3 Voice as a weaker cue for identity information 135
4.6 Limitations and Future Directions 137

4.7 Conclusion **138**

4.8 References **139**

5. General Discussion ***146***

5.1 Summary of the findings **146**

5.2 Connections of divergent findings between Studies 2 and 3 ..... **147**

5.3 Implications of emotional arousal in identity memory research **149**

5.4 Interactions between emotional expression and identity processing **150**

5.5 Methodological Implications **152**

5.5.1 Experimental paradigms 152

5.5.2. Testing platforms 153

5.6 Limitations and Future Directions **154**

5.6.1 Restricted use of stimulus database 154

5.6.2 Individual differences 154

5.6.3 Neural difference in recognizing high- and low-variability learned identities 156

5.6.4 Learning identities through multi-modal inputs 156

5.7 Conclusions **158**

General Reference List ***159***

Abstract

In daily interactions, people effortlessly recognize and identify familiar individuals through their faces and/or voices, even amidst the rich variability embedded in these social signals. The same task on unfamiliar and newly familiarized identities becomes, error prone and susceptible to perceptual variance, such as changes in emotional expression, which is a highly dynamic yet intrinsic component of face and voice. Emotional memory research has extensively demonstrated substantial memory benefit for emotional stimuli. However, such emotional benefits and the susceptibility to changes in emotional expression of person identity make the impact of emotional expression complex and nuanced.

This thesis aims to directly examine the influence of emotional expression on identity learning and recognition. Specifically, it explores how changes in emotional expression, and importantly, the extent of variability in the expressions during encoding, impact the learning and recognition of unfamiliar identities. Past research has examined memory of emotional faces extensively, but memory, especially identity memory of emotional voices received little attention. In addition, theoretical work tends to suggest large similarities in identity processing between the two modalities. Hence, I intend to examine the described research question in both faces and voices, through three closely connected studies.

Study 1 focuses on voice recognition, and investigates if, and to what extent, within-speaker changes in emotional expression affect subsequent speaker recognition. An additional source of variance - speech content - is added, aimed to provide a comparison to variance in emotional expression. Results from Study 1 demonstrate that speaker recognition is impaired when changes in either emotional expression or speech content is involved, and that higher encoding variability leads to a faster voice recognition. We continue investigating this encoding variance advantage in Study 2, in the aims of replicating (and expanding) the encoding variance advantage in Study 1, extending the research focus further to faces, and examining potential emotion-related factors that drives the advantage. Results reveal that low encoding variability with high-arousal emotional exemplars impairs identity recognition, but such a deficit can be compensated by high encoding variability, in both faces and voices. Finally, Study 3 examines how people explicitly recognize identities from novel emotional exemplars of previously encountered identities, at both the behavioral and neural levels. Results from the behavioral performance and neural activities in

regions within the Salience Network, indicate an improved and easier cross-expression recognition of the third novel exemplar for faces, but not for voices.

Collectively, the thesis demonstrates evidence across studies supporting the advantage of emotional exemplar variance on recognition of newly familiarized face and voice. The work also highlights an impaired recognition resulting from low variability learning with high-arousal emotional exemplars, compared to low-arousal ones. This arousal-based account can further help reconcile contradictory findings from past studies concerning categorical emotion specific influences on face recognition. Overall, this thesis helps contribute to a better understanding of the relationship between processing of identity information and emotional expression. In addition, it also provides methodological implications for face and voice memory paradigms using emotional stimuli, and for behavioral testing across different platforms.

Résumé

Dans les interactions quotidiennes, les gens reconnaissent et identifient sans effort les personnes familières grâce à leur visage et/ou à leur voix, même si ces signaux sociaux présentent une grande variabilité. La même tâche sur des identités non familières ou récemment familiarisées devient sujette à erreur et sensible à la variance perceptive, comme les changements dans l'expression émotionnelle, qui est une composante hautement dynamique mais intrinsèque du visage et de la voix. La recherche sur la mémoire émotionnelle a largement démontré les avantages substantiels de la mémoire pour les stimuli émotionnels. Toutefois, ces avantages émotionnels et la sensibilité aux changements d'expression émotionnelle de l'identité de la personne rendent l'impact de l'expression émotionnelle complexe et nuancé.

Cette thèse vise à examiner directement l'influence de l'expression émotionnelle sur l'apprentissage et la reconnaissance de l'identité. Plus précisément, elle se penche sur la façon dont les changements dans l'expression émotionnelle et, surtout, l'étendue de la variabilité des expressions pendant l'encodage, ont un impact sur l'apprentissage et la reconnaissance d'identités non familières. Les recherches antérieures ont largement examiné la mémoire des visages émotionnels, mais la mémoire, en particulier la mémoire de l'identité des voix émotionnelles, a reçu peu d'attention. En outre, les travaux théoriques tendent à suggérer de grandes similitudes dans le traitement de l'identité entre les deux modalités. Dans cette thèse, j'ai donc examiné la question de recherche décrite à la fois pour les visages et les voix, par le biais de trois études étroitement liées.

L'étude 1 se concentre sur la reconnaissance vocale et cherche à savoir si, et dans quelle mesure, les changements d'expression émotionnelle chez d'un locuteur affectent la reconnaissance ultérieure du locuteur. Une source supplémentaire de variance - le contenu du discours - est ajoutée, afin de fournir une comparaison avec la variance de l'expression émotionnelle. Les résultats de l'étude 1 démontrent que la reconnaissance du locuteur est altérée lorsque des changements dans l'expression émotionnelle ou le contenu du discours sont impliqués, et qu'une plus grande variabilité d'encodage conduit à une reconnaissance vocale plus rapide. L'objectif est d'étude de cet avantage de la variance d'encodage dans l'étude 2, dans le but de reproduire (et éventuellement d'élargir) l'avantage de la variance d'encodage de l'étude 1, d'étendre le champ de recherche aux visages et d'examiner les facteurs potentiels liés à l'émotion qui sont à l'origine de cet avantage. Les résultats révèlent qu'une faible variabilité d'encodage

avec des exemples émotionnels à fort niveau d'éveil nuit à la reconnaissance de l'identité, mais qu'un tel déficit peut être compensé par une forte variabilité d'encodage, tant pour les visages que pour les voix. Enfin, l'étude 3 examine comment les personnes reconnaissent explicitement des identités à partir d'exemples émotionnels nouveaux d'identités précédemment rencontrées, tant au niveau comportemental que neuronal. Les résultats des performances comportementales et des activités neuronales dans les régions du réseau de saillance indiquent une reconnaissance améliorée et plus facile de l'expression croisée du troisième nouvel exemplaire pour les visages, mais pas pour les voix.

En somme, la thèse démontre que les études soutiennent l'avantage de la variance des exemples émotionnels sur la reconnaissance des visages et des voix nouvellement familiers. Les travaux mettent également en évidence une altération de la reconnaissance résultant d'un apprentissage à faible variabilité avec des exemplaires émotionnels à fort niveau d'éveil, par rapport à ceux à faible niveau d'éveil. Cette explication basée sur l'éveil peut aider à réconcilier les résultats contradictoires d'études antérieures concernant les influences spécifiques des émotions catégorielles sur la reconnaissance des visages. Dans l'ensemble, cette thèse contribue à une meilleure compréhension de la relation entre le traitement de l'information sur l'identité et l'expression émotionnelle. En outre, elle fournit également des implications méthodologiques pour les paradigmes de mémoire des visages et des voix utilisant des stimuli émotionnels, et pour les tests comportementaux sur différentes plates-formes.

Acknowledgments

It has been such a journey to completing this dissertation, through the ups and downs, and all the challenges in between. I am indebted to all the people who helped me along this journey in many ways.

First and foremost, my deepest gratitude to my supervisor, Dr. Jorge Armony. Through countless meetings, discussions, and rounds and rounds of revisions and feedbacks, you constantly pushed me beyond my comfort zone and challenged me to think deeper and more independently. You always managed to reignite my curiosity and enthusiasm, providing guidance that pulled me out of dead-end moments. I am incredibly grateful for your guidance throughout these past years of research.

Then I would like to thank my advisory committee members, Drs. Signy Sheldon and Bratislav Misic, for offering invaluable insights and suggestions across multiple meetings, both related to the thesis work and about conducting research in general. Your questions and comments helped challenge myself to reflect and understand more of my research topic and methods.

Following that, I would like to extend my gratitude to the MRI technologists, Ron Lopez, David Costa, and late Louise Marcotte, for all the support and assistance during conducting my third study. And a special thanks to CRBLM's own Heather MacDougall for graciously handling and fulfilling all our requests, especially during the challenging times of COVID. Our work wouldn't be completed without the technical and practical supports from you.

To my lab mates, it has been an honor to work alongside with each one of you, Dr. Jocelyne Whitehead, Fumika Kondo, Soren Wainio-Theberge, Drs. Ignacio Spiousas and Peer Herholz. I have learned a lot from you, from inspiring academic discussions to casual chats. Your presence has made my time here stimulating and exciting. This journey would not have been as fun and enjoyable without you. On a personal note, I want to thank my dear friend Lei Liu, and my Soulstice family, for keeping my musical hobby alive and providing a distraction from the challenges of work at times. It was healing, and necessary.

Last but not least, I would like to thank my parents, who have provided me a worry-free environment to pursue this PhD path and the unconditional support, which helped me navigate and overcome some of the dark moments along this journey.

Contributions of original knowledge

1. We have demonstrated and replicated an emotional exemplar variance advantage for subsequent face and voice recognition across multiple experiments and testing platforms (i.e., in-lab and online).
2. We have shown that learning single repeated exemplar can affect subsequent identity recognition by the arousal level of the exemplars, namely that high aroused exemplars tend to interfere with identity learning and recognition.
3. We have demonstrated that the emotional exemplar variance advantage can be observed as early as after two distinct exemplars are learned in faces, but not significantly in voices.
4. We have found the recruitment of the Salience Network in response to faces perceived as compared to those perceived as old, especially after multiple encounters with the identities. This pattern may mirror the process of faces becoming familiar in a simplified way.
5. We have developed an effective encoding-recognition experimental paradigm to assess identity recognition using emotional exemplars. This is particular significant as current standardized tests for face and voice memory either do not utilize emotional exemplars, or incorporate them to a very limited extent.

Contributions of authors

This dissertation comprises the following 3 manuscripts (Chapters 2-4), of which I (HX) am the first author:

♦ **Chapter 2:** Xu, H.*, & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology, 74*(7), 1185-1201.

♦ **Chapter 3:** Xu, H.*, & Armony, J. L. (under review). Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition.

♦ **Chapter 4:** Xu, H.*, & Armony, J. L. (in prep). Cross-emotional-expression recognition in unfamiliar faces and voices – an fMRI study.

In each of the 3 manuscripts, the experimental design was developed collaboratively between HX and JLA. HX recruited subjects and collected behavioral and neuroimaging (fMRI) data in Chapters 2 through 4. Both behavioral and neuroimaging data was analyzed by HX in Chapters 2 to 4, under guidance from JLA. The first draft of each manuscript in Chapters 2-4 was written by HX, while subsequent drafts were reviewed and revised by JLA and HX. The final draft of each manuscript was approved by JLA. The inclusion of Chapters 2 to 4 in the current thesis were approved by both HX and JLA.

List of abbreviations

ACC	Anterior cingulate cortex
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
CEN	Central executive network
CI	Confidence interval
DDM	Drift diffusion model
DMN	Default mode network
FDR	False discovery rate
FFA	Fusiform face area
fMRI	Functional magnetic resonance imaging
fMRI-a	Functional magnetic resonance imaging-adaptation
FOV	Field-of-view
FRUs	Face recognition units
FWHM	Full-width-half-maximum
GLM	General Linear Model
GLMM	generalized linear mixed model
HRF	Hemodynamic response function
IAC	Interactive activation and competition
IOG	Inferior occipital gyrus
KDEF	Karolinska Directed Emotional Faces
LMM	Linear mixed model
MCC	Middle cingulate cortex
MNI	Montreal Neurological Institute
MPRAGE	magnetization-prepared rapid acquisition gradient echo
OFA	Occipital face area
PCC	Posterior cingulate cortex
PIN	Person identity nodes
pTFCE	Probabilistic threshold-free cluster enhancement
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
ROI	Region of interest
RP	Repetition priming
SD	Standard deviation
SE	Standard error
SFG	Superior frontal gyrus
SMA	Supplementary motor area
SN	Salience network
STG	Superior temporal gyrus
STS	Superior temporal sulcus
TE	Echo time
TP	Temporal pole
TR	Repetition time
TVA	Temporal voice area
VRUs	Voice recognition units

List of figures

Chapter 1. General Introduction

Figure 1-1. A demonstration of hierarchical models of face and voice processing 3

Chapter 2. Influence of emotional prosody, content and repetition on memory recognition of speaker identity

Figure 2-1. Recognition accuracy (a) and response times (b) in Experiment 1 31

Figure 2-2. Changes in response times (RTs) during encoding in Experiment 2 for the *Uni* and *Multi* conditions (relative to the first presentation) 38

Chapter 3. Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition

Figure 3-1. Procedure and example trial of the experimental task in the studies. 69

Figure 3-2. Averaged recognition accuracy and DDM-derived drift rates in the Main Study. 78

Figure 3-3. Confidence intervals (95%) of accuracy difference between *Multi* and *Uni* conditions from the Main Study using an n-jackknife subsampling approach 79

Figure 3-4. Scatterplot and estimated regression line of stimulus arousal intensity ratings against across-subject mean accuracy 81

Figure 3-5. Averaged recognition accuracy and DDM-derived drift rates in Follow-up Study 1 84

Figure 3-6. Averaged recognition accuracy and DDM-derived drift rates in Follow-up Study 2 86

Chapter 4. Cross-emotional-expression recognition in unfamiliar faces and voices – an fMRI study

Figure 4-1. Image samples of female and male face stimuli 119

Figure 4-2. Averaged recognition accuracy and DDM-derived drift rates of faces and voices 125

Figure 4-3. 2D renderings of the clusters under contrast P2-Correct vs. Incorrect for faces and parameter estimates in E2P1/E3P1 trials from ROIs 127

Figure 4-4. 2D renderings of the cluster under contrast P2-Correct vs. Incorrect for voices and parameter estimates in E2P1/E3P1 trials from the cluster 130

List of tables

Chapter 2. Influence of emotional prosody, content and repetition on memory recognition of speaker identity

Table 2-1. Descriptive statistics of recognition accuracy and response bias in Experiment 1 30

Table 2-2. Fixed effects from (G)LMM estimations on recognition response/bias/logRT in Experiment 1 32

Table 2-3. Descriptive statistics of recognition accuracy and response times (RTs) in Experiment 2 39

S.Table 2-1. Descriptive statistics and repeated-measures ANOVAs results of acoustic parameters for the stimuli used in Experiment 1.....58

Chapter 3. Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition

Table 3-1. Overview of the emotional expression combinations used in the studies 66

S.Table 3-1. Descriptive statitics of unbiased emotion recognition (ER) hit rates, ratings of valence and arousal from the selected face stimuli.....103

S.Table 3-2. Descriptive statitics of recognition measures from Main and two Follow-up studies 104

S.Table 3-3. Fixed factor effects from (G)LMM results on recongition accuracy and drift rates from the Main Study.....106

S.Table 3-4. Post-hoc pairwise comparisons on the variability-by-*Uni*-emotion interaction from the (G)LMM on old-identity trials of the Main study 107

S.Table 3-5. Fixed factor effects from the recognition confidence GLMM from the Main Study.....107

S.Table 3-6. Descriptive statistics of two cosine similarity indices of physical features of speech and face stimuli.....108

S.Table 3-7. Post-hoc pairwise comparisons on the emotional expression main effect of the LMM on stimulus-based cosine similarity of physical features.....109

S.Table 3-8. Fixed factor effects from the stimulus-level arousal rating model.....109

S.Table 3-9. Fixed factor effects from (G)LMM results on recongition accuracy and drift rates from the Follow-up Study 2.....109

Chapter 4. Cross-emotional-expression recognition in unfamiliar faces and voices – an fMRI study

Table 4-1. Group-level significant activation under the *F*-contrast of interest of the univariate analysis of face modality.....126

Chapter 1. General Introduction

Every day, we engage in the complex process of recognizing and identifying other people. Face and voice are two primary sources where we extract person identity information from. This critical function of human social cognition underpins our ability to form connections, communicate effectively, and maintain social relationships (Sidtis & Zäske, 2021). This process often gets interfered by other information that is conveyed through faces and voices. Indeed, a wealth of information can be extracted, such as a person's age, gender, emotional state, identity, and health. In order to successfully recognize or identifying a target individual, the extraction of the invariant features or signature of the face or voice while ignoring other changeable features is crucial. While people excel at recognizing familiar individuals both from both anecdotal and empirical evidence (e.g., Bruce, 1982), unfamiliar or newly familiarized individuals, which are most commonly used in lab-based experiments, are susceptible to changes in face images or vocal signals.

Emotional expression is another inherent component conveyed through both facial and vocal cues, and being continuously assessed during face and voice processing. Effective processing of emotional expressions is vital for human social interactions (Kringelbach & Berridge, 2009; Kreitewolf, Mathias, & von Kriegstein, 2017), allowing individuals to convey their internal states, intentions, and reactions to others. Besides its social significance, emotion is known to influence a wide range of cognitive processes, such as perception (e.g., Zadra & Clore, 2011; Niedenthal & Wood, 2019 for reviews), attention (e.g., Armony, Vuilleumier, Driver, & Dolan, 2001; Vuilleumier, 2005) and memory (e.g., LaBar & Cabeza, 2006; Tyng et al., 2017 for reviews).

Being two important and socially significant components embedded in facial and vocal information, the interplay between the processing of emotional expression and identity information has become a topic that interests many, especially in face research. Based on widely acknowledged models of face processing (Bruce & Young, 1986; Haxby, Hoffman & Gobbini, 2000) and voice processing (Belin, Fecteau & Bedard, 2004; Belin, Bestelmeyer, Latinus & Watson, 2011), I would like to expand current knowledge of such interactions between emotional expression and identity, from face to voice recognition, given the large similarities

proposed in the theoretical and empirical work (see Young, Frühholz, & Schweinberger, 2020 for a review).

Hence, it is the main focus of this dissertation to investigate the influence of changes in emotional expression on identity learning and memory in both faces and voices. A deeper examination of such influences, across modalities, will help us better understand the formation process of face and voice representation, and offer novel evidence to understand the relationship of emotional expression and identity information processing in face and voice. Lastly, I hope to address the similarities and differences between face and voice processing, by conducting the experiments with both face and voice stimuli so that results are comparable.

In this chapter, I first lay out the seminal models of face and voice processing, followed by the corresponding neural basis and theorized mental representation models that are well acknowledged and continuously investigated, to offer a clear picture of theoretical cross-modality similarities, and a foundation of the proposal of possible interactions between cognitive analysis of emotional expression and identity information when processing face and voice stimuli. Then I move on to discuss the flexibility of face and voice, providing a brief review on the influence of changeable stimulus features on face and voice recognition, that leads to a proposal of exemplar variance advantage, originated from representational theories. Lastly, I will narrow down the variance to the feature of interest, shared in both faces and voices – emotional expression, and review related research that has examined its influence on identity recognition. It opens up the intriguing question, that whether the emotional-expression-induced change is beneficial or detrimental for identity memory. The chapter finishes with a summary of thesis objectives, and a brief structural introduction of the experimental work of the following three chapters.

1.1 How do we recognize faces?

1.1.1 Face processing models

1.1.1.1 Functional face processing model

A cornerstone in the face research is the functional model of face processing by Bruce and Young (1986), which provides a parsimonious account of the cognitive processes involved in face processing and recognition.

The seminal model defines a sequence of cognitive stages, starting with a structural encoding of facial features (Figure 1-1). This stage involves a perceptual analysis of the presented face and extracts changeable and invariant features. Several pathways are separated from this stage, including expression analysis, facial speech analysis, directed visual processing, and face recognition units. The process of expression analysis is responsible for categorizing facial expressions from facial feature configurations, while facial speech analysis utilizes speech-related visual cues, particularly motions of lips and jaw, to assist understanding spoken languages. Directed visual processing is useful in guiding selective attention to the visual form of faces. The original model further proposes that these pathways are applicable to faces regardless of their familiarity, as facial identity is not relevant in the function of these pathways.

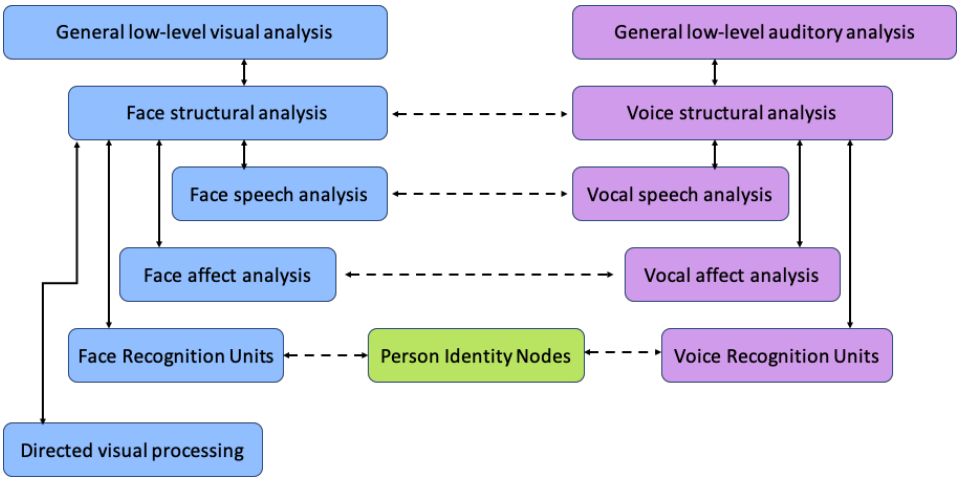

Figure 1-1. A demonstration of hierarchical models of face and voice processing (adapted from Belin, Fecteau & Bedard, 2004). Dash arrows indicate potential multimodal interactions.

The other pathway, however, going through face recognition units (FRUs), to the person identity nodes (PIN), eventually reaching name generation stage, is the main process responsible for face recognition and the retrieval of person-specific semantic information. According to Bruce and Young's model, the invariant face descriptions formed from the initial structural encoding stage, serve as inputs into FRUs, where all the known identities are stored. Successful recognition is achieved when a stored representation from the FRU is similar enough to the structural information extracted from the processed face. Upon a successful match, the FRU triggers the activation of associated PINs, enabling the retrieval of specific biographical information of the recognized identity, such as occupation, personal relationship, and name.

Information from other domains, such as equivalent counterparts of FRUs in voice (i.e., see voice recognition units below in Section 1.2.1), are expected to converge at the PIN.

The original face processing model also proposed that these pathways are functioning in parallel independently, which was supported by multiple early neuropsychological case studies (e.g., double dissociations of function impairment of recognition in face identity, but not emotional expression, or vice versa; Bruyer et al., 1983; Tranel et al., 1988). However, these results may be biased by methodological challenges (Calder & Young, 2005). In addition, more primate studies (e.g., Perrett et al., 1984; Hadj-Bouziane et al., 2008) and human neuroimaging studies (Sergent et al., 1994; Haxby, Hoffman & Gobbini, 2000) showed that, different brain regions, or cortical cell populations were selectively sensitive to either facial identity or expression, which is consistent with the concept of independent pathways. On the other hand, a growing body of studies presented results in support of an interacting or interdependent relationship between different processing pathways (see Fitousi & Wenger, 2013 for a review). For instance, an asymmetric relationship has been revealed that although perception of emotional expression was not modulated by changes in face identity, identity perception was on the contrary affected by changes in emotional expression in a perception experiment (Soto et al., 2015).

Following the Bruce & Young model, the Interactive Activation and Competition (IAC) model was further proposed and mainly elaborated on the FRU-PIN interaction in a connectionist network fashion (Burton, Bruce, & Johnston, 1990; Burton, 1994; Burton, Bruce & Hancock, 1999) based on the original model. In brief, the IAC model introduces the concept of cognitive competition, that the recognition of a familiar face involves a competitive process between different FRUs. Activation of target identity's FRU would accompany inhibitions of FRUs representing other individuals. Such competitions also extend into corresponding PINs. Thus, the IAC model can provide mechanical explanations to a range of face recognition phenomena observed in behavioral studies (see Burton, Bruce, & Johnston, 1990), with repetition priming being one of most frequently observed. Repetition priming is an effect that reflects a faster or easier recognition if the same identity had been previously seen (e.g., Bruce & Valentine, 1985; Ellis et al., 1987). The priming effect is strongest when the primer and test stimuli are identical, but were also found when the stimuli were different images of the same person. A strengthened FRU-PIN link due to the presence of the primer, would lead to a shorter time for the same person's PIN to reach activation. Furthermore, this effect is hypothesized to be

long lasting, as the strengthened link is not a transient change as the case in semantic priming. Indeed, experimental results did show a preserved repetition priming effect even when prime and test phases had minutes-long gaps (Ellis et al., 1987).

1.1.1.2 Neurological face processing model

The later developed neurological face perception model by Haxby, Hoffman and Gobbini (2000) provided fundamental frameworks for understanding the neural underpinnings of face processing. Haxby and colleagues' model proposed a network of brain regions, comprising two broad systems, namely the core system and the extended system, responsible of processing different aspects of face.

The core system includes the inferior occipital gyri (IOG), the superior temporal sulcus (STS), and the fusiform face area (FFA), responsible for different aspects of face processing. Specifically, the IOG is involved in the early-stage visual analysis of faces, while the STS processes changeable aspects of faces, such as facial expression and lip movement. The FFA, is proposed to process invariant aspects of face, including identity. The extended system involves brain regions that are not strictly dedicated to face processing but still plays a role in facilitating person perception, such as amygdala, insula, and other limbic regions. Both models from Bruce and Young (1986), and Haxby and collogues (2000), share the concept of distinct pathways for the visual analysis of identity and other changeable features, such as emotion analysis, at a cognitive and neural topographical level.

1.1.2 Familiar and unfamiliar faces

According to Bruce and Young's seminal model, FRUs hold stored representations of known faces, that are thought to encapsulate the invariant aspects of faces that distinguish one from another. Hence, recognition for familiar faces can be achieved through this pathway once a match is found between the stored representations and the inputs produced by structural encoding. Unfamiliar face, however, relies more on the visual codes produced by structural encoding and directed visual processing to compare or remember, as it lacks a stably formed face representation or an FRU.

Indeed, people excel in recognizing familiar faces, but had difficulty in unfamiliar (or less so, newly familiarized) faces. Psychological experiments, a majority of which were conducted on

young healthy adults (e.g., college students) due to the accessibility and convenience, have consistently demonstrated the discrepancies in recognition performance between familiar and unfamiliar faces (e.g., Johnston & Edmonds, 2009; Bonner, Burton, & Bruce, 2003). For example, Bruce (1982) tested recognition performance on both familiar and unfamiliar faces, when pose and/or expression was changed between test and study. Performance of unfamiliar face recognition, in both accuracy and response times, dropped when either component was changed (i.e., worse accuracy and slower speed), and was even significantly worse when both components were changed. Familiar face recognition, on the contrary, was not affected by such changes. Such a superior performance has also been seen in other familiar face identification and matching tasks (e.g., Burton et al., 1999; Bruce et al., 2001; Roark, O'Toole, & Abdi, 2003). Moreover, such a qualitative difference between familiar and unfamiliar face perception is also supported by results from neuropsychological (e.g., Ellis, Quayle, & Young, 1999; Malone et al., 1982) and neuroimaging studies (see Natu & O'Toole, 2011 for a review).

Understanding how faces become familiar is crucial, given the substantial distinctions between familiar and unfamiliar face recognition and identification. Exposure plays a vital role in fostering familiarity, as even a modest number of exposures can lead to rapid acquisition of familiarity (Bonner, Burton, & Bruce, 2003). Jenkins and Burton (2011) proposed that the familiarization involves an "exposure-driven refinement of the stored face representations", locate in the FRUs based on Bruce and Young's model. There have been two mainstream theories so far that are proposed to explain how these stored representations are formed and/or updated, namely the exemplar-based and prototype-based accounts.

1.1.3 Mental representations of face

The long-running debate between these two accounts started originally from category learning literature and is ongoing unresolved. The exemplar-based approach, proposes that individual exemplars are all stored and identity recognition is achieved by a successful match of the current face and previously stored exemplars (Knapp, Nosofsky & Busey, 2006; Longmore, Liu & Young, 2008). The prototype-based approach (or averaging model), on the other hand, claims that exemplar variation helps to construct a robust representation of encountered facial identities (e.g., Benson & Perrett, 1993; Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011), and that the representation becomes more stable when derived from more instances. The

latter proposal has received some support from experiments using either computer-generated faces (Bruce et al., 1991) or high quality images (Cabeza et al., 1999; Burton, Jenkins, Hancock & White, 2005), showing that participants would recognize an unseen prototype face, rather than a novel exemplar, after exposed to several exemplars of the same face that were used to generate the prototype. Cabeza and colleagues (1999) further found that the prototype effect can only survive in the same viewpoint or with small angular variations. As viewpoint plays a very important role in face processing (discussed in more detail in Section 1.3), this may point towards a possibility of a combination of both theoretical accounts used in face recognition (Bruce & Burton, 2002), that a number of prototypes of several viewpoints are stored. However, it is difficult to indeed disentangle one theoretical account from another, to directly probe which approach is the "real" representation model, as it is plausible to take one account and then reformulate it in the other one (e.g., Burton, Jenkins, & Schweinberger, 2011; see a similar argument in Zaki et al., 2003). And imperially, both accounts tend to provide converging predictions in face learning studies. Hence, it is not the intention of the current thesis to take on this long-running debate and seek experimental support in favor of one or the other account.

1.2 How do we recognize voices?

1.2.1 Voice processing model

Just like face, a wealth of information can be extracted from vocal signals, such as sex, region, identity, and emotion state. The leading model of voice processing, is proposed based on Bruce & Young's face processing model (Belin, Fecteau, & Bedard, 2004; Belin, Bestelmeyer, Latinus, & Watson, 2011), with a particular focus on identifying the neural correlates of different aspects of voice processing.

This model proposes that voices are processed in a hierarchical fashion, starting with low-level acoustic analysis of incoming sounds, which is thought to take place in subcortical nuclei and primary auditory cortex. This step is an equivalent counterpart as the structural coding stage in the face model. After this stage, speech-, affect-, and identity-specific information are then extracted and analyzed in three at least partially dissociable functional pathways (Figure 1-1), in a homologous way as in the face processing model (1986). These voice-signal specific analyses mostly take place in the Temporal Voice Areas (TVAs; Belin et al., 2000; Linden et al., 2011). Neuroimaging studies have consistently demonstrated the involvement of bilateral middle and

anterior superior temporal gyri/sulci (STG/STS) responding to vocal than non-vocal signals (e.g., Belin, Zatorre & Ahad, 2002; Pernet et al., 2015). Similar as in the face processing model (1986), voice recognition units is the pathway primarily in charge of processing invariant aspects of the voice, particularly speaker identity. Further examinations suggested functional specificity within the TVAs (see Belin, Bestelmeyer, Latinus & Watson, 2011). Particularly, the more anterior TVA regions, extending towards the temporal pole (TP), have shown to be involved in invariant representations of for both familiar and unfamiliar voice identities (e.g., Belin & Zatorre, 2003; Andics et al., 2010; Nakamura et al., 2001). Furthermore, different parts of the TP have shown distinct activation patterns to unfamiliar and familiar voices. The right superior TP seems sensitive to acoustic information for unfamiliar voices (Latinus, Crabbe, & Belin, 2009), which supports its proposed role in acoustic-based representation of unfamiliar voices. The inferior part of the TP, on the other hand, is related to storing non-verbal person-specific semantic information (Gorno-Tempini et al., 1998; Hailstone et al., 2010), which may be the neural correlates of the PIN. In addition, some supra-modal regions including precuneus, amygdala, inferior frontal gyrus, and anterior temporal lobe, are observed activated during voice processing as well (Blank, Wieland, & von Kriegstein, 2014).

There is evidence for both independence and interactions between these voice analysis pathways. Some neuropsychological findings from patient studies support the independence of speech and voice identity analyses. For example, studies in individuals with phonagnosia showed a disruption in speaker recognition or discrimination, while their ability to process emotion or speech related information remained intact (Garrido et al., 2009; Hailstone et al., 2010). The opposite cases were also reported, for example, stroke patients who suffered from aphasia, could still process voice identity information (van Lancker & Canter, 1982). On the other hand, some evidence in support of an inter-pathway interaction came from psychological studies in healthy participants. Studies on speech intelligibility reveal that listeners performed better in speech understanding from familiar speakers as opposed to unfamiliar ones (e.g., Goggin et al., 1991; Nygaard & Pisoni, 1998), and studies on speaker recognition showed an improved voice recognition and learning when listeners were exposed to stimuli in their native language (Perrachione, Pierrehumbert, & Wong, 2009; Perrachione et al., 2011; Orena, Theodore, & Polka, 2015). These results may point towards a more complex interactive yet partially independent relationship among these processing pathways.

1.2.2 Mental representation of voice

For voices, a popular view of voice representation is that, different voices are encoded in a multidimensional voice space in relation to a prototype voice (e.g., Baumann & Belin, 2010; Latinus et al., 2013; see Maguinness, Roswandowitz & von Kriegstein, 2018 for a review). A prototype is regarded as a representation of a very frequently encountered voice, or an average voice. The further away an individual voice is from this overall prototype voice, it is perceived as more distinctive (e.g., Mullennix et al., 2011). There have been inconsistent reports, however, on whether distinct or typical-like voices are easier to be remembered and recognized (e.g., Mullenix et al., 2011; Yarmey, 1991).

Many studies exploring the prototype-based coding mechanics focused on using different voices and between-speaker variability (e.g., Latinus et al., 2013; Latinus & Belin, 2011). To understand how a voice representation is formed or updated, a recent study directly tested specifically how a single voice identity is formed through variable exposures (Lavan, Knight & McGettigan, 2019). Participants first learned voice identities through multiple speech stimuli that were distributed in a ring-shape (away from the speaker's voice center) in a two-dimensional voice space, and they were then tested on the recognition with stimuli that were nested inside the ring-shape close to the voice center, and from the ring-shape. Results showed a superior recognition accuracy for stimuli around the center, and furthermore, a higher accuracy when the stimulus was closer to the center. Such results provide support to the prototype-based voice model, in the context of forming averaged abstract individual representations after exposures to various same-identity vocal signals, in a similar fashion as the prototype-based approach proposed in face representation formation (Burton, Jenkins, & Schweinberger, 2011; Valentine, 1991).

1.3 Flexibility of faces and voices

In both face and voice processing models, identity-specific and non-identity pathways tend to have a partially independent and partially interactive relationship. Considering the high flexibility and variability inherent in both face and voice stimuli, each modality introduces a variety of sources of variance that influence identity perception and recognition. This section

provides a brief review of some variances in each modality and how they have been shown to affect identity recognition.

1.3.1 Variance in faces

In daily life, faces are highly dynamic stimuli, containing transient changes such as expressions, makeup, hairstyle, pose, and less prominent changes like aging. As mentioned earlier, people can recognize familiar individuals despite changes from various sources, likely thanks to the invariant structural representation of familiar faces, while they do experience more mistakes and larger difficulty when recognizing or discriminating unfamiliar ones.

Inspired by debates in early object recognition literature (e.g., Marr, 1982; Bulthoff, Edleman, & Tarr, 1995; Hayward, 2003), viewpoint dependence has been extensively studied in face recognition. Its uniqueness also lies in its inherent feature of providing three-dimensional information while most of the other types of variance can only be perceived in a two-dimensional space (e.g., hairstyle, expression). A series of early studies tested recognition memory for faces when a viewpoint change was introduced at test. They consistently found poorer performance when faces with a changed viewpoint, rather than the same-view face images, were used for recognition test (Krouse, 1981), especially for unfamiliar faces (e.g., Baddeley & Woodhead, 1983; Hill & Bruce, 1996; O'Toole, Edelman, & Bulthoff, 1998; Longmore, Liu, & Young, 2008). Moreover, a larger viewpoint rotation between study and test views led to a worse recognition memory (Hill, Schyns, & Akamatsu, 1997). The subsequent question arises: can learning from multiple views compensate for the impaired recognition in a novel view? Exemplar-based approach may be useful for learning faces in this scenario, as prototype averaging effect was absent in face learning that involved large angular variations (Cabeza et al., 1999). However, learning from two different views did not guarantee a better recognition in a novel view, compared to learning from one view (Longmore, Liu & Young, 2008).

Other sources of variance have elicited similar results particularly in unfamiliar or newly familiarized face recognition. For example, changes in lighting condition (Braje et al., 1998), or image size (Kolers, Duchnicky, & Sundstroem, 1985) in the recognition test led to decreased recognition accuracy. These phenomena all points towards a possible explanation that learning one (as in most earlier studies) or a very limited amount of exposures per identity may not be

sufficient to construct a structural representation that are insusceptible to variance. Based on both exemplar-based and prototype-based account of face representation, more exposures are hypothesized to be beneficial.

In fact, recent research has realized that within-person variance, is a crucial component in establishing stable identity representation (Burton, 2013; Jenkins et al., 2011). Hence, a number of studies started using real-life ambient images (uncontrolled variability), taking advantage of the embedded within-person variability. Studies that contrasted learning conditions of high vs. low exemplar variability, indeed showed supporting evidence of an exemplar variance advantage for newly familiarized face recognition (Murphy et al., 2015; Ritchie & Burton, 2017; Matthews, Davis, & Mondloch, 2018; Gipson & Lampinen, 2020).

1.3.2 Variance in voices

Often times, voice is considered a weaker identity cue than face, which can get overshadowed or show interference effects in the co-presence of face stimuli (see Stevenage & Neil, 2014 for an overview). Although it is not applicable to straightforwardly assess the variability in the voice against face images, we do sense the higher variability in vocal signals, compared to a majority of face studies using static images. As only natural speech clips were used as experimental materials in the thesis, we focused on variances that occur in speech vocal signals (as opposed to non-speech vocalizations) and their influences on voice identity recognition.

As mentioned in Section 1.2, vocal speech analysis has been shown to interact, or interfere with vocal identity processing in some cases. For instance, improved voice recognition was reported when listeners learned voice stimuli in their native language, than in a foreign language (e.g., Perrachione, Pierrehumbert & Wong, 2009). Speech content is often manipulated as a within-speaker variance in voice learning experiments. Similar as the worse recognition performance resulted from viewpoint (and other source) changes in face, Zäske and colleagues (2014, 2017) observed reduced speaker recognition accuracy when speech content differed between study and test. Manipulating the length of presented audio excerpts, which is another unintentional way of manipulating speech content, also showed an influence on speaker recognition. For instance, Yarmey, Yarmey & Yarmey (1994) reported chance-level speaker recognition after a brief 15 second incidental exposure. While another study using about 1.5 minutes long audio materials with explicit instructions to remember speakers, yielded a clear

above-chance recognition (Papcun, Kreiman & Davies, 1989). Similar results of longer stimuli were found in other studies (e.g., Schweinberger et al., 1997; Kerstholt et al., 2004). From a different viewpoint, such results align with the exemplar variance advantage proposed earlier. It suggests that the longer exposure there is, the more (linguistic) information (i.e., more within-speaker variability) can be extracted to form or be compared to the target voice representation, which in turn leads to a superior recognition of encoded voice.

Although more sparse, a few studies also tested voice recognition using unsystematically controlled voice stimuli, similar to the ambient face image approach, to take advantage of the within-person variability. Several studies from Lavan and colleagues (2018; 2019a,b) used either voice materials extracted from TV shows, or lab-designed stimuli set that covered a variety of vocal recording scenarios (e.g., speaking styles and environments, recording sessions). These materials to some extent resemble the ambient face images in face research mentioned earlier. Results were less consistent in the few studies, providing weak support for the proposed exemplar variance advantage (Lavan et al., 2019a).

1.4 Emotional expression - common variance in face and voice

Belin and colleagues' voice processing model (2011) comprises homologous processing pathways as in Bruce and Young's face model (1986). For many face studies using static images as primary stimulus material, face-speech analysis is usually not at the center of research interest. That leaves the one common component that reflects changeable features shared in the processing of both modalities – emotional expressions. Efficient processing and accurate perception of emotional expressions are fundamental for effective social communication (Kringelbach & Berridge, 2009; Kreitewolf, Friederici, & von Kriegstein, 2014), and also have survival significance from an evolutionary perspective (Darwin, 1872; Leppänen & Hietanen, 2007; Sanders, Grandjean, & Scherer, 2005). Although we constantly receive and process emotional information from multiple modalities in complex daily interactions, we are also able to rely on single-modality inputs to process and decipher emotional expressions, which is often the cases in in-lab studies investigating modality-specific emotion perception (e.g., Schimer & Adophs, 2017; Bryant & Barrett, 2008; Paulmann & Uskul, 2014). As emerging evidence starts to support an (at least) partially independent, partially interacting functional relationship between the affect- and identity- processing pathways, I review, in this subsection, previous research, a

majority of which focused on emotional faces, that examined the mnemonic effects of emotional expression.

1.4.1 Recognition of emotional faces and voices

Many studies consistently showed that emotional faces (e.g., Sergerie, Lepage, & Armony, 2005; LaBar & Cabeza, 2006) and voices (e.g., Armony, Chochol, Fecteau, & Belin, 2007; Aubé, Peretz & Armony, 2013; Pichora-Fuller, Dupuis, & Smith, 2016) are better remembered and recognized than neutral ones. Markedly, this memory enhancement is stimulus specific and mostly occurs when study and test stimuli are identical. Hence, it was difficult to isolate identity recognition from image- or stimulus-based item recognition in many past memory studies, as recognition of the same stimuli and of the same identity can be two distinct tasks that involves a matching at the pictorial code level, or at the FRU level (Bruce & Young, 1986). This can be regarded as part of a more general phenomenon of emotional memory enhancement that includes emotionally charged objects and/or scenes (e.g., Kensinger, 2004; Kensinger & Schacter, 2005; Righi et al., 2012).

1.4.2 Identity recognition with varied emotional expressions

What happens when the emotional expression in the study and test materials changed? Prior studies reported a drop in explicit recognition accuracy, as well as longer response times (implicit measures) when the emotional expression was changed, compared to the no-change condition (e.g., Bruce, 1982; Chen & Liu, 2009; Liu, Chen & Ward, 2014; Nomi et al., 2013, for faces; Salove & Yarmey, 1980; Stevenage & Neil, 2014 for voices). This is similar to reported results with changes in other features (e.g., viewpoint and/or lighting condition changes in faces, speech changes in voices). This is not surprising, as for unfamiliar faces, the recognition tends to rely on the pictorial coding, or image/exemplar-based matching, and would become less image-dependent when familiarity to some extent is formed. We would then expect a recognition improvement if identity learning consists of a number of emotional expressions, according to the exemplar variance advantage mentioned above. With limited research conducted on this topic, evidence seemed to provide some weak support to the hypothesis in face (Liu, Chen & Ward; 2015; Liu et al., 2016), and not in voice (Lavan et al., 2019b).

1.4.3 Emotion-specific influence on face memory

Some work in face research suggested that the emotional influences on identity memory may be emotion-specific. One constantly studied effect is the "happy face advantage", that encoding happy faces could facilitate encoding or recognition of the face identity. Kottoor (1989) first reported the phenomenon, where participants learned individuals with three expressions (smile, pout, and neutral) and were tested on the same photos for recognition. Smiling faces were recognized better than faces with other two expressions. Granted, it suffered from the classic critique of possibly conflating image recognition with identity recognition due to the identical stimuli used in both study and test phases (Bruce, 1982). Nevertheless, later studies that specifically employed novel materials at test, still found a happy face advantage in subsequent recognition (D'Argembeau et al., 2003; D'Argembeau & van der Linden, 2007). A similar advantage of facilitated face recognition was also reported in faces previously studied in moderately positive expressions, compared to more intense happy or angry faces (Kaufmann & Schweinberger, 2004). Social and/or emotional significance was proposed to explain the phenomenon, linking happy/positive expressions with approval and satisfaction while angry/negative expressions with danger, threats, or disapproval (see discussion in D'Argembeau et al., 2003; D'Argembeau & van der Linden, 2007). However, there are studies suggesting otherwise. For example, Righi and colleagues (2012) reported a fearful, rather than happy, expression advantage in novel face recognition. Similarly, an advantage for angry faces was reported by Jackson, Linden and Raymond (2014). Moreover, Liu, Chen, & Ward (2014) carried out a rather comprehensive behavioral study that compared recognition performance on faces encoded in six basic emotion (joy, surprise, sadness, disgust, fear and anger). They observed a happy face training advantage, only in comparison to disgusted faces, but not other emotional faces. Overall, these studies challenged the proposed special effect of the happy expression, and the idea that whether the effect of emotional expressions could or should be interpreted based on specific emotions. Furthermore, little has been reported systematically regarding any similar advantage of specific emotions on voice identity memory (e.g., Saslove & Yarmey, 1980; Öhman, Erikson, & Granhag, 2013; Stevenage & Neil, 2014), which warrants more investigation in this avenue of research.

1.4.4 What's special about emotional expression?

Emotional expression is a highly dynamic feature in both face and vocal signals, like other features (e.g., viewpoints, lighting conditions and hairstyles in faces, see Bruce, 1982; Hill, Schyns & Akamatsu, 1997; Longmore, Liu & Young, 2008; Chen & Liu, 2009; vocalization [speech vs. non-speech], vowel, speech content and vocal style [spoken vs singing] in voices, Smith et al., 2018; Peynircioğlu, Rabinovitz, & Repice, 2017). However, emotional expression remains a unique source of variance in identity cue that receives extensive interest from researchers in multiple fields. We propose three important aspects that may set it apart from other sources of variance in face and voice.

Firstly, emotional expressions carry great biological significance and social relevance, that are usually not the case in other types of variance. Emotional expression, emerged as a product of evolution of social animals (Darwin, 1872; Zych & Gogolla, 2021), often conveys key information about an individual's internal state, intentions, and immediate reactions to environmental stimuli. Recognizing and interpreting these signals correctly can be crucial for survival, as they may indicate threats, friendly intentions, or the need for cooperation. The modern basic emotion theories (e.g., Darwin, 1872; Ekman, 1992) has identified a limited number of emotions that are biologically and psychologically fundamental for humans to handle life tasks, including fear, anger, joy, sadness, disgust, and surprise. Secondly, emotionally charged stimuli have demonstrated significant impacts on attention and memory processes (e.g., Armony, Vuilleumier, Driver, & Dolan, 2001; Kensinger & Schacter, 2005; Talmi et al., 2008). This can be better understood from another perspective of analyzing emotions – a dimensional model (Barrett & Russell, 1999; Russell, 2003). It suggests emotion to be interpreted in a meta-emotional space, rather than distinct emotion categorizations. The most common dimension researchers acknowledge nowadays are the arousal-valence two-dimensional space. Valence refers to whether an emotion is perceived pleasant (positive) or unpleasant (negative), while arousal depicts the intensity of the emotion. Both dimensions are often considered together or contrasted against each other in empirical work and are shown sometimes separate, sometimes interactive influences on in emotional stimuli processing (e.g, Robinson et al., 2004) or emotional memory (see Kensinger, 2007, 2009; Mather & Sutherland, 2011 for reviews). Lastly, emotional expression appears to be a cross-modality source of variance, that can be perceived separately, or integratively from face and voice (Schimer & Adolphs, 2017).

1.5 Overview of the current work

Current views of face processing and voice processing models convergingly suggest that, the processing of emotional expression and modality-specific identity tend to be partially interactive, hoisting the interests in examining the influence of encoding emotional expression and its change, on face and voice recognition. It is greatly valuable considering the sparse research in emotional voice and voice recognition. More recently, methodological awareness has been raised on separating identity memory from item memory, which usually involves novel test stimuli that are not exposed during initial study (encoding). It appears that for newly familiarized faces, studying exemplars with uncontrolled variability are beneficial for learning and recognizing faces later. However, evidence for a similar benefit in voice learning/recognition is weak and inconsistent, let alone when the encoding variance restricted within vocal or facial emotional expressions. Thus, the current thesis intends to examine and understand the core questions surrounding the effects of changes in emotional expression on face and voice recognition. Specifically, I aim to achieve three objectives through three presented studies:

(1) to examine whether exemplar variance from emotional expression, is beneficial for a better identity recognition (i.e., generalization to new exemplars);

(2) to understand whether emotional variance-related influences on identity recognition are emotion-specific, or modulated by other emotional-relevant features;

(3) to compare if emotional variance induced influences on voice and face recognition are similar or divergent, given that the proposed identity processing models and representation approaches between modalities have been often discussed and compared, but experimental outcomes were rarely tested within studies.

The experimental work comprises three chapters. Chapter Two focuses on the auditory modality, and examines the influence of changes and the extent of changes from two sources - emotional expression and speech content - on speaker recognition, in two behavioral experiments. Specifically, the first experiment of Chapter Two investigates whether speaker recognition suffers a decrease when encoding and test stimuli involves changes in emotional expression (prosody) and/or speech content, with a focus on fearful expression. The second experiment tests further on the encoding variance, that if encoding more variance from both

emotional expression and speech content, would help facilitate speaker recognition. Chapter Three follows up on the last experiment, and continues to investigate the influence of encoding variance from emotional expressions, on identity recognition, expanding the testing modality to both voice and face. It further tests the possibility that it is driven by certain emotional exemplars with two follow-up experiments. In Chapter Four, an fMRI study, further focuses on the scenario where people need to encode a variety of same-identity exemplars with distinct emotional expressions (as in one of the experimental conditions in Chapters Two and Three), and examines the explicitly behavioral recognition and neural activities when encountering novel emotional exemplars of previously seen individuals. It can be viewed as a simplified process of a face or voice becoming familiar, as they encode more novel emotional exemplars as the task goes on. Finally, Chapter Five summarizes the main findings across the three chapters, and discusses how these findings can advance our understanding on interactive relationships of processing between emotional expression and identity. I also used our proposed underlying mechanisms to help explain and reconcile some seemingly contradictory results from prior studies. I conclude by discussing methodological implications from the three studies that were conducted in multiple platforms and with various designs. Additionally, I explore certain limitations and suggest potential future directions that merit further investigation.

Chapter 2. Influence of emotional prosody, content and repetition on memory recognition of speaker identity

(Study 1)

Hanjian Xu[1,2,3], Jorge L. Armony[1,2,4]

[1]Douglas Mental Health University Institute, Verdun, Canada;
[2]BRAMS Laboratory, Centre for Research on Brain, Language and Music, Montreal, Canada;
[3]Integrated Program in Neuroscience, McGill University, Montreal, Canada;
[4]Department of Psychiatry, McGill University, Montreal, Canada

Xu, H., & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology, 74*(7), 1185-1201.

2.1 Abstract

Recognizing individuals through their voice requires listeners to form an invariant representation of the speaker's identity, immune to episodic changes that may occur between encounters. We conducted two experiments to investigate to what extent within-speaker stimulus variability influences different behavioral indices of implicit and explicit identity recognition memory, using short sentences with semantically neutral content. In Experiment 1 we assessed how speaker recognition was affected by changes in prosody (fearful to neutral, and vice versa in a between-group design) and speech content. Results revealed that, regardless of encoding prosody, changes in prosody, independent of content, or changes in content, when prosody was kept unchanged, led to a reduced accuracy in explicit voice recognition. In contrast, both groups exhibited the same pattern of response times (RTs) for correctly recognized speakers: faster responses to fearful than neutral stimuli, and a facilitating effect for same-content stimuli only for neutral sentences. In Experiment 2 we investigated whether an invariant representation of a speaker's identity benefited from exposure to different exemplars varying in emotional prosody (fearful and happy) and content (*Multi* condition), compared to repeated presentations of a single sentence (*Uni* condition). We found a significant repetition priming effect (i.e., reduced RTs over repetitions of the same voice identity) only for speakers in the *Uni* condition during encoding, but faster RTs when correctly recognizing old speakers from the *Multi*, compared to the *Uni*, condition. Overall, our findings confirm that changes in emotional prosody and/or speech content can affect listeners' implicit and explicit recognition of newly familiarized speakers.

Key words: speaker recognition; emotional prosody; exemplar repetition

2.2 Introduction

As is the case with faces (e.g., Bruce & Young, 1986), voices convey an array of important information about an individual (e.g., Schweinberger et al., 2014; Young, Frühholz, & Schweinberger, 2020). Whereas some of these cues depend on the speaker's current emotional state and intention (e.g., prosody and speech content), others are more stable, and help us recognize people we encountered in the past. This task requires the ability to extract, store, and match invariant characteristics of individuals' voices and disregard features that can vary upon different encounters. While this may appear effortless in the case of familiar individuals, it becomes more difficult for unfamiliar individuals whom we encountered only a handful of times (e.g., see Burton & Jenkins, 2011 for faces; Stevenage & Neil, 2014, Lavan et al., 2019a for voices). While there are many factors that can influence our ability to correctly distinguish previously encountered individuals from those who we met for the first time, existing memory literature – using mainly faces and, to a lesser extent, voice – highlights the importance of emotional expression, number and variety of exposures and, in the case of speech, content.

Emotion, as a natural feature of social stimuli, is known to facilitate long-lasting same-stimulus recognition accuracy and confidence (e.g., Kensinger, 2004; Kensinger & Schacter, 2005; LaBar & Cabeza, 2006; Righi et al., 2012). However, as a majority of studies of face (e.g., Sergerie, Lepage & Armony, 2005; LaBar & Cabeza, 2006) and voice (e.g., Armony, Chochol, Fecteau, & Belin, 2007; Aubé, Peretz & Armony, 2013; Pichora-Fuller, Dupuis, & Smith, 2016) memory primarily examined item memory for the exact same stimuli, it is difficult to disentangle the possible effects of emotion on item-specific memory from those on stimulus-independent identity memory. A recent behavioral study (Liu, Chen, & Ward, 2014) directly examined this issue by comparing the effect of six basic emotional expressions (i.e., happiness, sadness, fear, surprise, anger, and disgust) on long-term facial identity memory. Participants were shown faces of only one of the six expressions multiple times at training, and completed a standard old/new identity-recognition test afterwards on faces either with the same emotion (i.e., same stimulus), or with a neutral expression. Fear-, happy- and sad-trained identities were worse recognized when the test expression was neutral compared to when it was the same expression as during encoding, with no differences in the extent of the recognition impairment among these three types of training. Moreover, Redfern and Burton (2017a) found that participants tended to make

more mistakes when discriminating pictures from two individuals when they were emotionally expressive than when they depicted a neutral expression.

Saslove and Yarmey (1980) provided initial evidence that the change of emotional prosody from anger to neutral between training and test in a voice line-up task impaired subsequent recognition. However, another voice line-up experiment showed no emotion-change effect on listeners' voice memory, even with different testing delays (Öhman, Eriksson, & Granhag, 2013). The effect of prosody change was also examined in a same/different voice matching paradigm, in which participants were asked to make decisions on whether pairs of phrases presented in angry, happy, and neutral tones were produced by the same speaker or not (Stevenage & Neil, 2014). Results revealed a decline in performance when the emotional tone changed between two phrases. Thus, there is some evidence to suggest that changes in emotional prosody negatively influence working and/or episodic memory performance, although results are inconsistent.

Stimulus repetition is another factor that has been shown to influence identity memory. Although pure repetition may not be sufficient to form stable face representations that are stimulus-invariant (e.g., Bruce et al., 2001), several studies using faces show that subsequent recognition performance can be improved by learning from face images with a longer exposure duration (Memon, Hope, & Bull, 2003), and repetitions of the same face images (Roark et al., 2006) or of non-identical face images in neutral expression (Kaufmann, Schweinberger, & Burton, 2009). In addition to explicit recognition, stimulus repetition has been shown to enhance implicit memory, a phenomenon known as repetition priming (RP) and typically reflected in faster response times when responding about a given feature of a previously presented as a function of the number of repetitions of said item. RP effects for faces are observed for both familiar and, albeit to a lesser extent, for unfamiliar identities (Goshen-Gottstein & Ganel, 2000). In the case of unfamiliar faces, RP effects can be highly view-dependent (Martin et al., 2010), although some studies also found view-invariant RP effects with increased number of exposures (Martin & Greer, 2011; Clutterbuck & Johnston, 2005).

Although less studied, there is some evidence to suggest that memory for voice identity also benefits from multiple stimulus repetitions. For example, Neil and colleagues (see Stevenage & Neil, 2014) conducted a sequential same/different match task by increasing repetition times of the stimuli. Between each matching pair of voices, interference was introduced by adding 0 or 4

distractors. As expected, interference decreased matching performance, but repeatedly pre-exposed voices showed a resistance of the interference effect when compared to singly pre-exposed voices. Similarly, Zäske et al. (2014) showed that stimulus repetition strengthened subsequent voice identity recognition.

A related question is whether subsequent identity memory is better when the same stimulus is repeatedly encoded, compared to encoding different exemplars of the same individual. Two main representation models, largely based on faces, both predict an exemplar variation advantage. The pictorial coding model proposes that identity recognition is completed through comparisons with previously stored exemplars of the individual (e.g., Longmore, Liu, & Young, 2008); thus, the more variant exemplars encountered, the higher the chance of a successful match. The averaging model proposes that exemplar variation helps to construct a robust representation of encountered facial identities (e.g., Benson & Perrett, 1993; Jenkins & Burton, 2011), and that the representation becomes more stable when derived from more instances. Consistent with this hypothesis, Murphy et al. (2015) revealed a better identity recognition with novel face exemplars when face learning was enriched with multiple variant exemplars. Similar advantages were reported in name- and face-matching tasks after face learning with high within-identity variability, over low variability (Ritchie & Burton, 2017). Interestingly, Liu et al. (2015) found no difference in face identity recognition when comparing exposure to three different emotional expressions with that of only one expression during learning, but a better performance when contrasting the 3 emotional expression condition to one in which only neutral faces were presented. In contrast to the face literature, the possibility of a multiple exemplar advantage for voice identity memory has been little explored, with the few studies conducted providing only limited support for such an effect (Lavan et al., 2019c).

Finally, a few studies investigated memory for voice identity when the speech content was changed between encoding and recognition. As expected, better memory performance was observed when the content was kept the same (i.e., same stimulus), but there was nonetheless an above chance identity recognition for different-content stimuli (Zäske et al., 2014, 2017). Furthermore, identity recognition has been shown to be preserved even after manipulations that altered vocal quality or temporal-based phonetic information (Sheffert et al., 2002). Interestingly, better changed-content memory performance was reported for emotional compared to neutral

voices (Kim, Sidtis & Sidtis, 2019), suggesting that an interaction between emotion and content may exist.

Here, we report results from two studies designed to address some of the gaps and inconsistencies, as well as to extend findings, in the literature described above. Experiment 1 consisted of a between-group factorial design investigating how changes in emotional prosody (see Saslove & Yarmey, 1980; Öhman, Eriksson, & Granhag, 2013; Stevenage & Neil, 2014), content (see Zäske et al., 2014, 2017; Kim, Sidtis & Sidtis, 2019) and their interaction (see Kim, Sidtis & Sidtis, 2019) affect memory for voice identity. In Experiment 2, we applied a within-subject design in which the number of emotional speech exemplars was varied, in order to assess whether findings obtained in the implicit (repetition priming) and explicit (recognition) memory literature on faces (Martin & Greer, 2011; Murphy et al., 2015; Redfern & Benton, 2017a) also apply to voices. Furthermore, a comparison between Experiment 1 and Experiment 2 allowed us to test whether increasing the number of repetitions of a stimulus improves memory performance (e.g., Memon, Hope, & Bull, 2003; Roark et al., 2006).

2.3 Experiment 1

We employed a classic incidental old/new recognition task to investigate the effects of changed emotional prosody and content on subsequent voice identity recognition. We focused on fear, as previous studies from our group (Sergerie, Lepage, & Armony, 2005; Armony, Chochol, Fecteau, & Belin, 2007; Aubé, Peretz, & Armony, 2013) and others (e.g., LaBar & Cabeza, 2006; Pichora-Fuller, Dupuis, & Smith, 2016) have consistently shown enhanced memory accuracy for same-item fearful expressions, which has been ascribed to an amygdala-mediated preferential process of such stimuli that signal the potential presence of danger in the environment (Armony, 2013; Sangha, Diehl, Bergstrom, & Drew, 2020). Thus, according to this view, fearful prosody should serve as an emotionally arousing factor that facilitates processing and storing the voice identity; on the other hand, it introduces acoustic variability to the same identity, which would interfere with the memory encoding or retrieving process. Two groups of subjects participated in this experiment: one was exposed to fearful-prosody neutral-content sentences of various speakers at encoding and tested for identity memory using sentences from these speakers in both fearful and neutral prosodies (and with the same or different content). A second group underwent a similar paradigm but was exposed to neutral prosody sentences during

encoding. Within- and between-subject analyses were conducted to assess the effects of changing prosody and content on voice identity memory and whether encoding voices with fearful or neutral prosody led to changes in memory performance.

2.3.1 Methods

2.3.1.1 Participants

Sixty volunteers (34 female, aged 18-43 years) were recruited from the Greater Montreal Area, and participated in the experiment at the International Laboratory for Brain, Music, and Sound Research (BRAMS), Centre for Research on Brain, Language, and Music (CRBLM), or Douglas Mental Health University Institute at McGill University. A power analysis on our pilot data using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) indicated that 58 participants (N = 29 per group) would be sufficient to detect an expected effect of .48 with a power of .95 and an alpha level at .05. All of the participants were fluent in English, right-handed, had normal hearing and (corrected-to-) normal vision, and reported no previous diagnosis or treatment of psychiatric or neurological disorders. They provided written informed consent prior to participation and received monetary compensation after the experiment. The study was approved by the Faculty of Medicine Research Ethics Office at McGill University.

2.3.1.2 Stimuli

Auditory stimuli were selected from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018). They were audio-only recordings of 24 speakers (12 female) uttering two sample sentences of semantically neutral contents (“Kids are talking by the door” and “Dogs are sitting by the door”, hereafter referred as “kids” and “dogs” sentences, respectively), in neutral and strongly fearful prosodies, resulting in 48 speech stimuli in total (12 speakers × 2 prosodies × 2 contents). The two sentence samples share the same syntactic structure and same number of syllables and were rated similarly in terms of emotional intensity (see Table S1 of Livingstone & Russo, 2018). Speakers from the RAVDESS were native English speakers, with a neutral North American accent, to minimize the possible use of accent variability as a strategy to identify speakers (Gluszek & Dovidio, 2010). Only half of the stimuli were used in Experiment 1 (selection procedure described below), as a pilot memory test

using the full set of 24 speakers resulted in a chance-level memory performance. Loudness of all the speech stimuli was normalized with the Loudness Toolbox (Genesis S.A.) in Matlab 2017b.

2.3.1.3 Speaker Selection

We employed a speaker matching task to select a subset of the 12 most identifiable speakers when the speech prosody switched between fear and neutral, in order to reduce task difficulty and improve memory performance (Legge, Grosmann & Pierper, 1984). A separate group of eighteen participants (11 female; aged 18 – 32 years) participated in this experiment. Each participant completed the matching task sitting in front of a computer while listening to audio stimuli via Beyerdynamic DT 770/990 headphones. In each trial, a sentence in fearful prosody was presented, followed by another one with neutral prosody, with either the same or different speech content, from the same or a different (but same-sex) speaker, with a 200 ms inter-stimulus interval. Participants were asked to decide whether the two sentences were spoken by the same person by pressing the corresponding button on a keyboard. All possible same-sex speaker pairs of fearful and neutral sentences were divided in 6 runs. Each run consisted of 24 speakers (uttering a fearful sentence) paired with three individuals (speaking a neutral sentence): one being him-/her-self, the other two being pseudo-randomly assigned different same-sex speakers, ensuring content difference was counterbalanced. Each participant completed two out of the six runs, which were assigned pseudo-randomly so that in the end, each possible speaker pair was compared by 6 participants.

Average accuracy of matching performance was calculated for each of the twenty-four speakers across participants. Speakers were ranked by the matching accuracy in each sex separately (range: 0.48 - 0.79). The six male and six female speakers with the highest matching accuracy were selected for Experiment 1. No significant difference in accuracy was observed between the selected male (M = 0.66, SD = 0.07) and female (M = 0.71, SD = 0.04) speakers ($t(10) = 1.54$, $p = .15$, Hedges's $g_s = 0.81$). A post-hoc *t*-test confirmed that the selected twelve speakers were matched significantly more accurately than the unselected ones ($t(22) = 6.73$, $p < .001$, Hedges's $g_s = 2.65$).

2.3.1.4 Acoustic Features Analysis

To examine the acoustic (dis)similarity of the speech clips, we compared the acoustic differences

between stimuli as a function of their prosody and content. Seventeen physical acoustic parameters were included in the tests, which were extracted from each stimulus using Praat v6.1.04 (Boersma & Weenink, 2019); these included stimulus duration, and descriptive statistics (i.e., means and standard deviations) of the fundamental frequency F0, formant frequencies (F1-F4), and amplitude, as well as min, max and range of F0. While there is no consensus on which, and how many, parameters best represent vocal stimuli, those chosen here were selected from previous studies using shorter stimuli (e.g., Baumann & Belin, 2010; Latinus et al., 2013; Fecteau et al., 2007), and also included measures of within-stimulus variability (i.e., range and standard deviation) to account for the longer duration of the stimuli we used. These parameters have been previously shown to capture relevant aspects of speaker's identity and emotional expression. For instance, F0 and lower formant frequencies are important for voice identification (Xu et al., 2013; Matsumoto et al., 1973). Specifically, average fundamental frequency is an important source for listeners to distinguish or recognize speakers (Baumann & Belin, 2010; Chhabra et al., 2012) and their emotional state (Pichora-Fuller, Dupuis, & van Lieshout, 2016). Higher formant frequencies, especially F3 and F4, which relate to the size of a speaker's vocal tract, are thought to carry information about voice identity (e.g., Remez, Fellowes, & Rubin, 1997; Ghazanfar & Rendall, 2008) and remain invariant when uttering different vowels or tones (e.g., Kitamura et al., 2006; Takemoto et al., 2006).

A prosody-by-content repeated measures ANOVA on the 12 speakers (for full results, see supplementary S.Table 2-1) revealed significant main effects ($p < .05$, false discovery rate (FDR) corrected with the Benjamini-Hochberg approach; Benjamini & Hochberg, 1995) of prosody for min and max F0, and for mean F0, F1 and F2. In addition, there was a main effect of content for mean F3 and for standard deviation of F3, F4 and amplitude. No content-by-prosody interactions reached statistical significance.

Additionally, to relate the acoustic features with subjects' memory performance, we took these parameters as a feature array representing each stimulus in the multidimensional acoustic feature space (Armony, Chochol, Fecteau, & Belin, 2007; Baumann & Belin, 2010; Latinus et al., 2013). An average within-prosody distance for each stimulus was computed by averaging the Euclidean distances between the specific stimulus and the others from its prosody group. These mean Euclidean distances between two prosodies were compared in a Mann Whitney U test, to avoid the violation of variance homogeneity assumption. Fearful stimuli (Mean Rank (MR) =

33.29) were more distant among each other than neutral ones (MR = 15.71) in the multi-dimensional acoustic feature space (U = 77.00, $Z = 4.35$, $p < .001$, $\eta^2 = .39$). A similar analysis as a function of content reveled no significant differences in within-content distance between the "kids" (MR = 21.25) and "dogs" (MR = 27.75) sentences (U = 210.00, $Z = 1.61$, $p = .11$, $\eta^2 = .05$).

Finally, a complementary analysis on the speech similarity within each prosody was further conducted with a machine learning approach using the *caret* library (Kuhn, 2020) in R (version 4.0.0; R Core Team, 2020). Specifically, we trained a classifier to categorize speech prosody on the acoustic parameters extracted from different (not used in the experiment) exemplars of the 48 stimuli (12 speakers, 2 contents, and 2 prosodies), taken from RAVDESS, using support vector machine (SVM) with a linear kernel and a 10-fold cross validation procedure repeated 1000 times. The model was then used to identify the prosody of the stimuli we used in the study. All of the acoustic parameters were beforehand normalized due to the large discrepancies between their ranges. The trained model yielded an overall classification accuracy of 89.58%, significantly above chance level ($p < 10^{-8}$), with a kappa of 0.79. The prediction error was 20.83% among fearful clips, yet 0% in neutral clips. That is, results from the classifier were consistent with those from dissimilarity score comparisons, and together suggest that fearful speech clips were less similar to each other than neutral ones.

2.3.1.5 Procedure

Seated in front of a monitor, participants wore DT 770/990 headphones and used a keyboard to complete the task in a quiet room. They were instructed to press one of two keys (left/right) on the keyboard to answer the questions. Key assignment was counter-balanced across participants. Participants were asked to respond as quickly and accurately as possible. The experiment was self-paced; that is, once a response was made, it moved on to the next trial automatically, without an inter-trial interval (Steinborn et al., 2010). No break was taken throughout the experiment.

The experiment consisted of a short encoding session and a recognition test. During the encoding session, participants were asked to identify the sex of the speaker. Six speech clips, each produced by a different speaker (half male), were presented twice. Half of the participants were assigned to the *Fear* group, where all sentences presented were in a fearful prosody; the other half (*Neutral* group) listened to sentences with a neutral prosody instead (content

counterbalanced in both groups). The speaker recognition test took place immediately after encoding. Subjects were presented with 4 speech clips (2 prosodies × 2 contents) produced by each of the 6 speakers from the encoding session (i.e., old speakers) and 6 novel speakers, in a pseudo-randomized order. Each speech clip was followed by an old/new judgment question on voice identity. Participants were explicitly instructed to ignore any potential changes in the stimuli and only focus on speakers' identities. Response choice and time were recorded for each trial and submitted to analyses as described below.

2.3.1.6 Data Analysis

Encoding

Encoding response times (RTs) were examined for potential priming effects due to repetitions of the same voice identity, by implementing a regression coefficient analysis (RCA, Lorch & Myers, 1990) via linear mixed models. As we assumed a linear decrease trend in RTs as a function of repeated presentation (Xu, 2017), the slopes of RT change were estimated via linear regression. Based on the principle of RCA, we estimated the regression slopes at individual- and speaker-specific levels. These subject- and speaker-specific slopes were then analyzed in a linear mixed model (LMM) using the *lme4* library (Bates, Mächler, Bolker, & Walker, 2015) implemented in R, with group (*Fear*/*Neutral*) as the fixed between-subjects factor, and subject and speaker as random effects. Including speaker in the random effect structure can account for potential confounding speaker-specific effects and remove these from the fixed effects of interest (e.g., Baayen et al., 2008).

Recognition

Accuracy: Subjects' responses to each trial of previously presented speakers, coded as a binary variable (0 = "new", 1 = "old"), were fitted with a generalized linear mixed-effects model (GLMM) with a logit link function, with prosody (same/different, compared to encoding) and content (same/different) as the fixed within-subjects factors, and group as a between-subjects factor. For the specification of random effects, we used a maximal structure including both by-subject and by-speaker random intercepts and slopes of within-subjects fixed factors, in order to maximize the modelling generalizability (Barr, Levy, Scheepers, & Tily, 2013). When significant interaction effects were found, we conducted post-hoc *t*-tests (Bonferroni-corrected) to interpret the interactions using the *emmeans* R library (Lenth, 2020).

Additionally, to investigate whether effects of prosody on memory performance could be accounted for by acoustic (dis)similarity within and between prosody categories, a stimulus-based ANCOVA on the subject-averaged recognition accuracy was carried out, with emotional prosody (fearful/neutral) as a between factor and mean within-prosody distance as a covariate. A similar ANCOVA was conducted with within-content distance as a covariate.

Response Bias: To determine whether any differences obtained in the previous analysis could be accounted for, at least in part, to a different response strategy or bias as a function of the experimental manipulation, we computed the response bias (Br) for each subject and prosody, based on the 2-high threshold model (Snodgrass & Corwin, 1988): $Br = \frac{FA}{1-(H-FA)} - 0.5$, in which H and FA represent hit (correctly respond "old" when the voice identity was encountered before) and false alarm (falsely respond "old" when the voice identity was never encountered before) rates, respectively. Br is independent from memory performance, as it represents the tendency to respond "old" or "new" regardless of response accuracy. Positive values of Br indicate a tendency to respond "old", while a negative Br suggests a tendency to respond "new" (Sergerie, Lepage, & Armony, 2007). Br scores were analyzed with an LMM with prosody (same/different than encoding) as the only within-subjects fixed factor (as no same/different content could be assigned to new stimuli) and group as a between-subjects factor. The model also included subject random intercepts and slopes.

Response Times: We first applied a conventional RT cleaning procedure to exclude those shorter than 100 ms or longer than 3 standard deviations above the average per participant (e.g., Steinborn et al., 2010). We then applied a log transformation to remaining RTs to reduce the skewness of the distribution. Only correct trials of old speakers were included in the analysis. RTs were fitted with a linear mixed-effects model (LMM) with the same model structure as for response accuracy. Specifically, prosody (same/different) and content (same/different) served as fixed within-subjects factors, in addition to the between-subjects factor group. Random effects included intercepts and slopes for subject and speaker factors. Post-hoc tests with Bonferroni correction (*emmeans* R library) were conducted when necessary.

2.3.2 Results

2.3.2.1 Encoding

The LMM for the RT slopes (see Methods) revealed a significant effect for the intercept (*b* = -

0.16, SE = 0.04, *t(358)* = 4.40, $p < .001$), representing an overall decrease in RTs for the second presentation of a stimulus, compared to the first one, without a significant difference between groups, $b = 0.05$, SE = 0.04, *t(358)* = 1.43, $p = .15$.

2.3.2.2 Recognition

Response accuracies for all conditions in each group are summarized in Table 2-1. Overall accuracy across all conditions in both groups was significantly above chance level, *Fear* group: M = 0.60, SD = 0.07, $t(29)=8.28$, $p < .001$, Hedges's $g_s = 2.11$; *Neutral* group: M = 0.61, SD=0.09, $t(29) = 6.96$, $p < .001$, Hedges's $g_s = 1.77$, with no significant difference between groups, $t(58) = 0.34$, $p = .74$, Hedges's $g_s = 0.09$.

Table 2-1

Descriptive statistics of recognition accuracy and response bias in Experiment 1

Recognition Performance			***Fear* Group**	***Neutral* Group**
Accuracy	Overall		0.60 (0.07)	0.61 (0.09)
	Same Prosody	Same Content	0.92 (0.11)	0.86 (0.18)
		Different Content	0.72 (0.21)	0.67 (0.23)
	Different Prosody	Same Content	0.43 (0.24)	0.49 (0.35)
		Different Content	0.39 (0.21)	0.49 (0.33)
Response bias	Same Prosody		0.38 (0.17)	0.29 (0.24)
	Different Prosody		-0.15 (0.21)	-0.05 (0.31)

Values are reported in format: Mean (Standard Deviation).

Trial-by-trial response accuracy for old speakers was fitted with a GLMM with prosody (same/different), content (same/different) and group (*Fear/Neutral*) as fixed effects, as well as random intercepts and slopes for subject and speaker effects. Table 2-2 lists the estimated coefficient (*b*), standard error (SE), *z* score and *p* value for all of tested effects. Results showed a significant effect of prosody change ($p < .001$), reflecting a better recognition of old speakers when speech prosody remained the same between encoding and recognition. There was also an interaction between prosody and content ($p < .001$). Post-hoc tests revealed that recognition in

same-prosody trials was better when the content remained the same (SP/SC) than when it changed (SP/DC) (b = -1.41, SE = 0.24, z = 5.97, p < .001), but did not differ significantly as a function of content in different-prosody trials (DP/SC vs. DP/DC: b = -0.11, SE = 0.18, z = 0.62, p = .54) (illustrated in Figure 2-1a). Finally, there was an interaction between prosody and group (p = .037), due to a larger prosody effect in the *Fear* group (*Fear*: b = -2.32, SE = 0.33, z = 7.01, p < .001; *Neutral*: b = -1.49, SE = 0.32, z = 4.62, p < .001).

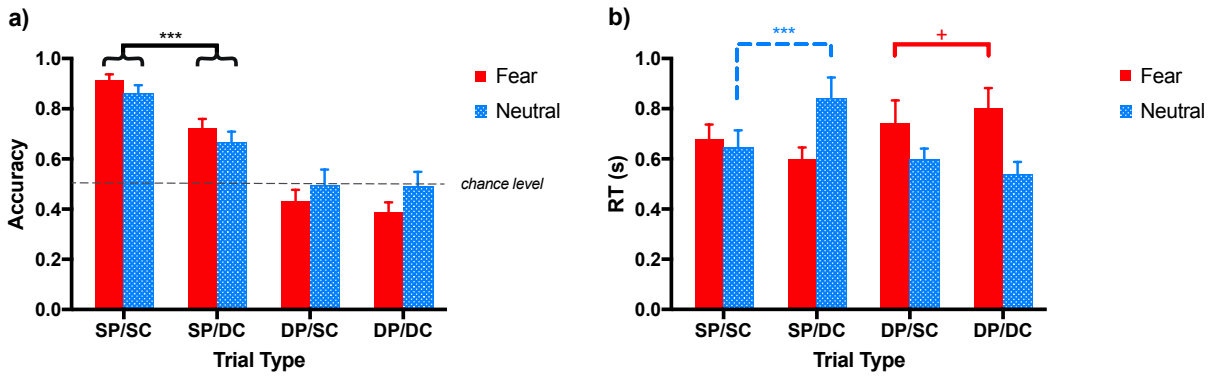

Figure 2-1. Recognition accuracy (a) and response times (b) in Experiment 1. Average (a) accuracy and (b) RTs for each trial type in each participant group. Horizontal lines show the significant differences between conditions in post-hoc tests. Horizontal dashed lines in (a) represents chance-level accuracy. Solid and dashed lines in (b) correspond to significant differences in the *Fear* and *Neutral groups*, respectively. Significance level: ***: p < .001; +: p = .06. (*Abbr:* SP = Same-Prosody; DP = Different-Prosody; SC = Same-Content; DC = Different-Content)

LMM estimation of response bias yielded a significant main effect of prosody (p < .001) and a prosody-by-group interaction (p = .024), as shown in Table 2-2. Post-hoc tests showed that these effects were due to the fact that, whereas both groups showed a significant positive bias (tendency to respond "old") for same-prosody trials (*Fear*: b = 0.38, SE = 0.03, t(57.3) = 11.04, p < .001; *Neutral*: b = 0.29, SE = 0.05, t(57.2) = 5.73, p < .001), only the *Fear* group showed a significant negative bias (tendency to respond "new") for different-prosody trials (*Fear*: b = -0.15, SE = 0.03, t(57.3) = 4.31, p < .001; *Neutral*: b = -0.05, SE = 0.05, t(29) = 0.89, p = .75).

Log-RTs of correct trials for old speakers in the recognition session were analyzed with an LMM with the same structure as the GLMM on response (see Table 2-2). We observed a trend for the main effect of content (p = .063) and a group-by-prosody interaction (p = .003). Post-hoc tests showed that *Fear* group participants responded faster to same-prosody stimuli (b = 0.19, SE

= 0.08, $t(34.7) = 2.41$, $p = .021$), with a trend for the opposite effect in the *Neutral* group (b = -0.15, SE = 0.08, $t(33.8) = 1.89$, $p = .071$). In addition, there was a triple interaction among group, prosody, and content ($p = .042$). Post-hoc tests were followed to disentangle the triple interaction: in the *Fear* group, participants' RTs showed no significant differences as a function of content when the recognition prosody was the same as in encoding (SP/SC vs. SP/DC: b = -0.06, SE = 0.09, $t(48.10) = 0.82$, $p = .41$), but when it was different, participants' response tended to be slower when speech content was also different (DP/SC vs. DP/DC: $b = 0.20$, SE = 0.11, $t(141.40) = 1.89$, $p = .061$). The *Neutral* group, however, displayed an opposite RT pattern: no significant difference from the content change was observed when the recognition prosody changed (DP/SC vs. DP/DC: $b = 0.09$, SE = 0.10, $t(107.50) = 0.87$, $p = .38$), but participants responded faster to same-content stimuli when the recognition prosody remained the same (SP/SC vs. SP/DC: $b = 0.17$, SE = 0.08, $t(51.80) = 2.23$, $p = .030$). A graphical summary of these effects is shown in Figure 2-1b.

Table 2-2

Fixed effects from (G)LMM estimations on recognition response, bias, and log-RTs in Experiment 1

Fixed Effects	*b*	SE	|*t*| or |*z*|	*p*
		Response Accuracy		
Intercept	0.69	0.15	4.72	< .001
Group	0.005	0.12	0.04	.97
Content	-0.38	0.08	4.99	< .001
Prosody	-0.95	0.13	7.37	< .001
Group × Content	-0.06	0.07	0.83	.41
Group × Prosody	-0.21	0.10	2.08	.037
Prosody × Content	0.33	0.07	4.54	< .001
Group × Prosody × Content	0.01	0.07	0.16	.87
		Response Bias		
Intercept	0.12	0.02	5.21	< .001
Group	-0.004	0.02	0.17	.87
Prosody	-0.21	0.02	10.58	< .001
Group × Prosody	-0.05	0.02	2.32	.024
		Recognition log-RT		
Intercept	-0.64	0.05	11.60	< .001

Group	0.02	0.05	0.41	.68
Content	-0.05	0.02	2.04	.063
Prosody	0.01	0.03	0.36	.72
Group x Content	-0.02	0.02	-0.70	.49
Group x Prosody	0.09	0.03	3.13	.003
Prosody x Content	0.02	0.02	1.03	.30
Group x Prosody x Content	0.04	0.02	2.04	.042

GLMM: generalized linear mixed-effects model; RT: response times; SE: standard error.

To assess whether differences in the acoustic parameters of the speech stimuli in the experiment were related to the behavioral effects described above, we examined the relation between the dissimilarity of each speech clip within its own emotional prosody and its overall recognition accuracy via an ANCOVA with emotional prosody as a between factor and average within-prosody Euclidean distance as a covariate. This analysis revealed a significant effect of distance ($F(1,45) = 8.64$, $p = .005$, $\eta_p^2 = .16$). Likewise, we observed a significant relation between stimulus accuracy and its mean distance to the other same-content stimuli ($F(1,45) = 6.34$, $p = .015$, $\eta_p^2 = .12$). That is, the less similar a stimulus was to the others within its own prosody or content group in the acoustic feature space, the more likely it was to be accurately identified as old or new.

2.3.3 Discussion

Results from Experiment 1 indicate that a change in emotional prosody between encoding and recognition had a detrimental impact on voice identity memory accuracy. This observed decline is consistent with prior findings using angry and neutral vocal phrases (Saslove & Yarmey, 1980; Read & Craik, 1995; Stevenage & Neil, 2014). Interestingly, and in agreement with Stevenage & Neil (2014), this recognition impairment was observed regardless of the encoding prosody, although there was a trend for a larger effect when the encoding prosody was fear. Additionally, reduced recognition in same-prosody stimuli was observed when the content changed across both groups, which replicated the results of impairment of voice recognition, from previous studies where speech content being the only experimental manipulation (Zäske et al., 2014, 2017). These results are also in line with previous studies reporting worse performance in speaker identification following changes in various voice properties, such as uttered languages (Wester,

2012; Winters, Levi, & Pisoni, 2008), speech type (i.e., spontaneous or read) (Smith et al., 2018), background noise (Smith et al., 2018), vocalization type (Lavan, Scott, & McGettigan, 2016), and vocalization approach (i.e., sung or spoken words, Peynircioğlu, Rabinovitz, & Repice, 2017). The worse performance for identity memory when prosody or content changed, was likely due, at least in part, to the within-speaker differences in key acoustic parameters as a function of changes in prosody and content (see S.Table 2-1). Indeed, we observed a significant positive correlation between a subject-averaged stimulus-based memory accuracy and its mean distance to the other stimuli in the acoustic parameter multidimensional space, confirming that the more dissimilar a stimulus was to the others in its prosody or content group, the better it could be correctly identified as new or old. This finding is consistent with the significant correlation between perceived speaker distinctiveness and distance-to-mean in the acoustic space reported by Latinus et al. (2013).

The response strategy indicated that both groups of participants shared, as could be expected, a common positive familiarity bias for same-prosody trials (i.e., participants tended to respond "old" to stimuli presented in the same prosody as those in the encoding session), while only subjects from the *Fear* group showed the opposite novelty bias for different-prosody trials (i.e., tendency to categorize neutral stimuli as "new"). The significant familiarity and novelty biases in the *Fear* group presented with fearful and neutral prosody, respectively, suggest that participants based their decisions of whether they had previously heard the speaker mainly on his/her emotional tone, even though they had been explicitly instructed to ignore this feature as irrelevant for the task.

Another measure of memory performance that was less discussed in previous studies is response times. RTs are often considered a proxy of response confidence in a memory test, as they have been shown to correlate strongly with subjective confidence ratings (Robinson, Johnson, & Herndon, 1997). Though they can also reflect or be influenced by task difficulty, effort or strategy (e.g., Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Pesonen, Hämäläinen, & Krause, 2007), there have been suggestions that in a memory recognition test, much of the information from explicit confidence ratings could be obtained in response times (Weidemann & Kahana, 2016). Intriguingly, groups showed opposite RT patterns with regard to same/different prosody between encoding and recognition. From another viewpoint, however, these findings show that both groups displayed a consistent RT pattern with respect to the actual prosody of

recognition stimuli (i.e., fearful vs. neutral), regardless of the prosody presented during encoding: participants were faster in responses to fearful than neutral stimuli, and keeping the same content consistency had a significant facilitating effect only in the case of neutral ones.

Several (non-mutually exclusive) possible explanations can help account for this pattern of response times shared by both groups. First, the facilitated response towards fearfully expressed stimuli may be a result of preferential processing of fearful voices due to their high salience. Emotional faces have been shown to either help (e.g., Phelps, Ling, & Carrasco, 2006; Chadwick et al., 2019) or impede (e.g., Eastwood et al., 2003; Hartikainen et al., 2000) performance in various perception tasks, the former being more likely in difficult tasks (for a discussion, see Chadwick et al., 2019). In our case, voice recognition was a rather difficult task, as evidenced by subjects' accuracy; thus, fearful prosody may have enhanced subjects' attention and/or arousal (e.g., Sutherland & Mather, 2012; Lin, Müller-Bardorff, Gathmann, et al., 2020), leading to a faster processing of those stimuli. Indeed, visual and auditory emotional, particularly fearful, expressions capture attention in an automatic fashion (Armony, Vuilleumier, Driver, & Dolan, 2001; Sanders, Grandjean, & Scherer, 2005) and thus, may lead to a more rapid detection and processing than neutral ones (Öhman & Mineka, 2001). In this context, more attentional resources would have been allocated towards the emotional prosody of the stimuli, and less was left for other characteristics, such as content. In contrast, content information was processed in neutral stimuli without competition from emotional expressions; hence, it contributed to subjects' recognition of previously heard speakers. This interpretation is also in line with the previously reported enhanced memory for the "gist" of emotional events, with no improvement for, or even at the expense of, their details (Christianson & Loftus, 1991; Bookbinder & Brainerd, 2017). Finally, differences in acoustic features between prosodies could have contributed to the observed RT pattern. As the acoustic analysis showed that fearful stimuli were acoustically more distant to each other than neutral ones, it is possible that these larger dissimilarities of fearful stimuli made it implicitly easier for listeners to distinguish speakers. Moreover, given the larger acoustic similarity within neutral prosody samples, any additional information, such as content, would have facilitated recognition of previously encountered speakers, thus resulting in a faster identification of same- than different-content neutral stimuli.

In summary, results from this experiment indicate that changes in speech prosody and content can have a deleterious effect on identity recognition accuracy, as well as an influence on how

participants decided which speakers they had not heard before (response bias). Moreover, response speed on correctly recognized speakers seemed to be dependent on the actual prosody of stimuli and, for neutral stimuli, on content change, in both groups of participants.

2.4 Experiment 2

Accuracy results from Experiment 1 suggest that the presentation of a single exemplar twice is not sufficient for forming a robust representation of an individual's voice that is immune to changes in identity-irrelevant features. In this experiment, we assessed whether increasing the number of exposures to each individual and, critically, the number of exemplars, could help improve voice identity memory performance. Specifically, we employed a within-subjects design in which participants were exposed to four presentations of each unfamiliar speaker. For half of the speakers, the same sentence expressed in fearful prosody (i.e., same stimulus) was always presented, whereas for the other half the samples were all different in terms of prosody (happy or fearful) and/or content ("kids" or "dogs"). In the recognition test, all speakers were presented in a neutral prosody. As mentioned above, we expected participants to exhibit a better voice identity recognition performance when they learned their identity through exposure to different exemplars of the same individual than when they only learned one example, especially when encountering them in a novel prosody (see Lavan et al., 2019c). Moreover, we hypothesized that memory performance for the four-repetition single-exemplar speakers in this experiment would be better than that observed in the *Fear* group of Experiment 1, where each stimulus was presented twice.

2.4.1 Methods

2.4.1.1 Participants

A different cohort of twenty-eight participants (18 female; aged 19 – 37 years) took part in this experiment at the same sites. Recruitment criteria were identical to those in Experiment 1.

2.4.1.2 Stimuli

All 24 speakers from the RAVDESS dataset (Livingstone & Russo, 2018) were used in this experiment. Each speaker uttered two different neutral-content sentences in three prosodies (neutral, strong fear, and strong happiness). The loudness normalization procedure was applied

in the same manner as in Experiment 1.

2.4.1.3 Procedure

The testing setup was the same as in Experiment 1; that is, it consisted of an incidental encoding session followed by a surprise speaker recognition test. During encoding, participants were asked to judge the age range of presented voices (based on pilot data, this task, more effortful than the sex discrimination one used in Experiment 1, improved memory accuracy). For each participant, 6 speakers (half female) were pseudo-randomly assigned to the *Multi* condition, where four distinct exemplars (2 contents x 2 prosodies: fear and happiness) of each speaker were presented once each. The other 6 speakers were assigned to the *Uni* condition, in which only one fearful exemplar per speaker was presented four times. Speech contents were counterbalanced within each condition, and the sequence was pseudorandomized so that the number of intervening trials between presentations of the same speaker were not differently distributed between the *Multi* and *Uni* conditions. As in Experiment 1, the recognition test took place immediately after encoding. Two neutral speech exemplars (2 contents) from each old speaker in both the *Uni* and *Multi* encoding conditions, together with 12 new speakers (2 contents in neutral prosody), were presented. Each exemplar was followed by an old/new judgment question. Response choice and time were recorded for each trial and submitted to subsequent analyses.

2.4.1.4 Data Analysis

We applied the same analysis approaches as used in Experiment 1. For encoding RTs, subject- and speaker-specific regression slopes were analyzed in an LMM, with condition (*Uni/Multi*) as the within-subjects fixed factor and a maximal random effect structure (intercept and slope) of subject and speaker.

Binary recognition responses were fitted in a GLMM with the fixed within-subjects factor of condition (*Uni/Multi*) and by-subject and by-speaker random intercepts and slopes. A single response bias (Br) per subject was calculated to identify an overall response strategy, as there was no sub-condition for new stimuli (i.e., neither *Multi* nor *Uni* condition had corresponding conditions among new-speaker trials). Recognition RTs were cleaned, and log transformed, following the same procedure as in Experiment 1. Log-RTs of correct trials for old speakers were fitted in an LMM with the within-subjects fixed factor condition (*Uni/Multi*).

To test the hypothesis of better memory accuracy when increasing encoding presentation numbers, we conducted a supplementary analysis comparing performance for different-prosody old-speaker trials from the *Fear* group in Experiment 1 (2 presentations of each stimulus) and *Uni* condition trials in Experiment 2 (4 presentations). These response data were fit in a GLMM, with experiment as the between-subjects fixed factor and random effects of subject and speaker.

2.4.2 Results

2.4.2.1 Encoding

Changes in encoding RTs across the four presentations of speakers are illustrated in Figure 2-2. Results from the LMM on RT slopes revealed a significant effect of condition (b = -0.15, SE = 0.07, t(46.12) = 2.31, p = .025), due to smaller slopes for the *Uni* compared to the *Multi* speakers. Post-hoc analyses for each condition separately revealed that the intercept was significantly negative for the *Uni* condition (b = -0.19, SE = 0.06, t(27.00) = -3.21, p = .003), but not the *Multi* condition (b = -0.04, SE = 0.04, t(15.93) = -0.90, p = .38) (regression lines illustrated in Figure 2-2). That is, only the *Uni* trials showed a significant decrease in RTs over repetitions of the same voice identity, which, in this case, consisted of the same stimulus.

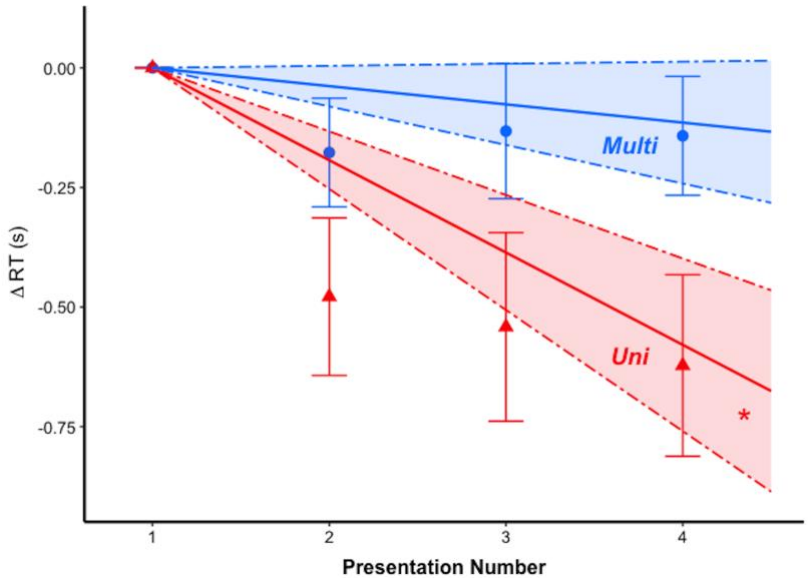

Figure 2-2. Changes in response times (RTs) during encoding in Experiment 2 for the *Uni* (red triangles) and *Multi* (blue circles) conditions (relative to the first presentation). The solid lines represent the subject- and speaker-averaged slopes obtained in the LMMs (see Methods for details). Dashed lines represent ±1SE of the mean slope.

* Slope significantly different from zero (p = .003)

2.4.2.2 Recognition

Response accuracy for each condition (overall, *Multi* and *Uni*) is shown in Table 2-3. The overall accuracy was significantly above chance level (overall: $t(27)$=4.21, $p < .001$, Hedges's g_s = 0.77), as well as both of old-speaker conditions (*Uni*: $t(27) = 2.70$, $p = .009$, Hedges's $g_s = 0.49$; *Multi*: $t(27) = 4.42$, $p < .001$, Hedges's $g_s = 0.81$). The GLMM on trial-by-trial responses for old speakers yielded no significant effect of condition (b = -0.04, SE = 0.22, $z = 0.18$, $p = .86$), suggesting that recognition accuracy of old speakers from the *Uni* and *Multi* conditions did not differ.

Table 2-3

Descriptive statistics of recognition accuracy and response times (RTs) in Experiment 2

Condition	Recognition Accuracy		Response Times (s)	
	M	*SD*	*M*	*SD*
Overall	0.56	0.07	0.75	0.29
Uni	0.59	0.18	0.83	0.38
Multi	0.62	0.14	0.68	0.32

RT: response times; M: mean; SD: standard deviation.

The comparison between the *Fear* group in Experiment 1 and *Uni* condition trials in Experiment 2 yielded a main effect of experiment (b = 1.00, SE = 0.25, $z = 4.05$, $p < .001$), due to a better recognition of speakers with changed prosody when they were presented 4 rather than 2 times during encoding. Moreover, unlike the case of the negative bias in different-prosody trials in Experiment 1, here we did not observed a significant response bias (Br = 0.06, $t(27)$ = 0.38, $p = .71$, Hedges's $g_s = 0.10$).

Recognition RTs with correct responses, shown in Table 2-3, were log-transformed and estimated in an LMM with condition (*Uni/Multi*) as the within-subjects fixed factor. This model revealed a significant effect of condition (b = 0.20, SE = 0.08, $t(664.07) = 2.43$, $p = .015$), which indicated that RTs for correctly recognized speakers previously encoded in the *Uni* condition (i.e., same fearful exemplar presented 4 times) were longer than those from the *Multi* condition

(i.e., four different exemplars varying in prosody and content).

2.4.3 Discussion

During the encoding session, repetition of the same stimulus resulted, as expected, in a linear reduction of response times, as typically shown in most repetition priming experiments (e.g., Bertelson, 1961; Pashler & Baylis, 1991). Interestingly, such a reduction was predominantly present in the *Uni* condition, with a substantially weaker (non-significant) effect for the repeated presentations in the *Multi* condition. Similar effects were found in previous studies: for instance, Manelis et al. (2013) compared the encoding RTs for object pictures in two repetition types (i.e., same-exemplar, resembling the *Uni* condition here; different-exemplar, where two presentation images were not identical but shared the same object gist, resembling the *Multi* condition). Although they observed a main effect of repetition on correctly recollected objects across both same- and different-exemplar conditions, post-hoc tests indicated the effect was driven by same-exemplar trials, with no significant priming for different-exemplar trials. Furthermore, similar attenuations in neural response were also reported in neuroimaging studies. Griffin et al. (2013) reported a neural activity decrease during the second presentation of images, but to a smaller extent in different-exemplar repetition, compared to same-exemplar repetition. This stimulus-specific, rather than individual-specific priming effect could be interpreted as subjects treating new exemplars of a repeated individual as new speakers. However, this seems unlikely, given the results for subsequent recognition RTs (discussed below).

Contrary to our hypothesis, increasing the variability of encoding exemplars did not improve recognition accuracy. Nonetheless, this finding is consistent with some previous studies. For instance, Liu and colleagues (2015) reported a similar lack of significant advantage in face identity recognition when presenting three different expressions over a single one during encoding. Similarly, Lavan et al. (2019c) also failed to find a clear benefit of high variability training in voice identity learning. One possible explanation, also put forward by Liu et al. (2015), is that four presentations of each voice, and without explicit feedback in terms of voice identity during encoding, were still insufficient to form a stable, prosody-invariant identity representation. Interesting, Liu et al. (2015) did observe a benefit of multiple-expression exposure but only when comparing it to a baseline condition containing neutral faces, and this

effect was only apparent when the faces at recognition were of a different expression from those presented at encoding. Thus, the lack of differences between our *Uni* and *Multi* conditions in our study could be due to the fact that in both cases the stimuli presented during encoding had an emotional prosody which, as mentioned in the Discussion of Experiment 1, could have overshadowed any potential small benefit on explicit recognition of multiple-prosody-encoding over single-prosody-encoding.

Despite the lack of a significant difference on identity recognition accuracy between *Multi* and *Uni* conditions, our findings suggest that presenting more than one exemplar of an individual's voice facilitates subsequent speaker's identity recognition, as reflected by the shorter RTs of correctly recognized old speakers from the *Multi* condition. Such reductions in RTs could reflect enhanced confidence (Weidemann & Kahana, 2016) and/or reduced difficulty (Jaeggi, Buschkuehl, Perrig, & Meier, 2010) when correctly identifying previously heard individuals who produced sentences in different emotional expressions and contents. This finding can, in turn, help address the stimulus- vs. individual-specific priming question raised above in the discussion on encoding RTs. That is, during encoding, presentation of new exemplars of a previously presented speaker may have required participants to find the corresponding matching individual among those already heard, resulting in longer RTs, and thus a smaller priming effect (for a similar argument, see Liu et al., 2015). Though we cannot directly determine which process actually took place, participants having shorter RTs when recognizing *Multi*-condition speakers than *Uni*-condition speakers provides evidence for an implicit advantage of multiple exemplar exposure on speaker memory, and therefore supports the latter proposed process. In summary, the RT results are consistent with the hypothesis, mainly established from studies using faces (e.g., Murphy et al., 2015), that exemplar variation may contribute to learning and subsequent recognition of newly familiarized speakers.

Performance in the *Uni* condition in Experiment 2 (4 presentations of each stimulus) was significantly better than that of the *Fear* group in Experiment 1 (2 presentations of each stimulus). This suggests that, as previously shown in both face (e.g., Roark et al., 2006; Murphy et al., 2015) and voice learning (Zäske et al., 2014), increasing the number of presentations of a stimulus improves its recognition. Interestingly, this enhanced memory was observed even if the number of individuals in Experiment 2 was twice that of Experiment 1, which has also been shown to affect memory performance (see Metzger, 2002 for faces). One caveat is that the

encoding tasks in the two experiments were different, and therefore it is possible that the more difficult task of Experiment 2 (age judgment) resulted in a deeper stimulus encoding than the easier task in Experiment 1 (sex judgment), and thus in a better memory performance, independently (for faces, see Bower & Karlin, 1974; Grady et al., 2002; Gur et al., 2002), or in addition to, the larger number of exemplar repetitions.

Taken together, findings from Experiment 2 indicate that speaker recognition across prosody can be improved by simply increasing repetition numbers, and exemplar variance could facilitate subsequent speaker recognition, though not necessarily in terms of explicit recognition accuracy, at least under the experimental setting used here.

2.5 General Discussion

This study investigated the influence of changes in emotional expression (i.e., prosody), content and exemplar variance on subsequent identity recognition of newly familiarized speakers. We examined these factors starting with the simplest scenario where individuals' speech prosody switched between neutral and fear and, orthogonally, content changed or remained the same (Experiment 1). We then extended the focus towards the number and variance of repeated encoding voices (Experiment 2). Whereas research on face memory extensively investigated the influence of within-person variability, from view point and facial expression, to unsystematic variability, using "ambient images" – a wide range of face photos taken in different real-life occasions (e.g., Ritchie & Burton, 2017; Redfern & Benton, 2017a, b, 2019), the majority of literature on voice identity recognition explicitly controlled and minimized most aspects of within-person variability, for example by using highly unified vocal content and tone (reviewed by Lavan et al., 2019b). Here, we took an approach similar to that previously used in studies of face identity recognition (Liu, Chen, & Ward, 2014); namely, we varied specific features of the voice stimuli within speakers (prosody and content), while minimizing other potential confounding factors that could influence memory, by using a well-controlled and validated laboratory-recorded audio-stimulus set.

Results from the two experiments revealed changes in explicit recognition performance (i.e., accuracy) between experimental conditions. Specifically, explicit recognition was impaired under certain experimental manipulations: when exposed to a novel prosody or a novel content at test (Experiment 1), or when the encoding exposure was rather limited and/or the encoding

processing depth was shallow (comparison between the two experiments; see discussion in Experiment 2). Particularly, impaired recognition of previously encountered speakers in Experiment 1 was observed in both *Fear* and *Neutral* groups, reflecting a difficulty in "telling people together" (Lavan et al., 2019a; see Burton, 2013 for faces), when speech exemplars were in a different, rather than same prosody from the one initially encoded. Change in content also interfered with successful recognition of individuals, but only when prosody remained constant. These findings are in line with prior studies using voice line-up (e.g., Saslove & Yarmey, 1980), speaker matching (e.g., Stevenage & Neil, 2014) and the recently developed identity sorting tasks (Lavan et al., 2019a).

However, we did not observe any difference as a function of the prosody presented during encoding (i.e., group effect) on accuracy in Experiment 1, which is consistent with the first two experiments described in Stevenage and Neil's review paper (2014). This was largely due to a common response bias, as participants tended to base their responses on the prosody of the speaker, particularly in the *Fear* group; that is, to categorize fearful voices as previously encountered and those presented with a neutral tone as never heard before. Meanwhile, contrary to our hypothesis, we failed to detect an advantage in memory accuracy, in Experiment 2, for voices that were encoded in two different prosodies (fearful and happy), compared to those encoded in only one (fearful). Speaker familiarity could play a potential role in the absence of such differences. For instance, subjects "told together" familiar speakers better than unfamiliar speakers (Lavan et al., 2019a), with similar findings observed for face identity (Burton et al., 2016). Since participants were only given the same limited amount of exposures to each speaker, a stable representation for each speaker might have been difficult to form, and easily influenced by expression variance. On the other hand, this paradigm helps rule out potential impact on subsequent recognition from another confounding factor, namely the amount of stimulus exposure. As already shown in face studies (e.g., Memon, Hope, & Bull, 2003), and old-speaker recognition performance between Experiment 1 and 2, more or longer exposures of an individual would lead to a better subsequent recognition. The voice sorting paradigm used by Lavan and colleagues did not allow to control the amount of time participants spent on each stimulus, which could have influenced their performance, especially for newly learned speakers.

Although accuracy did not show statistical differences between conditions, other measures of recognition performance, namely response bias (in Experiment 1) and response times (in both

experiments), did display differences between groups (Experiment 1) and presentation conditions (Experiment 2). In Experiment 1, RTs were influenced by stimuli's actual emotional prosody in the two groups, in addition to a content change effect only observed in responses to neutral prosody stimuli. As hypothesized in the discussion of Experiment 1, this RT pattern shared by both groups could be a result of how emotional stimuli are processed. Results from Experiment 2 demonstrated a facilitated response when training with both fearful and happy speech exemplars, rather than only fearful ones, which fits the prediction from exemplar variance advantage (Murphy et al., 2015). Lavan et al. (2019c) tested listeners' recognition performance on manipulating variability of voice stimuli (in a broader sense, not expressiveness variability in particular) and found no clear advantage for vocal identity training with high variability. They proposed that high variability advantage may be seen in situations when listeners are required to generalize to different unheard stimuli. Our results support, to some extent their proposal: although no advantage of recognition towards new unheard stimuli (in a different prosody) was detected, RTs did reflect a facilitation effect for multiple exemplar training. Nonetheless, it is worth pointing out the difference in the nature of the stimulus variability between studies when comparing the results. Whereas the manipulation in Lavan et al. (2019c) was in terms of recording sessions and speaker's speaking styles, ours was focused on prosody and content difference, with other audio settings being consistent (i.e., same recording facilities and spontaneous speaking). Whether such a distinction could account for the fact that we observed significant effects on RTs but not accuracy remains to be determined. Taken together, our findings of differences in RTs and response biases provide complementary insights and extend knowledge towards recognition of newly-familiarized speakers in addition to conventional identity recognition measures such as accuracy. More importantly, it highlights the relevance of these behavioral measures that were less studied in prior experiments, as they may reflect subtle influences of experimental manipulations that target implicit memory, without necessarily influencing explicit recognition accuracy.

In addition, our findings in voice are consistent with the updated facial processing model involving identity and expression processing and integration. There is a long history of research on the topics and in what manner the two processes take place, from the seminal Bruce and Young model (1986) that emphasized a functionally sequential processing manner, where

expression analysis takes place in a dedicated route which is ahead of identity processing via facial recognition unit, to the model proposed by Haxby et al. (2001), which divides facial perception into invariant features like identity, via a ventral temporal route involving the lateral fusiform gyrus and inferior occipital gyrus, and variable properties, including facial expressions, via another anatomical route involving superior temporal sulcus. The recent late bifurcation models (Calder, 2011) were based on these two models to explain integrated facial processing procedures, that both variant and invariant facial features are coded in a shared pathway before visual routes split for further finer processing. As our findings strongly indicated that speech prosody contributes to speaker recognition, they fit with the notion of an interactive mechanism for of vocal identity and vocal expression processing, in line with what the late bifurcation models propose for facial identification.

Lastly, our results showing prominent differences in response speed, which has been reported to exhibit a consistent relation to response confidence, may be relevant to the issue of reliability of earwitness in crime and court testimony. Empirical cases have shown that voice identifications in court can be accurate, but also highly unreliable (Sherrin, 2016). Laboratory studies also show that unfamiliar voice identification tasks are difficult and error-prone, and suffer from low accuracy rates (e.g. Stevenage et al., 2011; Yarmey, 2007). As Sherrin pointed out, it is common for speakers to employ expressive tones of voice during the commission of a crime. Our results of recognition decline due to the change in emotional prosody provide support for his suggestion that earwitnesses could be more reliable when they are exposed to the same tone of voice during the crime scene and the identification process.

2.6 Limitations

Here, we mostly focused on fear when exploring the influence of speech prosody change on identity recognition. This choice was based on previous work by us and others consistently showing an enhanced memory accuracy for emotional facial, vocal and musical expressions (for the same-item effect). While our results suggested similar impairment in voice recognition when the speech prosody changed between fear and neutral, like previous voice studies mostly on anger, parallel face studies have suggested a happy-face advantage (see Liu, Chen, & Ward, 2014). Whether this advantage is emotion- (or valence-) specific, and modality-specific, requires

further investigation. Likewise, more studies that include a wider variety of sentence contents are needed to fully characterize the influence of this factor on speaker identity recognition memory.

Although we interpreted the effects of our experimental manipulations on response speed as reflecting differences in response confidence, in line with an extensive existing literature (e.g., Robinson, Johnson, & Herndon, 1997; Weidemann & Kahana, 2016), we cannot rule out other possibilities, such as task difficulty or cognitive demands. Future studies including explicit measures of these variables should shed light on this issue.

As discussed in Experiment 2, an additional neutral *Uni* condition should help further test and characterize the observed exemplar variance advantage involving emotional expressions. However, increasing the number of conditions (and therefore stimuli) would likely further reduce the already weak memory performance. Further experiments including both within- and between-subject factors could overcome this challenge.

2.7 Conclusion

In summary, our studies offered a novel insight on understanding voice perception and recognition at the early stage of familiarization. Past research has focused largely on explicit recognition of voices and how changes in voices such as emotional prosody, speech content and exposure amount influence identity perception. Here we integrated these changes orthogonally in the experiments, and extended the behavioral repertoire measured to include response bias and response times. Our results indicated that the influence of these explicit and implicit recognition indices could be different, thus highlighting the usefulness of including behavioral measures other than response accuracy in future voice, and possibly face, identity memory or perception studies.

2.8 Acknowledgements

We are thankful to Dr. Signy Sheldon for insightful comments and suggestions. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC, 2017-05832) and the Canadian Institutes of Health Research (CIHR, MOP-130516), to JLA. HX received a CRBLM Graduate Student Stipend.

2.9 Conflict of interest

Authors declare no conflicts of interest.

2.10 References

Armony, J. L. (2013). Current emotion research in behavioral neuroscience: The role(s) of the amygdala. *Emotion Review, 5*(1), 104-115. http://doi.org/10.1177/1754073912457208

Armony, J. L., Chochol, C., Fecteau, S., & Belin, P. (2007). Laugh (or Cry) and You will be Remembered: Influence of Emotional Expression on Memory for Vocalizations. *Psychological Science, 18*(12), 1027–1029. https://doi.org/10.1111/j.1467-9280.2007.02019.x

Armony, J. L., Vuilleumier, P., Drive, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron, 30*(3), 829-841. https://doi.org/10.1016/S0896-6273(01)00328-2

Aubé, W., Peretz, I., & Armony, J. L. (2013). The effects of emotion on memory for music and vocalizations. *Memory, 21*(8), 981-990. https://doi.org/10.1080/09658211.2013.770871

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. http://dx.doi.org/10.1016/j.jml2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol. Res, 74*, 110–120. https://doi.org/10.1007/s00426-008-0185-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Benson, P. J., & Perrett, D. I. (1993). Extracting prototypical facial images from exemplars. *Perception*, *22*(3), 257-262. https://doi.org/10.1068/p220257

Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. *Quarterly Journal of Experimental Psychology, 13*(2), 90-102.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.04, retrieved 28 September 2019 from http://www.praat.org/

Bookbinder, S. H., Brainerd, C. J. (2017). Emotionally negative pictures enhance gist memory. *Emotion, 17*(1): 102–119. https://doi.org/10.1037/emo0000171

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology, 103*(4), 751–757. https://doi.org/10.1037/h0037190

Bruce, V., Herderson, Z., Newman, C., & Burton, A. M. (2011). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*, 207-218.

Bruce, V., Young, A. (1986). Understanding face recognition. *The British Psychological Society, 77,* 305-327.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*(8), 1467-85. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., Jenkins R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*, 202-23. https://doi.org/10.1111/cogs.12231

Burton, A. M., & Jenkins, R. (2011). Unfamiliar face perception. *The Oxford Handbook of Face Perception*, *28*, 287-306.

Calder, A. J. (2011). Oxford Handbook of Face Perception, Chapter 22: Does facial identity and facial expression recognition involve separate visual routes? (Calder, A. J., Rhodes, G., Johnson, M. H., & Haxby, J. V., Ed.). *Oxford University Press*, ISBN: 978-0-19-955905-3 (pp, 427-48).

Chadwick, M., Metzler, H., Tijus, C., Armony, J. L., & Grèzes, J. (2019). Stimulus and observer characteristics jointly determine the relevance of threatening facial expressions and their interaction with attention. *Motivation and Emotion*, *43*(2), 299-312. https://doi.org/10.1007/s11031-018-9730-2

Chhabra, S., Badcock, J. C., Maybery, M. T., & Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia, 50*, 2730–2735. https://doi.org/10.1016/j.neuropsychologia.2012.08.006

Christianson, S. A., & Loftus, E. F. (1991). Remembering emotional events: The fate of detailed information. *Cognition & Emotion*, *5*(2), 81–108.

https://doi.org/10.1080/02699939108411027

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*(1), 97-116. https://doi.org/10.1080/09541440340000439

Eastwood, J. D., Smilek, D., & Merikle, P. M. (2003). Negative facial expression captures attention and disrupts performance. *Perception & Psychophysics, 65*(3), 352–358. https://doi.org/10.3758/BF03194566

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fecteau, S., Berlin, P., Joanette, Y., & Armony, J. L. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage, 36*(2), 480-487. https://doi.org/10.1016/j.neuroimage.2007.02.043

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Curr. Biol, 18*, 457–460. https://doi.org/10.1016/j.cub.2008.03.030

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accent in communication. *Personality and Social Psychology Review, 14*(2), 214-237. https://doi.org/ 10.1177/1088868309359288

Goshen-Gottstein, Y., & Ganel, T. (2000). Repetition priming for familiar and unfamiliar faces in a sex-judgment task: Evidence for a common route for the processing of sex and identity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1198–1214. https://doi.org/10.1037/0278-7393.26.5.1198

Grady, C. L., Bernstein, L. J., Beig, S., & Siegenthaler, A. L. (2002). The effects of encoding task on age-related differences in the functional neuroanatomy of face memory. *Psychology and Aging, 17*(1), 7–23. https://doi.org/10.1037/0882-7974.17.1.7

Griffin, M., DeWolf, M., Keinath, A., Liu, X., & Reder, L. (2013). Identical versus conceptual repetition FN400 and parietal old/new ERP components occur during encoding and predict subsequent memory. *Brain Res, 1512*, 68-77. https://doi.org/10.1016/j.brainres.2013.03.014

Gur, R. C., Schroeder, L., Turner, T., McGrath, C., Chan, R. M., et al. (2002). Brain activation during facial emotion processing. *NeuroImage, 16*(3A), 651-62. https://doi.org/10.1006/nimg.2002.1097

Hartikainen, K. M., Ogawa, K. H., & Knight, R. T. (2000). Transient interference of right hemispheric function due to automatic emotional processing. *Neuropsychologia*, *38*(12), 1576-1580. https://doi.org/10.1016/S0028-3932(00)00072-5

Haxby, J. V., Gobbini, M. I., Furey, M. L., et al. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-30. https://doi.org/10.1126/science.1063736

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier B. (2010). The concurrent validity of the N-back task as a working memory measure. Memory, 18(4), 394-412. https://doi.org/10.1080/09658211003702171

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Phil. Trans. R. Soc. B., 366*, 1671-83. https://doi.org/10.1098/rstb.2010.0379

Kaufmann, J. M., Schweinberger, S. R., & Burton, A. M. (2009). N250 ERP correlates of the acquisition of face representations across different images. *Journal of Cognitive Neuroscience, 21*(4), 625-641. https://doi.org/10.1162/jocn.2009.21080

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, *15*(4), 241-252. https://doi.org/10.1515/REVNEURO.2004.15.4.241

Kensinger, E. A., & Schacter, D. L. (2005). Retrieving accurate and distorted memories: Neuroimaging evidence for effects of emotion. *Neuroimage, 27*(1), 167-177. https://doi.org/10.1016/j.neuroimage.2005.03.038

Kim, Y., Sidtis, J. J., & Sidtis, D. V. (2019). Emotionally expressed voices are retained in memory following a single exposure. *PLoS ONE, 14*(10). https://doi.org/10.1371/journal.pone.0223948

Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., & Honda, K. (2006). Cyclicity of laryngeal cavity resonance due to vocal fold vibration. *J. Acoust. Soc. Am, 120*, 2239–2249. https://doi.org/10.1121/1.2335428

Kuhn, M. (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

LaBar, K., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nat Rev Neurosci, 7***,** 54–64. https://doi.org/10.1038/nrn1825

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology, 23*(12), 1075-1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Burton, A. M., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019a). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240-48. https://doi.org/10.1177/1747021819836890

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019b). Flexible voices: Identity perception from variable vocal signals**.** *Psychon Bull Rev, 26*, 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019c). The effects of high variability training on voice identity learning. *Cognition, 193*, 104026. https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General, 145*(12), 1604–1614. https://doi.org/10.1037/xge0000223

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(2), 298–303. https://doi.org/10.1037/0278-7393.10.2.298

Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Lin, H., Müller-Bardorff, M., Gathmann, B. *et al.* (2020). Stimulus arousal drives amygdalar responses to emotional expressions across sensory modalities. *Science Report, 10,* 1898. https://doi.org/10.1038/s41598-020-58839-1

Liu, C. H., Chen, W. F., & Ward, J. (2014). Remembering faces with emotional expressions. *Front Psychol*, *5*, 1439. https://doi.org/10.3389/fpsyg.2014.01439.

Liu, C. H., Chen, W. F., & Ward, J. (2015). Effects of exposure to facial expression variation in face learning and recognition. *Psychological Research, 79*(6), 1042-53. https://doi.org/10.1007/s00426-014-0627-8

Livingstone, S.R., & Russo, F.A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in

North American English. *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77–100. https://doi.org/10.1037/0096-1523.34.1.77

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149-157. https://doi.org/10.1037/0278-7393

Manelis, A., Paynter, C. A., Wheeler, M. E., & Reder, L. M. (2013). Repetition related changes in activation and functional connectivity in hippocampus predict subsequent memory. *Hippocampus, 23*(1), 53-65. https://doi.org/10.1002/hipo.22053

Martin, D., Cairns, S. A., Orme, E., DeBruine, L. M., Jones, B. C., & Macrae, C. N. (2010). *Experimental Psychology, 57*(5), 338-345. https://doi.org/10.1027/1618-3169/a000040

Martin, D., & Greer, J. (2011). Getting to know you: from view-dependent to view-invariant repetition priming for unfamiliar faces. *The Quarterly Journal of Experimental Psychology, 64*(2), 217-223. https://doi.org/10.1080/17470218.2010.541266

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, *21*(5), 428-436. https://doi.org/10.1109/TAU.1973.1162507

Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: effects on eyewitness accuracy and confidence. *Br J Psychol, 94*(3), 339-54. https://doi.org/10.1348/000712603767876262

Metzger, M. M. (2002). Stimulus load and age effects in face recognition: A comparison of children and adults. *North American Journal of Psychology, 4*(1), 51–62.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-581. https://doi.org/10.1037/xhp0000049

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness memory. *Journal of Investigative Psychology and Offender Profiling, 10*(1), 57–70. https://doi.org/10.1002/jip.1381

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*(3), 483–522. https://doi.org/10.1037/0033-295X.108.3.483

Pashler, H., & Baylis, G. (1991). Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 33-48.

Pesonen, M., Hämäläinen, H., & Krause, C. M. (2007). Brain oscillatory 4-30 Hz responses during a visual n-back memory task with varying memory load. Brain Research, 1138, 171-177. https://doi.org/10.1016/j.brainres.2006.12.076

Peynircioğlu, Z. F., Rabinovitz B. E., & Repice J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice, 31*(2), 256.e13-17. https://doi.org/10.1016/j.jvoice.2016.06.004

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science*, *17*(4), 292-299.

Pichora-Fuller, M.K., Dupuis, K., & Smith, L. (2016). Effects of vocal emotion on memory in younger and older adults. *Experimental Aging Research, 42*(1), 14-30. https://doi.org/10.1080/0361073X.2016.1108734.

Pichora-Fuller, M. K., Dupuis, K., & van Lieshout, P. (2016). Importance of F0 for predicting vocal emotion categorization. *J. Acoust. Soc. Am., 140*(4), 3401-3401. https://doi.org/10.1121/1.4970917

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*(1), 6–18. https://doi.org/10.1037/1076-898X.1.1.6

Redfern, A. S., & Burton, C P. (2017a). Expressive faces confuse identity. *I-Perception, 8*(5), 1-21. https://doi.org/1177/2041669517731115

Redfern, A. S., & Burton, C. P. (2017b). Expression dependency in the perception of facial identity. *I-Perception, 8*(3), 1-15. https://doi.org/10.1177/2041669517710663

Redfern, A. S., & Burton, C. P. (2019). Representation of facial identity includes expression variability. *Vision Research, 157*, 123-131. https://doi.org/10.1016/j.visres.2018.05.004

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *J. Exp. Psychol. Hum. Percept. Perform., 23*(3), 651–666. Doi: 10.1037/0096-1523.23.3.651

Righi, S., Marzi, T., Toscani, M., Baldassi, S., Ottonello, S., & Viggiano, M. P. (2012). Fearful expressions enhance recognition memory: Electrophysiological evidence. *Acta Psychologica, 139*(1), 7-18. https://doi.org/10.1016/j.actpsy.2011.09.015

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 879-895. https://doi.org/10.1080/17470218.2015.1136656

Roark, D. A., O'Toole, A. J., Abdi, H., & Barrett, S. E. (2006). Learning the moves: The effect of familiarity and facial motion on person recognition across large changes in viewing format. *Perception, 35*(6), 761–773. https://doi.org/10.1068/p5503

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*(3), 416–425. https://doi.org/10.1037/0021-9010.82.3.416

Sanders, D., Grandjean, D., Pourtois, G., et al. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *Neuroimage, 28*(4), 848-58. https://doi.org/10.1016/j.neuroimage.2005.06.023

Sangha, S., Diehl, M. M., Bergstrom, H. C., & Drew, M. R. (2020). Know safety, no fear. *Neuroscience & Biobehavioral Reviews, 108*, 218-30. http://doi.org/10.1016/j.neubiorev.2019.11.006

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65*(1), 111-6. https://doi.org/10.1037/0021-9010.65.1.111

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews-Cognitive Science, 5*(1), 15-25. https://doi.org/10.1002/wcs.1261

Sergerie, K., Lepage, M., & Armony, J. L. (2005). A face to remember: emotional expression modulates prefrontal activity during memory formation. *Neuroimage, 24*(2), 580-5. https://doi.org/10.1016/j.neuroimage.2004.08.051

Serge**r**ie, K., Lepage, M., & Armony, J. L. (2007). Influence of emotional expression on memory recognition bias: a functional Magnetic Resonance Imaging study. *Biological Psychiatry, 62*(10), 1126-33. https://doi.org/10.1016/j.biopsych.2006.12.024

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1447–1469. https://doi.org/10.1037/0096-1523.28.6.1447

Sherrin, C. (2016). Earwitness Evidence: The Reliability of Voice Identifications. *Osgoode Hall Law Journal, 52*(3), 819-862.

Smith, H. M. J., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2018). Forensic voice discrimination: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272-287. https://doi.org/10.1002/acp.3478

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *J Exp Psychol Gen, 117*(1), 34–50. https://doi.org/10.1037//0096-3445.117.1.34

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2010). Differential effects of prolonged work on performance measures in self-paced speed tests. *Advances in cognitive psychology*, *5*, 105–113. https://doi.org/10.2478/v10053-008-0070-8

Stevenage, S. V., Howland, A., & Tipplet, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology, 25*(1), 112-8. https://doi.org/10.1002/acp.1649

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*(3), 266-281. http://dx.doi.org/10.5334/pb.ar

Sutherland, M. R., & Mather, M. (2012). Negative arousal amplifies the effects of saliency in short-term memory. *Emotion, 12*(6), 1367-1372. https://doi.org/10.1037/a0027860

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., & Honda, K. (2006). Acoustic roles of the laryngeal cavity in vocal tract resonance. *J. Acoust. Soc. Am, 120*, 2228–38. https://doi.org/10.1121/1.2261270.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science, 3*(4), 150670. https://doi.org/10.1098/rsos.150670

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*(6), 781-790. https://doi.org/10.1016/j.specom.2012.01.006

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual

talkers across languages. *The Journal of the Acoustical Society of America, 123*(6), 4524-38. https://doi.org/10.1121/1.2913046

Xu, C. (2017). The Effects of Response and Stimulus Repetition across Sequences of Trials in Go/No-go Tasks. Thesis at the University of Iowa.

Xu, M., Homae, F., Hashimoto, R., & Hagiwara H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Front. Psychol., 4*, 735. https://doi.org/10.3389/fpsyg.2013.00735

Yarmey, D. (2007). The Handbook of Eyewitness Psychology, Volume II: Memory for People - The Psychology of Speaker Identification and Earwitness Memory (Lindsay, R. C. L., et al, Ed.). *Mahwah, New Jersey: Lawrence Erlbaum Associates*, pp. 101-102.

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and Voice Perception: Understanding Commonalities and Differences. *Trends in Cognitive Sciences, 24*(5), 398-410. https://doi.org/10.1016/j.tics.2020.02.001

Zäske, R., Hasan, B. A. S., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex, 94*, 100-112. https://doi.org/10.1016/j.cortex.2017.06.005

Zäske, R., Volberg, G., Kovács, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *Journal of Neuroscience, 34*(33), 10821-31. https://doi.org/10.1523/JNEUROSCI.0581-14.201

2.11 Supplementary Materials

S.Table 2-1

Descriptive statistics and repeated-measures ANOVAs results of acoustic parameters for the stimuli used in *Experiment 1*

Acoustic Parameter	Descriptive Stats				Prosody Effect			Sentence Effect		
	Neutral	**Fear**	**"Kids"**	**"Dogs"**	F(1,11)	p	η_p^2	F(1,11)	p	η_p^2
Speech duration (s)	1.65 (0.22)	1.61 (0.18)	1.63 (0.21)	1.63 (0.20)	0.23	.64	.02	0.05	.83	.004
Min F0 (semitone)	0.63 (6.93)	9.79 (6.65)	5.69 (8.54)	4.73 (7.91)	14.64	.003*	.57	0.92	.36	.08
Max F0 (semitone)	14.63 (5.45)	22.02 (6.40)	18.51 (7.09)	18.14 (6.99)	20.16	<.001*	.65	0.19	.67	.02
Range F0 (semitone)	14.00 (8.29)	12.23 (5.57)	12.81 (7.82)	13.41 (6.34)	0.42	.53	.04	0.19	.67	.02
M F0 (semitone)	8.38 (5.38)	16.48 (5.61)	12.29 (6.98)	12.57 (6.77)	75.70	<.001*	.87	0.58	.46	.05
SD F0 (semitone)	3.20 (1.15)	2.64 (1.11)	2.73 (1.08)	3.11 (1.21)	2.15	.17	.16	3.96	.07	.26
M F1 (semitone)	628.76 (42.82)	732.38 (90.39)	663.76 (83.55)	697.38 (89.64)	24.71	<.001*	.69	3.95	.07	.26
SD F1 (semitone)	349.05 (84.55)	367.30 (121.16)	334.59 (87.62)	381.75 (114.74)	0.89	.37	.08	9.36	.01	.46
M F2 (semitone)	1726.22 (87.74)	1809.16 (94.03)	1754.23 (86.80)	1781.16 (110.53)	12.46	.005*	.53	3.74	.08	.25
SD F2 (semitone)	475.32 (61.46)	472.53 (96.70)	465.59 (74.02)	482.25 (86.65)	0.01	.91	.001	1.09	.32	.09
M F3 (semitone)	2715.05 (86.40)	2812.82 (107.05)	2727.78 (100.94)	2800.09 (104.86)	6.64	.03	.38	30.31	<.001*	.73
SD F3 (semitone)	507.41 (73.61)	464.32 (106.79)	453.76 (79.50)	517.97 (96.55)	3.24	.10	.23	11.69	.006*	.51

M F4 (semitone)	3838.46 (142.55)	3861.21 (128.66)	3838.56 (134.30)	3861.11 (137.26)	0.19	.67	.02	4.91	.05	.31
SD F4 (semitone)	455.19 (60.70)	460.53 (132.59)	433.20 (100.81)	482.53 (99.23)	0.02	.88	.002	19.78	<.001*	.64
M Amplitude (dB)	69.40 (1.79)	69.97 (1.73)	69.85 (1.59)	69.52 (1.95)	1.36	.27	.11	1.21	.29	.10
SD Amplitude (dB)	7.45 (1.30)	7.94 (1.43)	8.12 (1.44)	7.26 (1.19)	1.11	.31	.09	25.7	<.001*	.70
Median Amplitude (dB)	66.45 (2.79)	66.33 (1.75)	66.24 (2.31)	66.54 (2.34)	0.03	.87	.002	0.39	.55	.03

Abbr: Min/Max: minimum or maximum values of the corresponding parameter. M: mean of the corresponding parameter. SD: standard deviation of the corresponding parameter.

* $p < .05$ after the Benjamini-Hochberg correction of False Discovery Rate to account for multiple comparisons. None of the prosody-by-content interactions was statistically significant ($p \geq .90$, FDR corrected)

Connecting Chapters 2 to 3

In Chapter 2, we demonstrated an exemplar variance advantage in recognition speed, not in explicit recognition accuracy in Experiment 2. One argument on the lack of recognition accuracy difference we offered in the discussion was that, four exemplars per speaker with two distinct emotional expression in the encoding stage might not be sufficient for listeners to form a stable speaker representation. In other words, encoding variance may not be large enough. Indeed, previous face studies that have reported a strong advantage of learning variance (e.g., Murphy et al., 2015; Ritchie & Burton, 2017), employed ambient images, which incorporated a large amount of within-person variance. Another question we briefly mentioned in Chapter 2 was, whether the observed advantage can be simply due to certain stimuli that appeared only in the *Multi* condition. Specifically, in the paradigm, happy stimuli were only present in the *Multi* condition, and there lies the possibility that the advantage was simply because of a better learning from happy speech clips. With these unanswered questions in mind, we conducted Study 2 (Chapter 3) to clarify and expand previous findings with three modifications on the experimental design. Firstly, we removed speech content variance, to focus only on emotional expression, and to continue the original research question of the exemplar variance advantage in explicit recognition performance when more emotional variance was encoded. Secondly, we added a between-subjects factor into the design, to allow the single emotion in the *Uni* condition to alter between participants, in order to clarify (or rule out) the possibility that any exemplar variance effects can be driven by specific emotional exemplars. Lastly, we extended the same paradigm from voice to face. With the revised paradigm, Study 2 was able to continue examining the hypothesized emotional-exemplar variance advantage for identity memory in both modalities, and furthermore, to test if the effect is affected by specific emotion categories or emotion-relevant features, such as arousal or valence.

Chapter 3. Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition

(Study 2)

Hanjian Xu[1,2,3], Jorge L. Armony[1,2,4]

[1]Douglas Mental Health University Institute, Verdun, Canada;
[2]BRAMS Laboratory, Centre for Research on Brain, Language and Music, Montreal, Canada;
[3]Integrated Program in Neuroscience, McGill University, Montreal, Canada;
[4]Department of Psychiatry, McGill University, Montreal, Canada

Under review at *Motivation and Emotion*

3.1 Abstract

Emotional stimuli and events are better and more easily remembered than neutral ones. However, this advantage appears to come at a cost, namely a decreased accuracy for peripheral, emotion-irrelevant details. There is some evidence, particularly in the visual modality, that this trade-off also applies to emotional expressions, leading to a difficulty in identifying an unfamiliar individual's identity when presented with an expression different from the one encountered at encoding. On the other hand, past research also suggests that identity recognition memory benefits from exposure to different encoding exemplars, although whether this is also the case for emotional expressions, particularly voices, remains unknown. Here, we directly addressed these questions by conducting a series of voice and face identity memory online studies, using a within-subject old/new recognition test in separate unimodal modules. In the Main Study, half of the identities were encoded with four presentations of one single expression (angry, fearful, happy, or sad; *Uni* condition) and the other half with one presentation of each emotion (*Multi* condition); all identities, intermixed with an equal number of new ones, were presented with a neutral expression in a subsequent recognition test. Participants (N=547, 481 female) were randomly assigned to one of four groups in which a different *Uni* single emotion was used. Results, using linear mixed models on response choice and drift-diffusion-model parameters, revealed that high-arousal expressions interfered with emotion-independent identity recognition accuracy, but that such deficit could be compensated by presenting the same individual with various expressions (i.e., high exemplar variability). These findings were confirmed by a significant correlation between memory performance and stimulus arousal, across modalities and emotions, and by two follow-up studies (Study 1: N = 172, 150 female; Study 2: N=174, 154 female), which extended the original observations and ruled out some potential confounding effects. Taken together, the findings reported here expand and refine our current knowledge of the influence of emotion on memory, and highlight the importance of, and interaction between, exemplar variability and emotional arousal in identity recognition memory.

Keywords: Emotional Expression; Arousal; Exemplar Variance; Face Recognition; Voice Recognition

3.2 Introduction

Most people will immediately recognize the woman (and her scream) in Hitchcock's *Psycho* shower murder scene, even those who didn't see the actual movie. Consistent with this anecdotal observation, a large body of research has showed a superior memory recognition accuracy for emotionally expressed faces (e.g., Sergerie, Lepage, & Armony, 2005; LaBar & Cabeza, 2006) and voices (e.g., Armony, Chochol, Fecteau, & Belin, 2007; Aubé, Peretz, & Armony, 2013; Pichora-Fuller, Dupuis, & Smith, 2016), compared to neutral ones. This emotional memory advantage has also been reported for other emotional stimuli, such as objects, scenes, and events (e.g., Cahill et al., 1994; Kensinger, 2004; Kensinger & Schacter, 2005; Righi et al., 2012). Although the exact mechanisms of this emotional enhancement of episodic memory are still unclear, emotional arousal has been regarded as an important factor in influencing emotional memory (e.g., Mather & Sutherland, 2011; Ack Baraly, Hot, Davidson, & Talmi, 2016; Cahill et al., 1994). When an experience triggers an arousal response, memory formation stages — including encoding, consolidation, and retrieval — can be affected, possibly by amygdala-mediated processes (McGaugh, 2004; Qasim et al., 2023), increasing the likelihood of the experience being remembered (for a review, see Kensinger, 2009).

However, this emotion-mediated memory enhancement may come at a cost, namely decreased memory accuracy for emotionally-irrelevant or "extrinsic" aspects of the stimulus (that is, spatially, temporally, or conceptually distinct from the emotion-specific information; Kensinger, 2009). This phenomenon, sometimes referred to as the central-peripheral trade-off (Kensinger, 2007; Mather & Sutherland, 2011), is particularly pronounced for high arousal events, possibly due to a narrowing of the attention focus (Kensinger, 2009). Indeed, stimuli signaling threat are processed in a more automatic fashion (Armony, Vuilleumier, Driver, & Dolan, 2001; Sander et al., 2005), likely capturing attention (Mogg & Bradley, 1999; Dolcos et al., 2020), and thus resulting in a more effective encoding process (Talmi et al., 2008). This may reflect the evolutionary advantage prioritizing survival over other goals and motivations (e.g., Unkelbach, Alves, & Koch, 2020; Norris, 2021; Rozin & Royzman, 2001).

There is some evidence suggesting that the proposed trade-off also applies to emotional expressions; in that case, the memory enhancement for the previously encountered stimulus would be associated with a decrease in the recognition of the (expression-independent) individual's identity. Coming back to the *Psycho* example, how many of those who rapidly and

without hesitation recognize the shower scene would be able to identify the actress (Janet Leigh) if they saw or heard her in another movie? Whereas this question has been less investigated, some studies have shown a decrease in identity recognition of individuals, especially for unfamiliar ones, when the emotional expression changes between encoding and retrieval (see Bruce, 1982, Liu, Chen & Ward, 2014, Nomi et al., 2013 for faces; Saslove & Yarmey, 1980, Stevenage & Neil, 2014, Xu & Armony, 2021 for voices).

On the other hand, a number of identity representation models, largely based on face perception — such as the averaging/prototype (e.g., Benson & Perrett, 1993; Burton, Jenkins, Hancock, & White, 2005) and pictorial coding models (e.g., Longmore, Liu & Young, 2008) —, collectively propose that stimulus-independent person recognition may benefit from exposure to within-person exemplar variance (i.e., different instances of the same individual along a given feature dimension, such as view-point for faces or speaking style for voices), which allows for the formation of a more stable identity representation, or a greater likelihood of a match between a novel encounter and previously stored instances. Studies using uncontrolled within-person exemplar variance in learning of faces (ambient face images, e.g., Murphy et al., 2015; Ritchie & Burton, 2017; Matthews, Davis, & Mondloch, 2018) and voices (Lavan et al., 2019a) provide support for this proposal.

Interestingly, there is some empirical data, though somewhat limited and not always consistent, suggesting a similar benefit after encoding exemplars in multiple emotional expressions of the same individual. Liu and colleagues (2015) examined this hypothesized advantage using static face images, by comparing face recognition in a novel emotional expression after they were encoded with a single exemplar (i.e., only one emotional expression) or with multiple exemplars (i.e., several different emotional expressions). Though the expression-independent face recognition was overall above-chance, the advantage of encoding multiple emotional expressions was only significant when compared to identities encoded with a single neutral expression. In their follow-up work using both static images and dynamic face videos (Liu et al., 2016), they found that exposure to single or multiple dynamic expressions led to a similar identity recognition presented with a new expression. In contrast, exposure to static faces resulted in a significantly weaker recognition of the new-expression identities. This is not surprising as research has shown that people are more efficient in learning and recognizing dynamic faces (e.g., O'Toole, Roark & Abdi, 2002; Xiao et al., 2014). One noticeable feature in

both studies is that the emotions in the exemplars encoded in both multiple-expression and single-expression conditions were randomly chosen from six basic emotions, rather than using fixed ones for each participant. Whereas this could provide more generalizable conclusions that are not emotion-specific, it may have also obscured possible emotion-dependent influences, which could have contributed to the small magnitude of the effects observed. Nonetheless, these results collectively point to an emotional expression advantage for face identity memory, albeit with a weaker magnitude compared to the effect obtained with ambient images.

The relation between emotional expressions and identity memory has been less studied for voices. Consistent with the central-peripheral trade-off hypothesis, Lavan and colleagues (2019b) found that learning high expressive unfamiliar voices interfered with identity categorization in a sorting task, compared to learning low expressive voices. Regarding exemplar variability, a recent study reported an advantage in recognition speed, though not in recognition accuracy, for speakers who were encoded with two (fearful and happy), compared to one (fearful), prosodies (Xu & Armony, 2021).

In summary, emotional expressions, or at least some of them, seem to influence subsequent memory: single expressions lead to an enhancement of same-identity/same-emotion (i.e., same stimulus) recognition, but a reduction of same-identity/different-emotion accuracy. However, exposure to multiple emotional expressions may overcome the latter deficit, at least to some extent, by helping form a stable emotion-independent identity representation. Nonetheless, the putative mnemonic advantage of multiple emotional within-person exemplars in both face and voice has not been thoroughly tested. The limited number of studies, as mentioned above, did provide, mostly indirect, and sometimes conflicting, support for this multiple exemplar advantage. Moreover, it remains unclear whether such effects are specific to certain categories of emotion, such as high arousing ones (e.g., fear), or are generalizable to others. This is relevant as not every expression poses the same level of salience or recruits the same level of attentional resources (e.g., Lundqvist, Bruce & Öhman, 2015; Vuilleumier & Huang, 2009).

Here, we report three studies to directly test whether presentation of multiple emotions during encoding leads to a better identity recognition, compared to the repeated presentation of a single emotion and, if so, whether this effect depends on the specific emotions presented (see Table 3-1 for the expression combinations used in each study). We did so through a recognition memory task in which subjects encoded identities through speech clips or face images (in separate

modules). Half of the identities were presented with four different emotions (*Multi* condition), while for the other half, a single emotion was presented four times (*Uni* condition). The specific emotion for the *Uni* condition was varied across subjects. Two additional studies were subsequently conducted to further clarify and help interpret the findings obtained in the main study. Choice (Older/Younger and New/Old during encoding and recognition, respectively; see Methods) and the corresponding response times (RTs), as well as response confidence during recognition, were collected. We analyzed recognition accuracy and an integrated accuracy-speed index using drift diffusion models, as described in the General Methods. We hypothesized that participants would recognize identities encoded through multiple expressions better than those with a single exemplar. Moreover, we expected this memory advantage to be influenced by the specific emotions presented, either in terms of threat-signaling relevance (i.e., anger and fear) or stimulus arousal.

Table 3-1

Overview of the emotional expression combinations used in the studies.

Study	**Encoding**		**Recognition**
	Uni	*Multi*	
Main Study (N=547)	angry / fearful / happy / sad *(between subjects)*	angry, fearful, happy, sad	neutral
Follow-up Study 1 (N=172)	1.1 fearful, sad	-	neutral
	1.2 fearful, neutral	-	happy
Follow-up Study 2 (N=174)	angry / fearful *(between subjects)*	angry, fearful, happy	neutral

3.3 General Methods

3.3.1 Participants

Young healthy adults were recruited from the McGill Psychology extra-credit participant pool and from a McGill University community (students and alumni) Facebook group. Participants completed the studies online on a JATOS server (Lange, Kühn, & Filevich, 2015), hosted by the International Laboratory for Brain, Music, and Sound Research (BRAMS), using the jsPsych library (de Leeuw, 2015). They were all fluent in English, had normal hearing and (corrected-to-)

normal vision, reported no prior diagnosis or treatment of psychiatric/neurological conditions, and received course credits or monetary compensation (CAD10). Each subject took part in only one of the studies. The studies were approved by the Faculty of Medicine Research Ethics Office at McGill University.

3.3.2 Stimuli

Vocal stimuli were audio-only recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018). The recordings consisted of 24 speakers (12 female, aged 21-33) uttering a semantically neutral sentence ("Kids are talking by the door"), in four emotional (fear, happiness, sadness, and anger, the strongly expressive version) and neutral expressions (i.e., prosody). Accent variability was already minimized given that the database was designed and created with a neutral North American accent, to avoid participants using accent as a voice recognition shortcut (Gluszek & Dovidio, 2010). Speech clips were trimmed to exclude silence in the beginning and end of the original recordings in Praat v6.1.04 (Boersma & Weenink, 2019). Loudness was then normalized using the Loudness Toolbox (Genesis S.A.) in Matlab 2017b. The final speech clips had a mean duration of 1.75 s (SD = 0.30 s; range: 1.26 – 2.91 s).

Face images were taken from the Karolinska Directed Emotional Faces (KDEF) database (Lundqvist et al., 1998). Forty-eight actors (24 female, aged 20-30) were selected based on highest emotion categorization accuracy on the four emotional expressions (Goeleven et al., 2008), taken into consideration together with the criteria of as little presence of significant features/marks on images that enabled easy recognition as possible (for the list of selected actors and additional information of the stimuli, see Supplementary Information [SI]). The exterior of faces (e.g., shoulder) was removed using Adobe Photoshop CS5.1 (Adobe Systems, San Jose CA), to achieve a uniform face size, contrast and resolution (see Sergerie, Lepage, & Armony, 2005; 2006).

Normative values for emotional judgments on the face and voice stimuli, obtained from three validation studies (Goeleven et al., 2008, Sutton, Herbert & Clark, 2019 for faces; Livingstone & Russo, 2018 for voices), are provided in S.Table 3-1.

3.3.3 Procedure

Participants were instructed to complete the online studies on a computer browser in a quiet environment. There was no interaction between participants and experimenters during the procedure. The experiment was designed to work in the most common browsers and operating systems. Access to a keyboard and headphones was required for participation.

In all three studies, each participant completed two single-modality modules (one face, one voice) in full-screen mode sequentially, with the order counterbalanced. Each module contained an encoding and a recognition phase. Each phase was programmed to automatically start three seconds after the completion of the preceding one. While there were self-paced text instructions at beginning of each phase, participants were told not to take long pauses on the instruction pages. Post-hoc tests also confirmed that the time spent on instructions between encoding and recognition phases did not have a significant influence on performances. In the face module, participants were presented a series of emotional face images from a number of identities (half female), each for 1.5 s, and were asked to judge the age bracket of each face (younger or older than 30 years) by pressing a key. Immediately after the encoding phase, an old/new recognition test took place, where a novel-expression face image of previously encoded identities and an equal amount of novel identities was presented in pseudo-random order. Additionally, subjects were asked to give a binary confidence rating (*Sure/Unsure*) on each response they made during recognition. In both phases, the following trial started 0.5 s after a response was recorded. The structure of the voice module was identical to the face module. Participants were instructed beforehand that there was no relation between the stimuli of both modules. A schematic of the experimental trials is shown in Figure 3-1. Study-specific manipulations on encoding conditions are further described in detail under each study and summarized in Table 3-1.

Throughout the online experiment, the assignment of response key in both the encoding (age) and recognition trials was randomized within subject (e.g., a 'Older than 30' or 'Old' response may be linked with 'J' key in one trial, but with 'K' key in another). This was to avoid potential subject-specific response lateralization tendency (e.g., tendency to press a given key). However, the key assignment for the confidence rating was fixed within subject and randomized across subjects. Lastly, a short practice session was implemented before the actual experiment, containing the encoding and recognition phases of both modality modules, with fewer and entirely different stimuli (for details of the practice session, see Practice Session section in SI).

Feedback on recognition accuracy was given at the end of each practice trial to ensure participants understood the task and were comfortable with the stimuli.

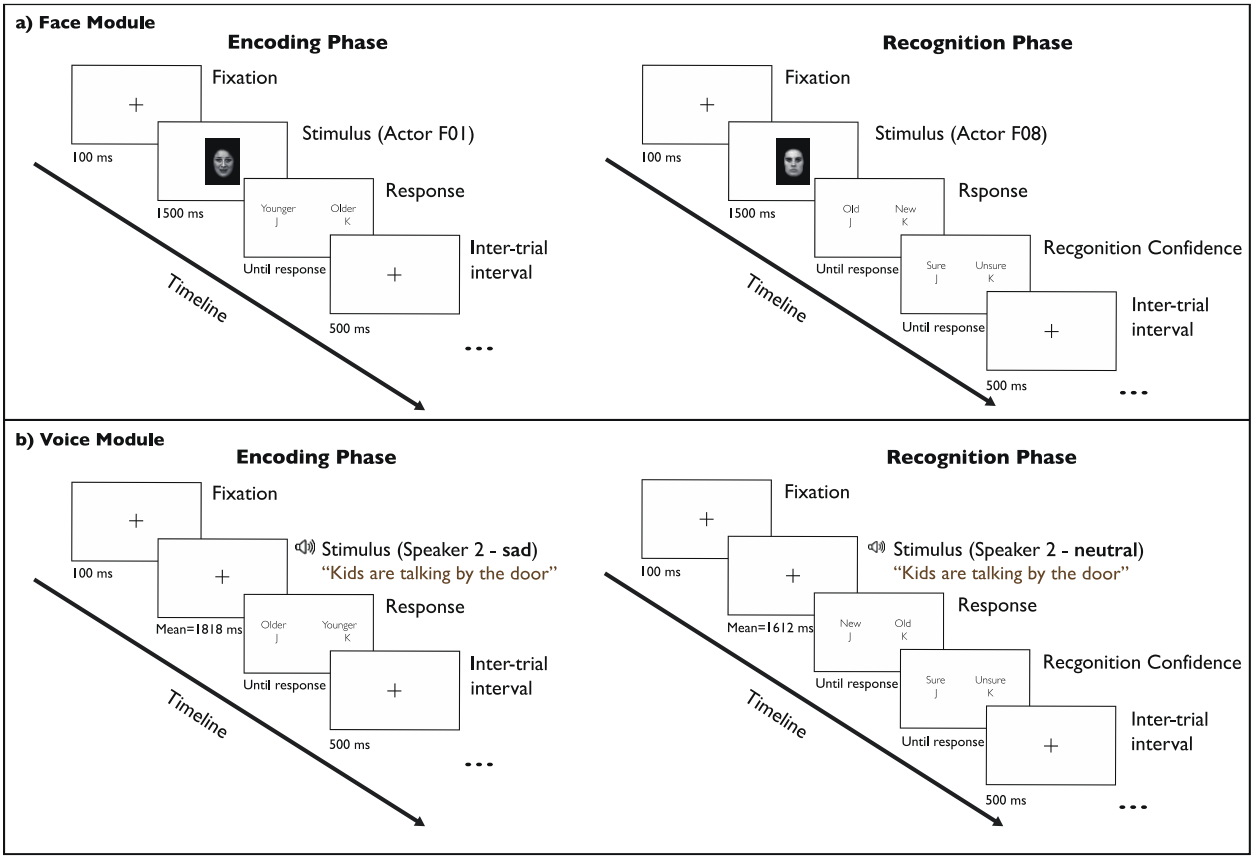

Figure 3-1. Procedure and example trial of the experimental task in the studies. a) and b) depict the example trials for both encoding and recognition phases, in face and voice modules respectively.

3.3.4 Dependent Measures

Response choices and response times (RTs) for both encoding and recognition tasks, as well as the binary confidence ratings during the recognition test, were recorded using customized jsPsych plugins for subsequent analysis. Data used for the following analyses are available from the authors upon request.

3.3.5 Data Analysis

3.3.5.1 Recognition Accuracy

Binary measures of accuracy (correct/incorrect) and confidence (Sure/Unsure) during the recognition test were analyzed on a trial-by-trial level, using a generalized linear mixed model (GLMM) with a logit link function. Fixed-effect factors were defined according to the experimental design, as described in the study-specific Methods sections below. The random effect structure included random intercepts of subject and stimulus identity, with the latter accounting for possible differences in identity-specific distinctiveness and to allow for the generalization of findings beyond the material used here (Clark, 1973; Baayen, Davidson, & Bates, 2008). The analysis was implemented in R with the *lme4* R package (Bates et al., 2015). The omnibus effects of modelled fixed factors were obtained via *Anova* function (type III) from *car* package (Fox & Weisberg, 2019).

3.3.5.2 Drift Diffusion Models (DDMs)

We employed DDMs (Ratcliff & McKoon, 2008) to analyze subjects' binary choices, combining their trial-specific responses and RTs, as a process in which evidence is accumulated in each trial towards one of the two possible responses, until a threshold is reached, and a decision is made. A key parameter of this model is the drift rate (v), defined as the rate of information accumulation, determined by the quality of the extracted stimulus information. DDMs have been extensively used to examine cognitive processes such as memory and decision making (for a review, see Ratcliff et al., 2016) and, more recently in emotion research (e.g., Williams et al., 2023; Mueller & Kuchinke, 2016, for emotional face and word perception, respectively). In the context of recognition memory, drift rate can be interpreted as the quality of match between the tested stimulus and memory, which is usually expected to vary among different experimental conditions. The other main parameters of the model are the starting point (z), usually used to model the changes of response proportions, and the boundaries (a), which are assumed to vary in cases of an instruction change (e.g., emphasis on response speed or accuracy).

We adapted a recent variation of diffusion models, D*M (Verdonck & Tuerlinckx, 2016), using the DstarM package in R. Specifically, we ran D*Ms on trial-level response and RT data for each module of each participant separately, to estimate the parameters v, a, z, and within-subject variability of v (sv) and z (sz). Drift rates were modeled to vary among recognition

conditions, while the other parameters were held constant as they were not hypothesized to change across experimental conditions. The estimated subject-condition-specific drift rates were then submitted to linear mixed models (LMMs), with fixed factors including modality and other study-specific variables. The random effect structure contained a random intercept of subject. Results of the LMM was obtained via *anova* function (type III) from *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

For both LMM and the GLMM results, whenever significant interactions were present, post-hoc pairwise tests (for categorical variables) or post-hoc tests of linear trends (for continuous variables) were conducted using the *emmeans* R package (Lenth, 2020) (Bonferroni-Holms correction implemented when needed). Effect sizes were obtained with the *RESI* package (Vandekar, Tao, & Blume, 2020) for fixed omnibus effects in the models (denoted as S index), and *emmeans* package for post-hoc pairwise tests on interactions: odds ratio [OR] for post-hoc tests in GLMMs, and Cohen's *d* in LMMs.

3.3.5.3 Stimulus-based physical feature analysis

In order to investigate the possibility that (part of) the results could be explained by intrinsic stimulus physical characteristics, we conducted a stimulus-based similarity analysis for voice and face stimuli separately, as a function of emotional expression. For voices, we extracted 17 acoustic parameters from each speech clip, including stimulus duration and multiple formant frequency descriptive statistics (see Xu & Armony, 2021 for details) using Praat v6.1.04 (Boersma & Weenink, 2019). In the case of faces, we conducted a pixel-based principal component analysis on the 240 face images for dimension reduction. We kept 56 principal components ("eigenfaces"), explaining 90% of the total variance, with the corresponding coefficients per stimulus characterizing their physical attributes. Each stimulus was therefore represented in a multidimensional (17 for voices, 56 for faces) feature space (Armony, Chochol, Fecteau, & Belin, 2007; Baumann & Belin, 2010; Latinus et al., 2013; Sergerie, Lepage & Armony, 2005).

First, we examined whether stimuli within certain emotions were more similar among each other than other emotions, by calculating the average cosine similarity between each stimulus and all the other same-emotion stimuli and. Those values were then analyzed in a one-factor

(emotion) LMM, with a random identity intercept. Next, we examined whether each identity's physical change from a given emotion to its neutral counterpart differed among emotions. To do so, we computed the cosine similarity between each emotional (angry, fearful, happy, and sad) and neutral expression, both within and between identities. Then, the difference of within- and between-identity similarity per identity per emotion, representing an identity's distinctiveness from other identities in one emotion, was entered in a one-factor (emotion) LMM, with a random identity intercept.

3.4 Main Study

3.4.1 Methods

3.4.1.1 Participants

A cohort of 556 healthy individuals (523 and 33 from the Participant pool and Facebook post, respectively; see General Methods) took part in the study. One participant was excluded because of repeated participation. Eight participants were excluded due to incomplete participation. Thus, the final sample consisted of 547 participants (481 female aged 18 – 31 years: Mean [M] = 20.5, Standard Deviation [SD] =1.5; 66 male aged 18 – 32 years: M = 21.1, SD = 2.3).

3.4.1.2 Procedure

In the face module, participants first viewed 96 face images from 24 randomly selected identities (half female); that is, during this encoding phase, each identity was presented four times, intermixed with presentations of others. Twelve identities were randomly assigned to the *Multi* (exemplar variability) condition, and thus presented in four distinct emotional expressions (fearful, happy, angry and sad) once each. The other 12, assigned to the *Uni* condition, were presented with a single emotional expression (i.e., same stimulus) four times. The single emotion used in this condition was pseudo-randomized across participants, resulting in four groups of participants (i.e., *Uni*-Fearful [N=145], *Uni*-Happy [N=144], *Uni*-Sad [N=129], *Uni*-Angry groups [N=129]). The *Uni* and *Multi* trials were pseudo-randomly intermixed, and the distribution of lags (i.e., number of intervening stimuli) between same-identity exemplars did not differ between the *Uni* and *Multi* conditions. Moreover, both presentation order of emotions and first-order transition of same-identity emotional stimuli, were balanced across participants among *Multi* trials. The old/new recognition test consisted of the 24 encoded identities and 24

novel ones, all with a neutral expression. The voice module shared the identical structure but with 12 identities presented during encoding, and 24 (half new) during recognition. We used a smaller number of identities to ensure adequate memory performance, as determined by pilot experiments and previous studies (Xu & Armony, 2022). The assignment of identities to each category (*Uni*, *Multi* or *New*) was random and counterbalanced across participants.

3.4.1.3 Data Analysis

Encoding

Response times (RTs) from the encoding task (age judgment) were analyzed to test whether there were stimulus- and/or identity-specific implicit memory (priming) effects. RTs were cleaned by excluding trials with RTs that were 3 SD beyond mean value per subject. Then, a linear regression on same-identity RTs was performed for each encoded identity per subject (Lorch & Myers, 1990). The regression coefficients (RT-slopes) were then submitted to an LMM, with encoding variability (*Multi*/*Uni*) and modality (visual/auditory) as fixed within-subjects factors, and *Uni*-emotion as fixed between-subjects factor.

We also computed within-identity consistency in the age judgment responses across the 4 presentations, the Age Consistency Score (ACS), as follows: ACS = 1-[#presentations - max(#Old, #Young)]. That is, the possible values were 1 (all 4 age responses were the same), 0 (3 out of 4 responses were the same), and -1 (2 out of 4 responses were the same).

The subject- and identity-specific RT-slopes and ACSs were then entered as covariates in the recognition accuracy model, described below, to determine whether priming and/or age consistency predicted subsequent identity recognition memory and, if so, whether this effect interacted with the encoding condition (*Multi*/*Uni*) .

Recognition of old identities

Overall recognition accuracy (across both new- and old-identity trials) was analyzed in a simple mixed ANOVA with modality as a within-subjects factor and *Uni*-emotion group as a between-subjects factor. Three (G)LMM models were constructed within old identity trials, namely on recognition accuracy, DDM-derived drift rates, and recognition confidence.

(1) Recognition accuracy

In this GLMM, fixed effect structures consisted of two within-subject factors, encoding exemplar variability condition (*Uni/Multi*) and modality (visual/auditory), and one between-subjects factor, *Uni*-emotion group (*Uni*-Angry/*Uni*-Fearful/*Uni*-Happy/*Uni*-Sad). The two encoding indices representing identity priming and age consistency (RT-slope and ACS) were entered into the model as covariates.

To ensure that any difference (or lack thereof) in memory accuracy between *Uni* and *Multi* conditions was truly reflective of the entire dataset (as opposed to the possibility of being driven by a small number of extreme or high-leverage subjects), we performed a series of n-jackknife resampling cross-validations: for each participant group, 1000 random subsets of n subjects of the original sample were selected (n = 5-95% of the original sample, in steps of 5%). At each data fraction, the mean jackknife estimate of the *Multi*-*Uni* accuracy difference and its 95% confidence interval were calculated.

(2) Drift rates

Subject- and condition-specific drift rates obtained from the drift diffusion model were entered into an LMM with condition and modality as within-subject factors and *Uni*-emotion as a between-subject one.

(3) Recognition confidence

Recognition confidence was analyzed in a GLMM with encoding variability condition, modality and accuracy as within-subjects, and *Uni*-emotion as between-subjects, factors. In order to validate the hypothesized relationship between confidence rating and response times (Shaw, McClure, & Wilkens, 2001; Robinson, Johnson & Herndon, 1997; Weidemann & Kahana, 2016), we added log-transformed RTs as covariate in the model.

Detection of new identities

For new identity trials, (G)LMMs with a similar structure were constructed on the same three measures: recognition accuracy, DDM-derived drift rates, and recognition confidence. Importantly, as new identity trials did not have the encoding exemplar variability condition, the models only contained modality and *Uni*-emotion as within- and between-subjects fixed factors, respectively. As with old identities, we included response accuracy and log-RT as covariates in the recognition confidence GLMM.

Stimulus-based arousal-rating analysis

Finally, to assess the potential impact of stimulus arousal on identity memory, we performed linear models on recognition accuracy of *Uni* identities, using the arousal ratings of stimuli from two prior validation studies (Sutton, Herbert & Clark, 2019 for faces; Livingstone & Russo, 2018 for voices). To account for the discrepancies between rating ranges of the two datasets (1-5 points for voices, and 1-9 points for faces), the rating data were normalized within each modality prior to modelling. Linear models were then performed on subject-averaged recognition accuracy with arousal rating, emotion, and modality as fixed factors.

3.4.2 Results

3.4.2.1 Encoding: Implicit Memory (Priming)

The LMM on estimated identity-repetition RT slopes (an index of priming; see Methods for more details) as a function of modality (visual/auditory), encoding variability (*Multi*/*Uni*) and *Uni*-emotion (participant groups) revealed that in all cases the slopes were significantly negative (*Uni*: b = -0.067, SE = 0.004, z = -17.60, p < .001, Cohen's d = 0.26; *Multi*: b = -0.052, SE = 0.004, z = -13.59, p < .001, Cohen's d = 0.20), reflecting identity-specific priming (i.e., faster responses as a function of repeated presentation of the same individual). Nonetheless, a main effect of encoding variability (F[1,19133] = 15.63, p < .001, S = 0.31) indicated that, as expected, this effect was stronger in the *Uni* than *Multi* condition. Additionally, there was a main effect of modality (F[1,19133] = 139.85, p < .001, S = 0.59), due to the steeper slopes for voices than faces.

3.4.2.2 Recognition: Identity Memory

Overall accuracy

A 2 (modality) by 4 (*Uni*-emotion) mixed ANOVA on overall accuracy revealed a significant modality effect (F[1,543] = 739.90, p < .001, $\eta_p^2 = .58$), due to a better memory for faces than voices, with no *Uni*-emotion main effect (F[3,543] = 0.65, p = .58, η_p^2 = .004) or interaction (F[3,543] = 0.19, p = .90, η_p^2 = .001). Planned t-tests confirmed that overall memory accuracy was significantly above chance level for both modalities (voice: t[546] = 130.21, 95% CI = [0.53, 0.55], p < .001; face: t[546] = 172.28, 95% CI = [0.69, 0.70], p < .001; descriptive stats in S.Table 3-2).

Recognition of old identities

Results of the GLMM (see S.Table 3-3 for full results) on recognition accuracy with encoding variability, modality as within-subject, and *Uni*-emotion as between-subject factors, as well as encoding RT-slope and Age Consistency Score as covariates, revealed significant main effects of modality ($\chi^2[1] = 7.54$, $p = .006$, S = 0.11; better memory for faces than voices), encoding exemplar variability ($\chi^2[1] = 6.60$, $p = .010$, S = 0.10; better memory for *Multi* than *Uni* identities), RT-slope ($\chi^2[1] = 8.51$, $p = .004$, S = 0.12; stronger identity priming predicted better recognition), and ACS ($\chi^2[1] = 23.17$, $p < .001$, S = 0.20; higher intra-identity consistency in age judgment predicted higher recognition accuracy). Importantly, the exemplar variability effect was qualified by a significant variability-by-emotion interaction, $\chi^2(3) = 11.25$, $p = .010$, S = 0.12. Post-hoc tests on encoding condition (see S.Table 3-4, and Figure 3-2a for visualization) indicated that this interaction was due to the memory advantage for *Multi* identities in all groups (p's $\leq .003$, OR's ≥ 1.29) except the *Uni*-Sad group ($b = -0.003$, SE = 0.07, $z = 0.05$, $p = .96$, OR = 1.00).

To rule out the possibility that the difference between *Uni*-Sad and the other groups could have been driven by subject-specific confounding effects (e.g., extreme or high-leverage data points), we performed an n-jackknife supplementary analysis. This analysis showed that the encoding variability effect (or lack thereof) was stable even for small random subsets of the original sample (Figure 3-3). Specifically, for the *Uni*-Angry, Fearful and Happy groups, the *Multi* advantage remained significant with only about 50% or less of the original sample sizes. In contrast, for the *Uni*-Sad group, the 95% CIs included zero for all subsamples, indicating a reliable lack of accuracy difference between encoding conditions.

Subject- and condition-specific drift rate estimates from the diffusion models (D*Ms) are summarized in Figure 3-2b (descriptive stats in S.Table 3-2). The LMM for old identities revealed both main effects of encoding condition ($F[1,1581.1] = 39.03$, $p < .001$, S = 0.26) and modality ($F[1,1583.5] = 300.94$, $p < .001$, S = 0.74), as well as emotion-by-variability ($F[3, 1580.9] = 4.31$, $p = .005$, S = 0.13) and modality-by-condition ($F[1, 1579.6] = 6.95$, $p = .008$, S = 0.10) interactions (S.Table 3-3). Consistent with the accuracy results, the emotion-by-variability interaction was due to a significantly larger drift rate in the *Multi* than *Uni* condition, in all (p's $< .01$, Cohen's $d > 0.2$) but the *Uni*-Sad group ($b = 0.05$, SE = 0.072, $t[1603] = 0.76$, $p = .45$,

Cohen's d = 0.068) via post-hoc pairwise tests (S.Table 3-4). The modality-by-variability interaction was due to the fact that although both modalities showed a larger drift rate for *Multi* than *Uni* identities, the effect was larger for faces (voice: b = 0.13, SE = 0.05, t[1595] = 2.55, Cohen's d = 0.16; face: b = 0.31, SE = 0.05, t[1594] = 6.29, Cohen's d = 0.39).

The GLMM on response confidence (*Sure*/*Unsure*) yielded a significant effect of accuracy (χ^2[1] = 21.79, p < .001, S = 0.19), modality (χ^2[1] = 11.06, p = .001, S = 0.14), and log-transformed RTs (χ^2[1] = 722.20, p < .001, S = 1.15), as well as variability-by-modality (χ^2[1] = 4.62, p = .032, S = 0.081), modality-by-accuracy (χ^2[1] = 11.51, p < .001, S = 0.14), and variability-by-modality-by-accuracy (χ^2[1] = 6.08, p = .014, S = 0.096) interactions (see S.Table 3-5). The significant effect of RTs indicated faster responses for *Sure* trials, supporting the hypothesis that they can serve as a proxy for response confidence (Shaw, McClure, & Wilkens, 2001; Robinson, Johnson & Herndon, 1997; Weidemann & Kahana, 2016). Post-hoc tests on the three-way interaction indicated that participants were significantly more confident when successfully recognizing *Multi*-encoded than *Uni*-encoded identities for faces (b = 0.13, SE = 0.06, z = 2.35, p = .019, OR = 1.14), but not for voices (b = 0.06, SE = 0.08 z = 0.82, p = .41, OR = 1.06). No difference in incorrect response confidence between *Multi* and *Uni* identities was found in either modality (voice: b = 0.001, SE = 0.09, z = 0.01, p = .99, OR = 1.00; face: b = -0.09, SE = 0.08, z = -1.22, p = .22, OR = 0.91).

Detection of new identities

Similar analyses were conducted for performance on new identities, but with only two fixed factors, modality and *Uni*-emotion group, as these identities were not related to the expression-variability manipulation in the encoding phase. The accuracy GLMM (full stats in S.Table 3-3) revealed a significant effect of modality (χ^2[1] = 37.12, p < .001, S = 0.28; higher accuracy for faces than voices), but no main effects interactions involving *Uni*-emotion groups (p's > .5, S < 0.001).

The drift-rate LMM (full stats in S.Table 3-3) also yielded a modality effect (F[3,535.4] = 533.38, p < .001, S = 0.98): whereas the drift rates for new faces were significantly negative (i.e., towards a "New" response; b = -0.81, SE = 0.03, t[1061] = -27.18, p < .001, Cohen's d = 1.25), those for voices were positive (b = 0.10, SE = 0.03, t[1060] = 3.43, p < .001, Cohen's d = 0.16), reflecting a tendency to respond "Old" towards new voices (Figure 3-2b).

In the case of the binary confidence responses, the GLMM yielded main effects of accuracy ($\chi^2[1] = 11.56$, $p < .001$, S = 0.14) and modality ($\chi^2[1] = 67.14$, $p < .001$, S = 0.35), as well as an interaction between the two ($\chi^2[1] = 39.63$, $p < .001$, S = 0.27). Post-hoc tests on the interaction showed that whereas correct responses of faces were made with higher confidence than incorrect ones ($b = 0.57$, SE = 0.05, $z = 12.35$, $p < .001$, OR = 1.77), the opposite was true for voices ($b = -0.45$, SE = 0.06, $z = -7.10$, $p < .001$, OR = 0.64). There was also a significant effect of the log-RT covariate ($\chi^2[1] = 742.64$, $p < .001$, S = 1.16; see S.Table 3-5), confirming the relation between implicit (faster RTs) and explicit (Sure/Unsure choice) confidence measures.

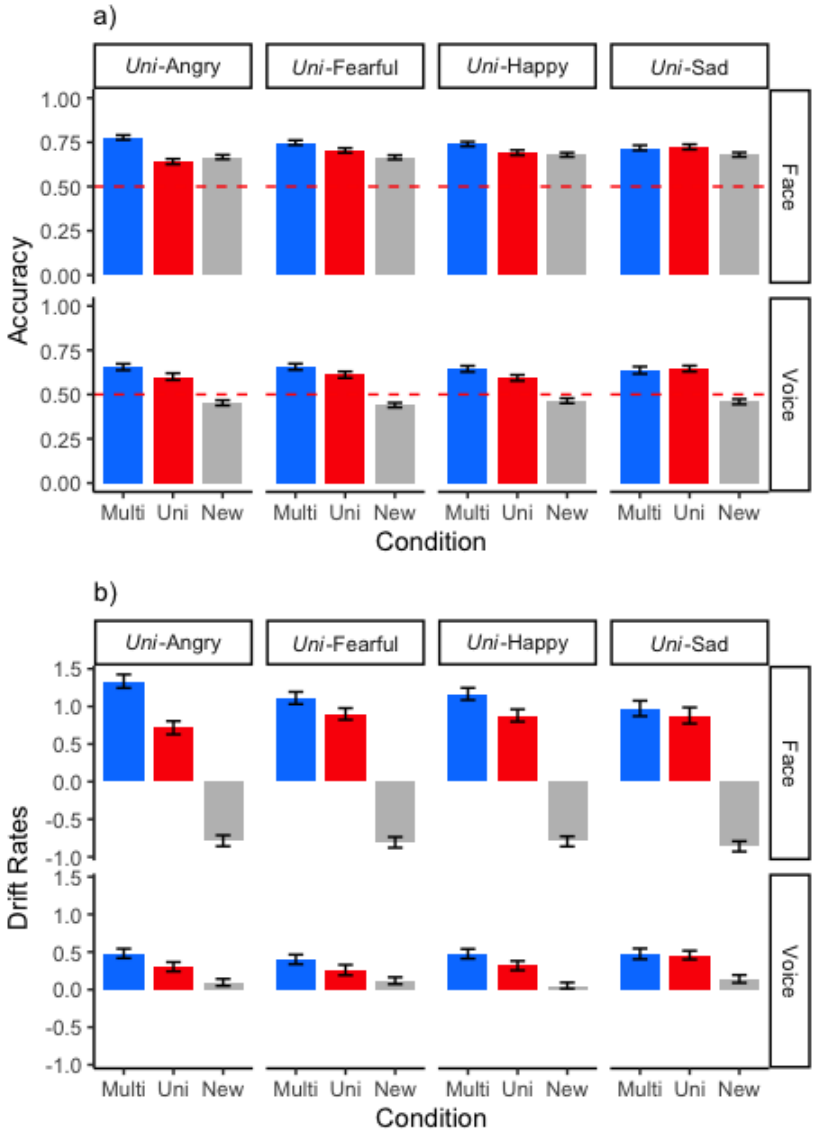

Figure 3-2. Averaged recognition accuracy (*a*) and DDM-derived drift rates (*b*) by encoding variability in each *Uni*-emotion group and modality (dashed horizontal line in *a* indicated chance-level accuracy).

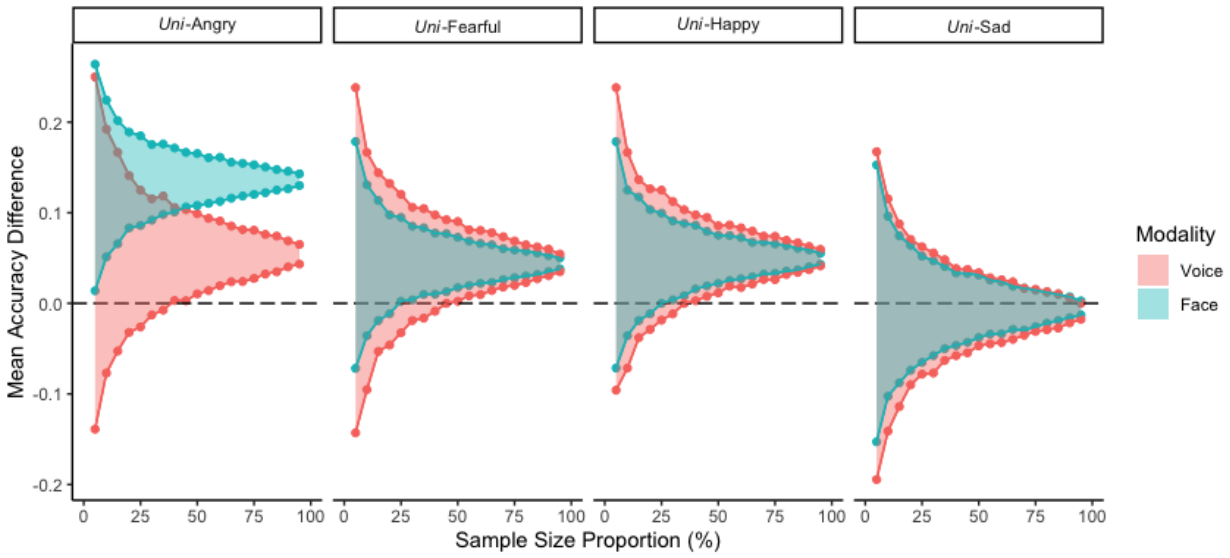

Figure 3-3. Confidence intervals (95%) of accuracy difference between *Multi* and *Uni* conditions from the Main study, using an n-jackknife subsampling approach with 5% step-wise sample sizes from 5% to 95%.

Stimulus-based physical features

A one-way LMM on the within-emotion similarity for vocal stimuli showed a significant effect of emotional prosody ($F[4,92] = 9.24$, $p < .001$, S = 0.52) (descriptive statistics of computed cosine similarity in S.Table 3-6). Post-hoc pairwise *t*-tests revealed that this was driven by within-neutral similarity being significantly larger than within-sad, fearful and angry similarities ($t[92]$'s > 3.5, p's < .003, Cohen's d's > 1.1) as well as within-happy similarity being larger than within-fearful similarity ($t[92] = 3.06$, $p = .029$, Cohen's $d = 0.88$; complete pairwise test statistics listed in S.Table 3-7). A one-way LMM on the difference of within- and between-identity similarity between emotional and neutral stimuli did not reveal a significant effect of prosody ($F[3,69] = 1.12$, $p = .35$, S = 0.054), suggesting that the speaker distinctiveness in acoustic similarity between emotional and neutral speech did not differ among the four emotions.

The same analyses on face-image derived principal components (see S.Table 3-6 for cosine-similarity descriptive statistics) also showed a significant effect of emotional expression on within-emotion average similarity ($F[4,188] = 2.52$, $p = .043$, S = 0.16), which was driven by a significantly higher within-happy than within-neutral face similarity ($t[188] = 3.04$, $p = .027$,

Cohen's $d = 0.62$; see complete pairwise test statistics in S.Table 3-7). For the emotional-to-neutral identity distinctiveness, the LMM did not reveal a significant emotion effect ($F[3,141] = 0.20$, $p = .89$, $S < 0.001$). Taken together, these results support the notion that the behavioral recognition patterns reported above were unlikely to be due to intrinsic physical differences of the stimuli in either modality.

Stimulus-based arousal rating

The linear model with stimulus-level accuracy of all *Uni* recognition trials as dependent variable and emotion, modality and stimulus arousal intensity rating as factors, yielded a main effect of modality ($F[1,269] = 17.73$, $p < .001$, $S = 0.22$; higher accuracy for faces than voices) and arousal rating ($F(1,269] = 22.15$, $p < .001$, $S = 0.27$; descriptive statistics and fixed factor effects of the model are shown in S.Tables 3-1 and 3-8, respectively). The latter effect reflected a negative correlation between stimulus emotional arousal and identity recognition memory for both faces and voices, independent of the specific emotion expressed, as shown in Figure 3-4.

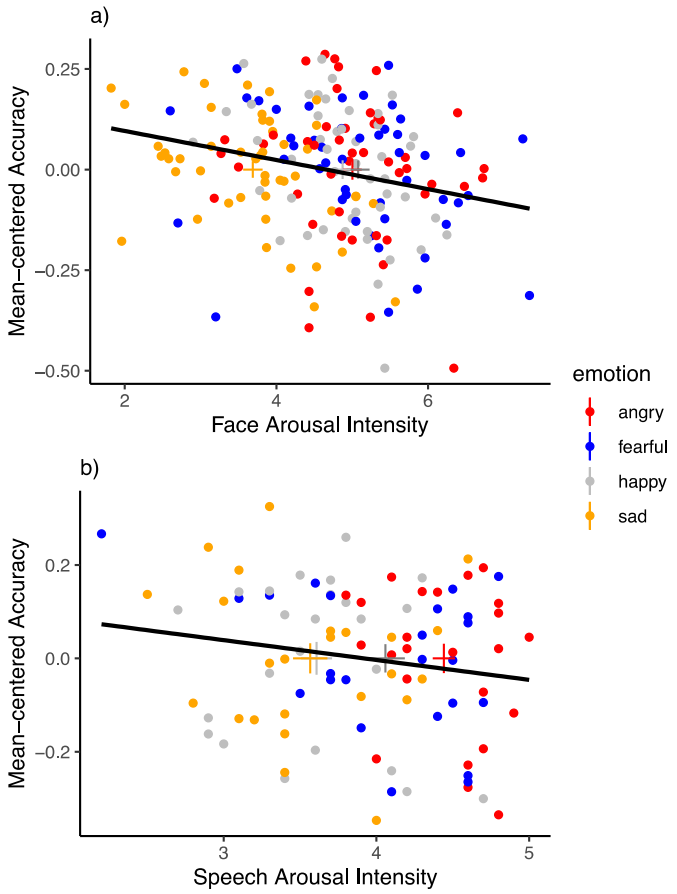

Figure 3-4. Scatterplot and estimated regression line of stimulus arousal intensity ratings against across-subject mean accuracy after accounting for accuracy difference among *Uni*-emotion groups (crosses and the arm lengths represent mean and standard error of intensity ratings per stimulus emotion category).

3.4.3 Summary

Results from the Main Study revealed a significant advantage of emotional expression variability (*Multi* condition) during encoding for the subsequent recognition of facial and vocal identities, relative to the repeated presentation of angry, happy and fearful, but not sad expressions (*Uni* condition). In the *Uni*-Sad group, there was no significant difference in recognition accuracy between *Multi* and *Uni* encoding conditions. As the *Multi* condition was the same for all *Uni* emotion groups, we speculated that the absence of an advantage for this condition in the *Uni*-Sad group was driven by whatever feature distinguishes this emotion from the others (happiness, anger and fear), one likely candidate being emotional intensity or arousal (e.g., Kensinger, 2009). This possibility was supported by the significant negative correlation between stimulus arousal, or emotional intensity, and identity recognition accuracy across expressions and modalities. That

is, individuals whose emotional expressions were judged to be less arousing were better recognized later, when presented in a novel, neutral expression, than those considered to express high intensity emotions. However, the between-group nature of our design did not allow for directly testing this possibility; we therefore conducted two additional studies, each consisting of two related experiments, focusing on specific combinations of emotional expressions based on their arousal level (Table 3-1).

3.5 Follow-up Study 1: High vs. low arousal emotions

3.5.1 I: Fearful vs. Sad

In this experiment, we directly compared recognition performance of identities presented with only one expression during encoding, either sadness or fear, in a within-subject design. Based on the results from the Main Study, we predicted a memory recognition advantage for sad-encoded identities, compared to those encoded with a fearful expression.

3.5.1.1 Methods

Eighty-six (18 – 26 years: M = 20.5, SD = 1.6; 74 female) new participants were recruited through the McGill Psychology extra-credit participant pool. The overall procedure was similar to the original experiment in the *Uni*-Sad group, consisting of both face and voice separate modules, but with the original *Multi* encoding condition replaced by a *Uni* condition in fearful expressions. Hence, all encoded identities (same amount as in the Main Study) were presented four times, half with a single sad exemplar and the other with a fearful one. To test our hypothesis, we focused our analysis on the trial-level recognition response and DDM-derived drift rates of old identity trials. The modeling structure was similar as in the Main Study, but with only modality, and the encoding emotion (fearful vs. sad) as within-subject factors.

3.5.1.2 Results

The GLMM on recognition accuracy yielded main effects of modality ($\chi^2[1] = 8.99$, $p = .003$, S = 0.30) and encoding emotion ($\chi^2[1] = 8.34$, $p = .004$, S = 0.29) without a significant interaction ($\chi^2[1] < .001$, $p = .99$, S < 0.001), suggesting an overall better recognition for face, and a better recognition for sad than fearful encoded identities (Figure 3-5a). For subject-level condition-

specific drift rates, the LMM revealed significant effects of both modality (F[1,252.61] = 50.46, p < .001, S = 0.76) and encoding emotion (F[1,251.31] = 6.35, p = .012, S = 0.25), without a significant interaction (F[1,251.29] = 0.28, p = .60, S < 0.001), due to larger drift rates for faces than voices, and for sad- than fearful-encoded identities (Figure 3-5b; descriptive stats in S.Table 3-2).

3.5.2 II: Fearful vs. Neutral

3.5.2.1 Methods

To further confirm, and extend, the results from the previous experiment, we conducted a similar experiment but using neutral expressions, instead of sad, as the low-arousal expression. To keep the emotional expressions different between encoding and recognition, as in the previous experiments, we used happy faces/voices in the recognition test. The recruitment process remained the same, with a group of 86 new participants completing the experiment (18 – 29 years: M = 20.3, SD = 1.7; 76 female). Analysis on recognition response and drift rates was the same as in the Fearful vs. Sad experiment.

3.5.2.2 Results

The GLMM on recognition accuracy yielded main effects of modality (χ^2[1] = 10.15, p = .001, S = 0.33) and encoding emotion (χ^2[1] = 49.05, p = .003, S = 0.31) without a significant interaction (χ^2[1] = 0.63, p = .43, S < 0.001), driven by an overall better recognition for faces, and for neutral- than fearful-encoded identities (Figure 3-5). Likewise, the LMM on DDM-derived drift rates revealed significant main effects of modality (F[1,251.37] = 62.45, p < .001, S = 0.84; larger drift rates for faces than voices) and encoding emotion (F[1,251.36] = 8.62, p = .004, S = 0.30; larger drift rates for neutral- than fearful-encoded identities), without a significant interaction (F[1,251.37] = 0.82, p = .37, S < 0.001).

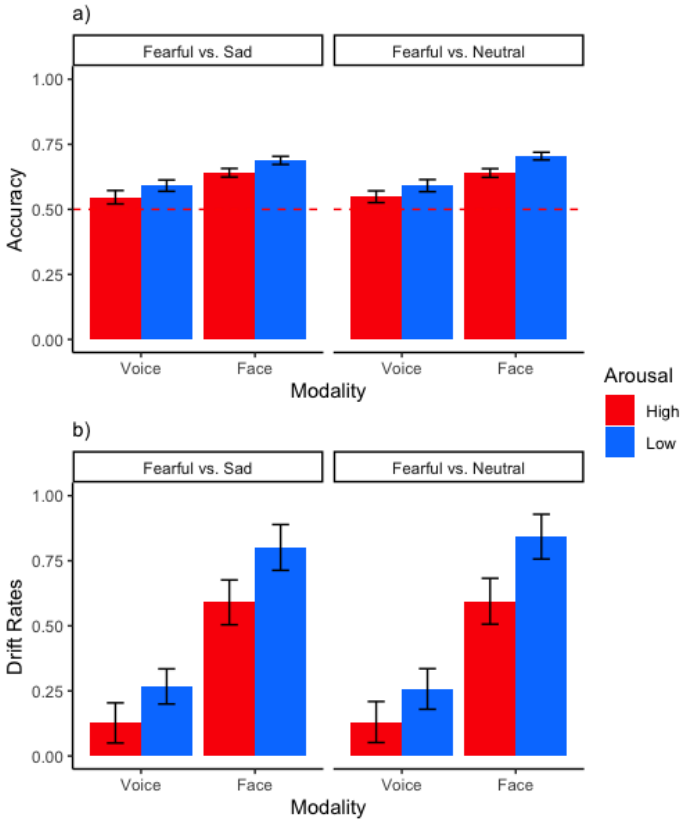

Figure 3-5. Averaged recognition accuracy (*a*) and DDM-derived drift rates (*b*) by encoding emotion arousal level per modality from Follow-up Study 1 (dashed horizontal line in *a* indicated chance-level accuracy).

3.6 Follow-up Study 2: High-arousal *Multi* vs. *Uni*

Results from the Main Study and Follow-up Study 1 point to a key role of arousal in identity recognition. Because the *Multi* condition in the Main Study included both high- (anger, fear and happiness) and low- (sadness) arousal emotions, there remains the question of whether the better performance for this condition, compared to high-arousal *Uni* ones, was due to the presence of the low arousal expression in the former. Thus, to directly test this possibility, we conducted an additional online experiment with the same paradigm and recruitment criteria, without using any sad stimuli. Specifically, each identity was presented three, instead of four, times, during encoding. In the *Multi* condition, three emotional exemplars (happiness, fear, and anger) of each identity were presented once; in the *Uni* condition, either one angry or fearful exemplar for each identity was presented three times, which served as a between-subjects variable (i.e., *Uni*-Fearful vs. *Uni*-Angry). We expected the multiple exemplar advantage to be present for both groups.

3.6.1 Methods

A total of 174 new participants (18 – 31 years: M = 20.5, SD= 1.7; 154 female) were recruited via the McGill Psychology extra-credit participant pool under the same criteria as previous studies. Half of the participants (N=87) were presented with angry exemplars in the *Uni* condition, while the other half with fearful exemplars. We analyzed recognition accuracy and DDM-derived drift rates of old identity trials in the same manner as before.

3.6.2 Results

The GLMM on recognition accuracy, with modality and exemplar variability as within-subject and *Uni*-emotion as between-subjects factors (S.Table 3-9), revealed significant modality ($\chi^2[1]$ = 8.46, p = .004, S = 0.21) and variability ($\chi^2[1]$ = 20.28, p < .001, S = 0.33) effects, driven by, as in the case of the main study, a more accurate recognition for faces, and for *Multi*-encoded identities (Figure 3-6a). The LMM on drift rates yielded a similar pattern, with significant effects of modality (F[1,503.41] = 127.12, p < .001, S = 0.33) and variability (F[1, 502.05] = 20.42, p < .001, S = 0.85). Additionally, there was a significant *Uni*-emotion effect (F[1, 167.14] = 4.24, p = .041, S = 0.14), due to overall larger drift rates in the *Uni*-Fearful than *Uni*-Angry groups, as illustrated in Figure 3-6b.

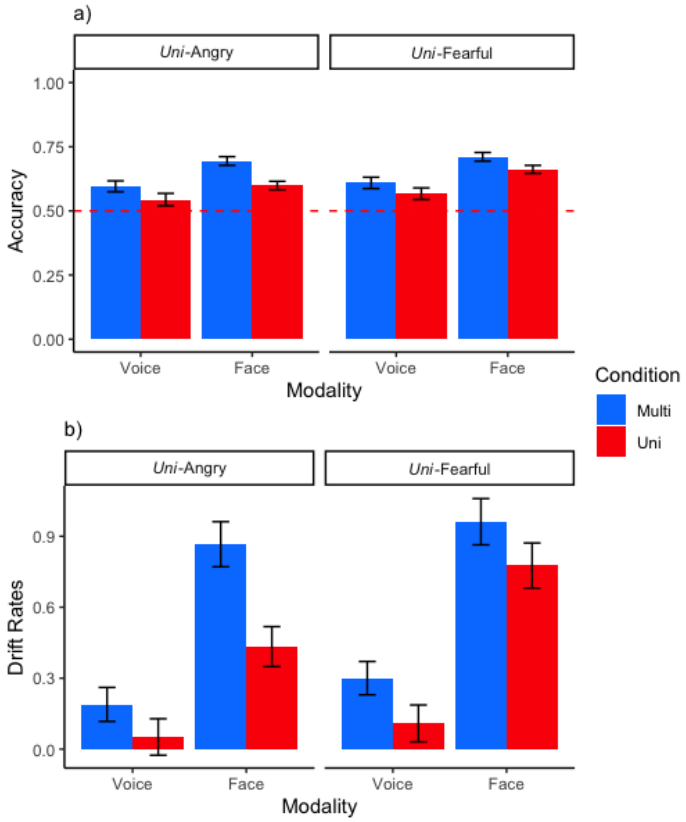

Figure 3-6. Averaged recognition accuracy (*a*) and DDM-derived drift rates (*b*) by encoding variability per modality from Follow-up Study 2 (dashed horizontal line in *a* indicated chance-level accuracy).

3.7 General Discussion

3.7.1 Implicit and explicit identity memory

As expected, we observed significant priming in repeated presentations of the same face and voice expression, as indexed by negative RT slopes of the age judgment task. Notably, a similar effect, albeit weaker, was observed when the same individual was presented with different emotional expressions. That is, over the course of the encoding session, participants developed expression-independent implicit memory for previously-encountered individuals.

Interestingly, the magnitude of the priming effect during encoding positively predicted subsequent recognition memory for both faces and voices. These findings contribute to a literature that has yielded conflicting results (e.g., Li & Jiang, 2020; Miyoshi et al., 2014), providing new and strong support for a direct relation between implicit and explicit memory processes (Turk-Browne, Yi, & Chun, 2006; Gagnepain et al., 2008). Further evidence comes

from the task itself: age judgment consistency predicted recognition accuracy. That is, when a participant judged the same individual more consistently as being younger or older than 30 years of age, regardless of whether appeared with the same or different emotional expression, the more likely they were to remember that individual later. As the stimulus identities were presented with a never-encountered neutral expression in the recognition test, their successful recognition required the formation of a stable expression-independent representation of that individual. Consistently assigning an identity to a given age bracket during repeated encounters, either with the same or different expression, could be taken as evidence, albeit indirect, that such an individual-specific representation was being formed.

Most of the effects discussed here and in the next sections were present in both face and voice modules. Nonetheless, some modality differences were observed. In general, accuracy was higher for faces than voices; this is not surprising given that voice is considered a weaker identity cue and voice recognition is more error-prone (e.g., Stevenage & Neil, 2014; Barsics, 2014; Young et al., 2020; Hanley et al., 1998; Damjanovic & Hanley, 2007). In the case of new identities, in addition to being more accurate, subjects were more confident when correctly identifying a never-encountered individual through their face than their voice. This was reflected by the fact that whereas drift rates for new faces were negative (i.e., towards a "new" response), those for new voices were positive, indicating a bias to consider voices as previously encountered individuals (i.e., a familiarity bias). Nonetheless, the drift rates for true old identities remained larger than those for new ones. Hence the subjects were attempting, despite this bias and with limited success, to correctly identify the individuals as new or old based on their voice and not just following this putative bias.

3.7.2 High arousal interferes with emotion-independent identity memory

Results from the Main Study point, albeit indirectly, towards a relative recognition impairment in performance, reflected in reduced accuracy and less efficiency of memory processing - as indexed by the diffusion model drift rates, argued to represent the strength of memory traces that are used to discriminate between old and new items -, of identity recognition memory for high-arousal emotions in both modalities. This possible relation was then directly tested, and confirmed, by a strong correlation between stimulus-specific recognition accuracy and the corresponding subjective arousal ratings across all emotions and modalities. We further

confirmed this impairment effect in two additional experiments (Follow-up Study 1) comparing, in a within-subject design, high- and low-arousal expressions, namely Fearful vs. Sad (with Neutral recognition) and Fearful vs. Neutral (with Happy recognition). Again, results revealed a significantly reduced recognition accuracy for identities encoded with a high-arousal expression, both for faces and voices, ruling out any possible confounding effects due to potential overall performance differences among *Uni*-emotion groups from the Main Study.

Thus, the overall pattern of results across studies, measures and indices, strongly supports the conclusion of a reduced expression-independent identity recognition, when encoded in high-arousal emotions. This is consistent with the central/peripheral (or intrinsic/extrinsic) trade-off phenomenon mentioned in Introduction (for reviews, see Buchanan & Adolphs, 2002; Kensinger, 2009; Mather & Sutherland, 2011). Following this notion, we suggest that exposure to individuals expressing a high-arousal emotion enhances recognition of that specific identity-emotion combination (i.e., same stimulus), but interferes with the formation of an expression-invariant representation of the individual's identity. This could result in a worse recognition when the identity is presented in a novel expression. Notably, the arousal influence on memory was still present despite the task irrelevance of emotional expression, both during encoding (age judgment) and recognition. This finding fits well with the notion that arousal modulates selective attention towards significant or goal-relevant aspects of stimuli (Mather & Sutherland, 2011), and evidence that arousing (particularly negative) aspects of events capture attention in a relatively automatic manner (e.g., Armony, Vuilleumier, Driver, & Dolan, 2001; Dolan & Vuilleumier, 2003; Sanders et al., 2005). Moreover, these results, and their interpretation, are also consistent with studies reporting better performance in visual detection and memory tasks when in a low-arousal mood state, such as sadness (e.g., Jefferies et al., 2008; Hills, Werno & Lewis, 2011). The proposed hypothesis would not only help explain our current findings, but also integrate them with the seemingly contradictory results in the past showing an enhanced memory for same-stimulus recognition of emotional faces or voices (e.g., Sergerie, Lepage & Armony, 2005, Aubé, Peretz & Armony, 2013). That is, emotional arousal would strengthen the encoding of the core features of an emotional stimulus, thus resulting in a superior recognition of the stimulus per se. However, this would be accomplished at the expense of the encoding of its "emotionally-irrelevant" features, which would then reduce the ability to identify them in a

different exemplar of the same individual. This in turn, would be reflected in a difficulty to successfully generalize individuals to a novel expression, as was the case here.

3.7.3 Multiple exemplar memory advantage compared to repeated high-arousal expressions

Results from the *Multi* condition from the Main Study and Follow-up Study 2 collectively suggest that encoding exemplar variance help overcome the arousal-related weaker identity representation formation, discussed above, even when most (Main Study), or even all (Follow-up Study 2), of the expressions convey high arousal information. This is likely achieved through integration of different exemplars into an expression-independent representation of the individual. Indeed, as mentioned above, a significant priming effect in encoding RTs over repeated individual presentations was found in both *Uni* and *Multi* conditions, and, more importantly, the magnitude of this priming, as well as the degree of age judgment consistency, predicted recognition success during the test phase in which individuals were presented in a never-encountered-before neutral expression. It is important, however, to consider an alternative explanation, namely that this *Multi* advantage was simply due to the presence of sad stimuli in that condition in all groups. Results from the Follow-up Study 2, in which we used a reduced *Multi* condition, without this low-arousal expression, rule out this possibility, by consistently showing the advantage on recognition performance.

Results from the DDM-derived drift rates, integrating response choice and times, suggest facilitated responses towards correctly recognized *Multi* identities. A speeded response has been proposed to reflect an increased confidence in the recognition of previously encountered identities (Robinson, Johnson & Herndon, 1997; Weidemann & Kahana, 2016; Xu & Armony, 2021). Here, we confirmed this hypothesis by including an explicit measure of response confidence (*Sure/Unsure*). As expected, confident responses showed significantly shorter RTs. Furthermore, participants reported being more confident of their response when correctly recognizing *Multi* than *Uni* identities, but only for faces. This lack of significant confidence difference, compared to the DDM results, may suggest that the usage of a binary confidence rating is not precise or sensitive enough, to detect weaker or nosier difference. These findings confirm the validity of using response times as a proxy for response confidence which, in fact, may be more sensitive and efficient than implementing a binary explicit confidence rating.

Overall, our findings are consistent with previously reported memory advantage of encoding variability on identity recognition for faces (Murphy et al., 2015; Ritchie & Burton, 2017; Matthews, Davis, & Mondloch, 2018) and voices (Lavan et al., 2019a; Xu & Armony, 2021). They also extend, and help generalize, those findings in several ways. First, from a methodological standpoint, they confirm that, despite some concerns on data collection and quality (reviewed by Finley & Penningroth, 2015), online memory experiments can yield meaningful results, comparable to the ones obtained from in-person studies, including those related to response times. Indeed, the effect sizes for recognition accuracy as a function of exemplar variability (*Multi* vs. *Uni*) were comparable between studies ($d = 0.17$ for voices of *Uni*-Fearful group, $d = 0.13$ for data from Xu & Armony, 2021). One of the advantages of porting an in-person study to an online platform is that it typically allows for the recruitment of larger sample sizes which, are more likely to yield more robust, statistically significant effects, as it was the case here. Nonetheless, a necessary trade-off is a reduced control over the testing environment. Additionally, our results demonstrate a strong across-modality correspondence in the effects of emotion, both in terms of facilitation and impairment, on identity memory. This goes along with the proposed similarity in voice and face identity processing (see Belin et al., 2011 for a review).

Our findings may also provide some insights into the controversial topic of eye/ear-witness reliability (Magnussen et al., 2010; Sherrin, 2016). In such cases, the suspected criminals are encountered in a few, usually only one, instances, under less-than-ideal conditions. In contrast, the recognition (line-up) test is typically conducted under tightly controlled conditions, requiring the witness to identify the target individual among a group of similar foils (with neutral expressions). Recent research suggests that encoding conditions (e.g., close/far, short/long exposure) have a substantial impact on recognition confidence and accuracy (Molinaro, Charman, & Wylie, 2021). Our results further confirm and extend this observation, showing that emotion-related variables, including expression and variability, also play a significant role in subsequent recognition accuracy and confidence.

3.8 Limitations and Future Directions

As is the case in many psychology studies, when conducted either online or in-person, female participants were over-represented in our sample, and we were therefore unable to explore sex

differences. As some sex- and gender-based individual differences have been reported in emotion and memory studies (e.g., see Armony & Sergerie, 2007; Skuk & Schweinberger, 2013; Herlitz & Lovén, 2013; Patel, Fredborg, & Girard, 2023 for memory research; see Montagne et al., 2005; Filkowski et al., 2017; Kret & de Gelder, 2012 for emotion processing), it is important in future studies to ensure enough diversity to be able to assess whether the observed effects are modulated by these factors.

Whereas we tested the encoding of in Follow-up Study 2 in three different high-arousal emotions (anger, fear and happiness/joy), we only used sadness as a low-arousal expression (as well as neutral ones in the Follow-up Study 1). This was a limitation of the datasets employed which, as most available ones, restrict the set of emotions to so-called basic ones (i.e., the ones mentioned here, plus disgust and, sometimes, surprise). Although our conclusions are supported by the direct correlation analysis between recognition accuracy and arousal scores across emotions and modalities, it would be important to further confirm our findings with other low-arousal expressions. This could be achieved by employing other emotions (e.g., contempt) and/or through morphing procedures.

We tested recognition in each modality separately, thus our results cannot directly speak to the question of emotional memory for multimodal stimuli. Nonetheless, given that we observed similar patterns for face and voice, we can speculate that these effects would also hold for audio-visual expressions, although this remains to be confirmed. Such studies would be an important contribution to the memory literature, which reports largely conflicting observations. For example, different studies have found the simultaneous presentation of face and voice stimuli help (e.g., Maguinness, Schall, & von Kriegstein, 2021; von Kriegstein et al., 2008; Zäske, Mühl, & Schweinberger, 2015) and hinder (e.g., Lavan et al., 2023; Cook & Wilding, 2001; Tomlin, Stevenage, & Hammond, 2017) voice identity learning and recognition.

3.9 Conclusion

We investigated how emotion influences face- and voice-based recognition of individuals subsequently encountered with a different expression. Specifically, our findings demonstrated that encoding identities with high-arousal expressions hindered subsequent different-expression recognition for both modalities. However, this effect could be overcome by the presentation of multiple exemplars with different expressions during encoding. In fact, these opposing effects

led to a similar recognition accuracy between repeated presentation of a single low-arousal expression (e.g., sad) and multiple high-arousal ones (e.g., angry, fearful, and happy). To illustrate these findings, we can return to the movie example mentioned in the Introduction; imagine that, back in the 60's, two friends — one of whom saw *Psycho* several times while the other, in addition, watched *Touch of Evil* and *Holiday Affair* — ran into Janet Leigh when she was ordering coffee at the local diner. According to our results, the fan who saw Ms. Leigh starring in different films should have been more likely to recognize her, and therefore get her autograph.

3.10 References

Ack Baraly, K., Hot, P., Davidson, P., & Talmi, D. (2016). How Emotional Arousal Enhances Episodic Memory. In J. H. Byrne (Ed.), Learning and Memory: A Comprehensive Reference (2nd ed.). Elsevier BV

Armony, J. L., & Sergerie, K. (2007). Own-sex effects in emotional memory for faces. *Neuroscience Letters*, *426*(1), 1-5. https://doi.org/10.1016/j.neulet.2007.08.032

Armony, J. L., Chochol, C., Fecteau, S., & Belin, P. (2007). Laugh (or cry) and you will be remembered: Influence of Emotional Expression on Memory for Vocalizations. *Psychological Science, 18*(12), 1027–1029. https://doi.org/10.1111/j.1467-9280.2007.02019.x

Armony, J. L., Vuilleumier, P., Drive, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron, 30*(3), 829-841. https://doi.org/10.1016/S0896-6273(01)00328-2

Aubé, W., Peretz, I., & Armony, J. L. (2013). The effects of emotion on memory for music and vocalizations. *Memory, 21*(8), 981-990. https://doi.org/10.1080/09658211.2013.770871

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244–254. https://doi.org/10.5334/pb.ap

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*, 110-120. https://doi.org/10.1007/s00426-008-0185-z

Belin, P, Bestelmeyer, P. E. G., Latinus, M., & Watson R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711-725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Benson, P. J., & Perrett, D. I. (1993). Extracting prototypical facial images from exemplars. *Perception*, *22*(3), 257-262. https://doi.org/10.1068/p220257

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program].

Version 6.1.04, retrieved 28 September 2019 from http://www.praat.org/

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Buchanan, T., & Adolphs, R. (2002). The role of the human amygdala in emotional modulation of long-term declarative memory. In S. Moore & M. Oaksford (eds.), *Emotional cognition: From brain to behavior*. London, UK: John Benjamins.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Cahill, L., Prins, B., Weber, M., *et al.* (1994). β-Adrenergic activation and memory for emotional events. *Nature, 371*, 702–704. https://doi.org/10.1038/371702a0

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92*(4), 617-629. https://doi.org/10.1348/000712601162374

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition, 35*, 1205–1210. https://doi.org/10.3758/BF03193594

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

Dolan, R. J., & Vuilleumier, P. (2003). Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences*, *985*, 348–355. https://doi.org/10.1111/j.1749-6632.2003.tb07093.x

Dolcos, F., Yuta Katsumi, Y., Moore, M., et al. (2020). Neural correlates of emotion-attention interactions: From perception, learning, and memory to social cognition, individual differences, and training interventions. *Neuroscience & Biobehavioral Reviews*, *108*, 559-601. https://doi.org/10.1016/j.neubiorev.2019.08.017.

Filkowski, M. M., Olsen, R. M., Duda, B., Wanger, T. J., & Sabatinelli, D. (2017). Sex differences in emotional perception: Meta analysis of divergent activation. *NeuroImage, 147*, 925-933. https://doi.org/10.1016/j.neuroimage.2016.12.016

Finley A. J., & Penningroth, S. L. (2015). Online versus in-lab: pros and cons of an online prospective memory experiment. In A. M. Columbus (Ed.), *Advances in Psychology Research*, vol. 113 (pp. 135-162). Hauppauge, NY: Nova Science Publishers, Inc.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd edition. Sage, Thousand Oaks, CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gagnepain, P., Lebreton, K., Desgranges, B., & Eustache, F. (2008). Perceptual priming enhances the creation of new episodic memories. *Consciousness and Cognition, 17*(1), 276–287. https://doi.org/10.1016/j.concog.2007.03.006

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accent in communication. *Personality and Social Psychology Review, 14*(2), 214-237. https://doi.org/ 10.1177/1088868309359288

Goeleven, E., de Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion*, *22*(6), 1094-1118. https://doi.org/10.1080/02699930701626582

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of *familiar-only* experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 51A*(1), 179–195. https://doi.org/10.1080/027249898391819

Herlitz, A. & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: a meta-analytic review. *Visual Cognition, 21*(9-10), 1306-1336. https://doi.org/10.1080/13506285.2013.823140

Hills, P. J., Werno, M. A., & Lewis, M. B. (2011). Sad people are more accurate at face recognition than happy people. *Consciousness and Cognition, 20*(4), 1502-17. https://doi.org/10.1016/j.concog.2011.07.002

Jefferies, L. N., Smilek, D., Eich, E., & Enns, J. T. (2008). Emotional valence and arousal interact in attentional control. *Psychological Science, 19*(3), 290-295. https://doi.org/10.1111/j.1467-9280.2008.02082.x

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, *15*(4), 241-252. https://doi.org/10.1515/REVNEURO.2004.15.4.241

Kensinger, E. A. (2007). Negative emotion enhances memory accuracy: behavioral and neuroimaging evidence. *Current Directions in Psychological Science, 16*(4), 213-218. https://doi.org/10.1111/j.1467-8721.2007.00506.x

Kensinger, E. A. (2009). Remembering the details: effects of emotion. *Emotion Review*, *1*(2), 99-113. https://doi.org/10.1177/1754073908100432

Kensinger, E. A., & Schacter, D. L. (2005). Retrieving accurate and distorted memories: Neuroimaging evidence for effects of emotion. *NeuroImage, 27*(1), 167-177. https://doi.org/10.1016/j.neuroimage.2005.03.038

Kret, M. E., & de Gleder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia, 50*(7), 1211-1221. https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13.

LaBar, K., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7***,** 54–64. https://doi.org/10.1038/nrn1825

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE*, *10(*7): e0134073. https://doi.org/10.1371/journal.pone.0134073

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology, 23*(12), 1075-1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019b). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240–2248. https://doi.org/10.1177/1747021819836890

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019a). The effects of high variability training on voice identity learning. *Cognition, 193*, 104026.

https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Ramanik Bamaniya, N., Muse, M. M., et al. (2023). The effects of the presence of a face and direct eye gaze on voice identity learning. *British Journal of Psychology*, *114*(3), 537–549. https://doi.org/10.1111/bjop.12633

Lenth, R. V. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Li, B, & Jiang, L. (2020). The relationship between perceptual priming and subsequent recognition memory: an event-related potential study. *NeuroReport, 31*(17), 1175-79. https://doi.org/10.1097/WNR.0000000000001533

Liu, C. H., Chen, W. F., & Ward, J. (2014). Remembering faces with emotional expressions. *Frontiers in Psychology*, *5*, 1439. https://doi.org/10.3389/fpsyg.2014.01439.

Liu, C. H., Chen, W. F., & Ward, J. (2015). Effects of exposure to facial expression variation in face learning and recognition. *Psychological Research, 79*(6), 1042-53. https://doi.org/10.1007/s00426-014-0627-8

Liu, C. H., Chen, W. F., Ward, J., & Takahashi, N. (2016). Dynamic emotional faces generalize better to new expression but not to a new view. *Scientific Reports*, *6*, 31001. https://doi.org/10.1038/srep31001

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77–100. https://doi.org/10.1037/0096-1523.34.1.77

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149-157. https://doi.org/10.1037/0278-7393

Lundqvist, D., Bruce, N., & Öhman, A. (2015) Finding an emotional face in a crowd: Emotional and perceptual stimulus factors influence visual search efficiency. *Cognition and Emotion*, *29*(4), 621-633. https://doi.org/10.1080/02699931.2014.927352

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Magnussen, S., Melinder, A., Stridbeck, U., & Raja, A. Q. (2010). Beliefs about factors affecting the reliability of eyewitness testimony: A comparison of judges, jurors and the general public. *Applied Cognitive Psychology, 24*(1), 122–133. https://doi.org/10.1002/acp.1550

Maguinness, C. , Schall, S. , & von Kriegstein, K. (2021). Prior audio-visual learning facilitates auditory-only speech and voice-identity recognition in noisy listening conditions. *PsyArXiv* . 10.31234/osf.io/gc4xa

Mather, M., & Sutherland M. (2011). Arousal-biased competition in perception and memory. *Perspectives on Psychological Science*, *6*(2), 114-133. https://doi.org/10.1177/1745691611400234

Matthews, C. M., Davis, E. E., & Mondloch, C. J. (2018). Getting to know you: the development of mechanisms underlying face learning. *Journal of Experimental Child Psychology*, *167*, 295-313. https://doi.org/10.1016/j.jecp.2017.10.012

McGaugh J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual review of neuroscience*, *27*, 1–28. https://doi.org/10.1146/annurev.neuro.27.070203.144157

Miyoshi, K., Minamoto, T., & Ashida, H. (2014). Relationships between priming and subsequent recognition memory. *SpringerPlus*, *3*(546). https://doi.org/10.1186/2193-1801-3-546

Mogg, K., & Bradley, B. P. (1999). Selective attention and anxiety: A cognitive–motivational perspective. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 145–170). John Wiley & Sons.

Molinaro, P. F., Charman, S. D., & Wylie, K. (2021). Pre-identification confidence is related to eyewitness lineup identification accuracy across heterogeneous encoding conditions. *Law and human behavior*, *45*(6), 524–541. https://doi.org/10.1037/lhb0000452

Montagne, B., Kessels, R. P. C., Frigerio, E., de Haan, E. H. F., & Perrett, D. I. (2005). Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? *Cognitive Processing, 6,* 136-141. https://doi.org/10.1007/s10339-005-0050-6

Mueller, C.J., & Kuchinke, L. (2016). Individual differences in emotion word processing: A diffusion model analysis. *Cogn Affect Behav Neurosci, 16*, 489–501.

https://doi.org/10.3758/s13415-016-0408-5

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-581. https://doi.org/10.1037/xhp0000049

Nomi, J. S., Rhodes, M. G., & Cleary, A. M. (2013). Emotional facial expressions differentially influence predictions and performance for face recognition. *Cognition and Emotion, 27*(1), 141-149. https://doi.org/10.1080/02699931.2012.679917

Norris, C. J. (2021). The negativity bias, revisited: Evidence from neuroscience measures and an individual differences approach. *Social Neuroscience, 16*(1), 68-82. https://doi.org/10.1080/17470919.2019.1696225

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences, 6*, 261–266. https://doi.org/10.1016/S1364-6613(02)01908-3

Patel, R., Fredborg, B. K., & Girard, T. A. (2023). Modulation of emotion-enhanced recollection by gender and task instructions. *Emotion, 23*(6), 1764–1772. https://doi.org/10.1037/emo0001196

Pichora-Fuller, M.K., Dupuis, K., & Smith, L. (2016). Effects of vocal emotion on memory in younger and older adults. *Experimental Aging Research, 42*(1), 14-30. https://doi.org/10.1080/0361073X.2016.1108734.

Qasim, S., E., Mohan, U.R., Stein, J.M. et al. (2023). Neuronal activity in the human amygdala and hippocampus enhances emotional memory encoding. *Nat. Hum. Behav., 7*, 754–764. https://doi.org/10.1038/s41562-022-01502-8

Ratcliff R, McKoon G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput, 20*(4), 873-922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Righi, S., Marzi, T., Toscani, M., et al. (2012). Fearful expressions enhance recognition memory: Electrophysiological evidence. *Acta Psychologica, 139*(1), 7-18. https://doi.org/10.1016/j.actpsy.2011.09.015

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 879-895. https://doi.org/10.1080/17470218.2015.1136656

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*(3), 416–425. https://doi.org/10.1037/0021-9010.82.3.416

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sanders, D., Grandjean, D., Pourtois, G., et al. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *NeuroImage, 28*(4), 848-58. https://doi.org/10.1016/j.neuroimage.2005.06.023

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65*(1), 111-6. https://doi.org/10.1037/0021-9010.65.1.111

Sergerie, K., Lepage, M., & Armony, J. L. (2005). A face to remember: emotional expression modulates prefrontal activity during memory formation. *NeuroImage, 24*(2), 580-5. https://doi.org/10.1016/j.neuroimage.2004.08.051

Sergerie, K., Lepage, M., & Armony, J. L. (2006). A process-specific functional dissociation of the amygdala in emotional memory. *Journal of Cognitive Neuroscience, 18*(8), 1359-67. https://doi.org/10.1162/jocn.2006.18.8.1359

Shaw, J. S. III, McClure, K. A., & Wilkens, C. E. (2001). Recognition instructions and recognition practice can alter the confidence–response time relationship. *Journal of Applied Psychology*, *86*(1), 93-103. https://doi.org/10.1037/0021-9010.86.1.93

Sherrin, C. (2016). Earwitness Evidence: The Reliability of Voice Identifications. *Osgoode Hall Law Journal, 52*(3), 819-862.

Skuk, V. G. & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research, 296*, 131-140. https://doi.org/10.1016/j.heares.2012.11.004

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*(3), 266-281. http://dx.doi.org/10.5334/pb.ar

Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance ratings for facial stimuli. *Quarterly Journal of Experimental Psychology, 72*(8), 2046–2055.

https://doi.org/10.1177/1747021819829012

Talmi, D., Anderson, A. K., Riggs, L., et al. (2008). Immediate memory consequences of the effect of emotion on attention to pictures. *Learning & Memory*, *15*(3), 172-182. https://10.1101/lm.722908

Tomlin, R. J., Stevenage, S. V., & Hammond, S. (2017). Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition, 25*(4-6), 629–643. https://doi.org/10.1080/13506285.2016.1245230

Turk-Browne, N. B., Yi, D. J., & Chun, M. M. (2006). Linking implicit and explicit memory: common encoding factors and shared representations. *Neuron*, *49*(6), 917-27. https://doi.org/10.1016/j.neuron.2006.01.030

Unkelbach, C., Alves, H., & Koch, A. (2020). Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 62, pp. 115–187). Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2020.04.005

Vandekar, S., Tao, R., & Blume, J. (2020). A Robust Effect Size Index. *Psychometrika*, *85*(1), 232–246. https://doi.org/10.1007/s11336-020-09698-2

Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time data: Theory and implications. *Psychological Review, 123*(2), 208–218. https://doi.org/10.1037/rev0000019

von Kriegstein, K., Dogan, O., Grüter, M., et al. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6747–6752. https://doi.org/10.1073/pnas.0710826105

Vuilleumier, P, & Huang, Y. (2009). Emotional attention: Uncovering the mechanisms of affective biases in perception. *Current Directions in Psychological Science, 18*(3), 148-152. https://doi.org/10.1111/j.1467-8921.2009.01626.x

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science, 3*(4), 150670. https://doi.org/10.1098/rsos.150670

Williams, W. C., Haque, E., Mai, B., & Venkatraman, V. (2023). Face masks influence emotion judgments of facial expressions: a drift–diffusion model. *Scientific Reports, 13*, 8842.

https://doi.org/10.1038/s41598-023-35381-4

Xiao, N. G., Perrotta, S., Quinn, P. C., et al. (2014). On the facilitative effects of face motion on face recognition and its development. *Frontiers in Psychology,* ***5***(633). https://doi.org/10.3389/fpsyg.2014.00633.

Xu, H., & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology, 74*(7), 1185-1201. https://doi.org/10.1177/1747021821998557

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in cognitive sciences*, *24*(5), 398-410. https://doi.org/10.1016/j.tics.2020.02.001

Zäske, R. , Mühl, C. , & Schweinberger, S. R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLoS One, 10*(11), 1–12. https://doi.org/10.1371/journal.pone.0143151

3.11 Supplementary Information

Methods

Stimuli

For the face stimuli used in the Main and two Follow-up studies, 48 KDEF actors were selected based on the highest emotion recognition rates (ER rates, data taken from Appendix 2 in Geoleven et al., 2008) of the four emotion categories used in our studies (i.e., angry, fearful, happy, and sad). The actor list was shown below. The descriptive stats of the unbiased hit rates of emotion recognition (Geoleven et al., 2008), as well as valence and arousal ratings (Sutton, Herbert & Clark, 2019) of selected stimuli are summarized in S.Table 3-1.

Female: AF01, AF02, AF03, AF05, AF06, AF07, AF08, AF09, AF11, AF13, AF16, AF20, AF21, AF23, AF25, AF26, AF27, AF28, AF29, AF30, AF32, AF33, AF34, AF35.

Male: AM01, AM03, AM05, AM06, AM08, AM09, AM10, AM11, AM12, AM13, AM15, AM16, AM17, AM18, AM19, AM22, AM23, AM24, AM25, AM27, AM28, AM29, AM31, AM35.

S.Table 3-1

Descriptive statitics of unbiased emotion recognition (ER) hit rates, ratings of valence and arousal from the selected face stimuli.

Emotion	Faces (KDEF)			Voices (RAVDESS)
	Unbiased ER hit[+]	Valence[*]	Arousal[*]	Arousal[++]
Angry	0.28 (0.18)	1.85 (0.56)	4.97 (0.87)	4.44 (0.36)
Fearful	0.59 (0.30)	2.08 (0.60)	5.11 (1.05)	4.06 (0.63)
Happy	0.92 (0.25)	6.61 (0.52)	4.87 (0.72)	3.61 (0.50)
Sad	0.54 (0.15)	1.81 (0.50	3.69 (0.89)	3.57 (0.55)

Format: Mean (Standard Deviation);

+: data from Geoleven et al., 2008;

*: data from Sutton, Herbert & Clark, 2019, rating ranged 1-9.

++: data from Livingstone & Russo, 2018, rating ranged 1-5.

Practice Session

Stimuli

Selected face images with four emotional expressions (fearful, angry, sad and happy) and neutral expression from the Japanese Female Facial Expression dataset (JAFFE, Lyons, Kamachi, & Gyoba, 1998) were used. Specifically, we used the first exemplar of each emotional category of actors coded KA, TM, UY, and NM. The images were preprocessed in the same fashion as the experimental stimuli (e.g., exterior face removal and size/resolution adjustment).

German speech clips were selected from the Berlin Database of Emotional Speech (EMODB, Burkhardt et al., 2005), including four speakers (09 [F], 10 [M], 13 [F] and 15 [M]) uttering the same short sentence (a07) in fear, disgust, happiness, boredom prosodies. As the database lacked one speaker's neutral sample uttering the a07 sentence, we adapted to use a different sample sentence (a01/02/04) in the neutral tone from the four speakers as the recognition test stimuli. Loudness of the speech clips were adjusted to be at a comparable level as the RAVDESS samples.

Procedure

To ensure that they fully understood the task, participants underwent a practice session right before the actual experiment, which was a shortened version of both modules, using different stimulus sets of faces and voices (see above). Particularly, only 2 identities were presented in the encoding phase for each module, one in the *Multi* and the other in the *Uni* condition (8 trials in total). Then four identities (2 old, 2 new) were tested in the recognition phase using exemplars in a neutral expression or prosody. At the end of each trial after both the recognition response and confidence rating, a correct/incorrect text feedback was given on the screen. Participants would repeat the practice session until reaching a 100% accuracy for the recognition tests, in order to enter the actual experiment.

S.Table 3-2

Descriptive statitics of recognition measures from Main and two Follow-up studies.

Participant Group	**Auditory**				**Visual**			
	Main Study							
	Accuracy							
	Multi	*Uni*	New	Overall	*Multi*	*Uni*	New	Overall
Uni-Angry	0.66	0.60	0.45	0.54	0.78	0.64	0.67	0.69
(N = 129)	(0.21)	(0.22)	(0.17)	(0.11)	(0.15)	(0.18)	(0.14)	(0.10)

Uni-Fearful	0.66	0.61	0.44	0.54	0.75	0.70	0.66	0.70
(N = 145)	(0.21)	(0.22)	(0.16)	(0.10)	(0.16)	(0.16)	(0.15)	(0.09)
Uni-Happy	0.64	0.59	0.46	0.54	0.74	0.69	0.68	0.70
(N = 144)	(0.21)	(0.20)	(0.17)	(0.09)	(0.16)	(0.17)	(0.14)	(0.09)
Uni-Sad	0.64	0.65	0.46	0.55	0.72	0.72	0.68	0.70
(N = 129)	(0.23)	(0.19)	(0.17)	(0.10)	(0.16)	(0.17)	(0.14)	(0.09)
				Drift Rates				
Uni-Angry	0.48	0.30	0.10	-	1.33	0.72	-0.79	-
(N = 129)	(0.71)	(0.70)	(0.53)		(0.99)	(1.00)	(0.83)	
Uni-Fearful	0.40	0.26	0.12	-	1.11	0.90	-0.81	-
(N = 145)	(0.76)	(0.83)	(0.55)		(0.97)	(0.93)	(0.85)	
Uni-Happy	0.48	0.32	0.05	-	1.17	0.88	-0.80	-
(N = 144)	(0.77)	(0.74)	(0.32)		(0.97)	(0.98)	(0.81)	
Uni-Sad	0.48	0.46	0.14	-	0.97	0.88	-0.86	-
(N = 129)	(0.79)	(0.65)	(0.59)		(1.16)	(1.20)	(0.77)	
				Follow-up Study 1				
				Accuracy				
	High	Low	New	Overall	High	Low	New	Overall
Fearful vs. Sad	0.55	0.59	0.50	0.54	0.64	0.69	0.72	0.69
(N=86)	(0.24)	(0.20)	(0.16)	(0.11)	(0.15)	(0.15)	(0.14)	(0.09)
Fearful vs Neutral	0.55	0.59	0.49	0.53	0.64	0.70	0.68	0.68
(N=86)	(0.21)	(0.22)	(0.18)	(0.11)	(0.16)	(0.14)	(0.13)	(0.09)
				Drift Rates				
Fearful vs. Sad	0.13	0.27	-0.03	-	0.59	0.80	-0.93	-
(N=86)	(0.71)	(0.63)	(0.51)		(0.79)	(0.81)	(0.64)	
Fearful vs Neutral	0.13	0.26	0.01	-	0.60	0.84	-0.77	-
(N=86)	(0.73)	(0.72)	(0.56)		(0.82)	(0.79)	(0.76)	
				Follow-up Study 2				
				Accuracy				
	Multi	*Uni*	New	Overall	*Multi*	*Uni*	New	Overall
Uni-Angry	0.60	0.54	0.48	0.53	0.69	0.60	0.64	0.64
(N=87)	(0.20)	(0.23)	(0.18)	(0.11)	(0.16)	(0.16)	(0.14)	(0.08)
Uni-Fearful	0.61	0.57	0.47	0.53	0.71	0.66	0.65	0.67
(N=87)	(0.21)	(0.21)	(0.18)	(0.10)	(0.16)	(0.15)	(0.12)	(0.09)
				Drift Rates				
Uni-Angry	0.19	0.052	0.025	-	0.87	0.43	-0.60	-
(N=87)	(0.67)	(0.72)	(0.54)		(0.88)	(0.78)	(0.58)	

Uni-Fearful (N=87)	0.30 (0.65)	0.11 (0.72)	0.037 (0.57)	-	0.96 (0.90)	0.78 (0.89)	-0.64 (0.61)	-

* Format: Mean (Standard Deviation)

S.Table 3-3

Fixed factor effects from (G)LMM results on recongition accuracy and drift rates from the Main Study.

Fixed Effects	Accuracy				Fixed Effects	Drift Rate (v)			
	χ^2	Df	p	S		(Df1, Df2)	F	p	S
Old-identity (G)LMMs									
Var	6.60	1	.010	0.10	**Var**	1, 1581.1	39.03	<.001	0.26
Emo	2.37	3	.50	< 0.001	**Emo**	3, 528.1	0.17	.91	< 0.001
Mod	7.54	1	.006	0.11	**Mod**	1, 1583.5	300.94	<.001	0.74
Var × Emo	11.25	3	.010	0.12	**Var × Emo**	3, 1579.6	4.31	.005	0.13
Var × Mod	2.35	1	.13	0.050	**Var × Mod**	1, 1583.4	6.95	.008	0.10
Emo × Mod	2.05	3	.56	< 0.001	**Emo × Mod**	3, 1583.4	1.96	.12	0.072
Var × Emo × Mod	2.21	3	.53	< 0.001	**Var × Emo × Mod**	3, 1579.5	1.61	.19	0.058
Slope	8.51	1	.004	0.12					
ACS	23.17	1	< .001	0.20					
ACS × Var	3.14	1	.076	0.063					
ACS × Emo	2.68	3	.44	< 0.001					
ACS × Mod	0.004	1	.95	< 0.001					
Slope × Var	0.59	1	.44	< 0.001					
Slope × Emo	7.66	3	.054	0.092					
Slope × Mod	0.070	1	.79	< 0.001					
ACS × Var × Emo	3.59	3	.31	0.033					
ACS × Var × Mod	0.62	1	.43	< 0.001					
ACS × Mod × Emo	2.24	3	.52	< 0.001					
Slope × Var × Emo	3.44	3	.33	0.028					
Slope × Var × Mod	0.37	1	.54	< 0.001					
Slope × Mod × Emo	4.40	3	.22	0.050					
ACS × Mod × Emo × Var	0.47	3	.93	< 0.001					
Slope × Mod × Emo × Var	1.87	3	.60	< 0.001					
New-identity (G)LMMs									

Emo	2.15	3	.54	<.001	**Emo**	3, 536.8	0.09	.97	< 0.001
Mod	37.12	1	< .001	0.28	**Mod**	1, 535.4	533.38	<.001	0.98
Emo × Mod	0.75	3	.86	<.001	**Emo × Mod**	3, 535.4	0.70	.55	< 0.001

Abbreviation: Var = variability, Emo = *Uni*-emotion, Mod = modality, Slope = encoding response time slope, ACS = age consistency score.

S.Table 3-4

Post-hoc pairwise comparisons on the variability-by-*Uni*-emotion interaction from the (G)LMM on old-identity trials of the Main Study.

Post-hoc tests	***b***	**SE**	**|*z*| or |*t*|(df)**	***p***	***Effect size***
		Accuracy GLMM			
Contrast: Multi > Uni		**Var × Emo interaction**			**Odds ratio**
Uni-Angry	0.53	0.07	7.49	< .001	1.70
Uni-Fear	0.29	0.07	3.82	.003	1.29
Uni-Happy	0.29	0.07	4.29	.003	1.34
Uni-Sad	-0.003	0.07	0.05	.96	1.00
		Drift rate LMM			
Contrast: Multi > Uni		**Var × Emo interaction**			**Cohen's *d***
Uni-Angry	0.41	0.071	5.77 (1594)	< .001	0.51
Uni-Fear	0.18	0.067	2.62 (1591)	.009	0.22
Uni-Happy	0.23	0.067	3.36 (1590)	.001	0.28
Uni-Sad	0.05	0.072	0.76 (1603)	.45	0.068

S.Table 3-5

Fixed factor effects from the recognition confidence GLMMs from the Main Study.

Fixed Effects	χ^2	**Df**	***p***	**S**
		Old-identity GLMM		
Var	0.20	1	.65	< 0.001
Emo	0.18	3	.98	< 0.001
Acc	21.79	1	< .001	0.19
Mod	11.06	1	.001	0.14
Var × Emo	2.53	3	.47	< 0.001
Var × Acc	0.16	1	.69	< 0.001
Emo × Acc	1.54	3	.67	< 0.001

Var × Mod	4.62	1	.032	0.081
Emo × Mod	1.50	3	.68	< 0.001
Acc × Mod	11.51	1	< .001	0.14
Var × Emo × Acc	1.30	3	.73	< 0.001
Var × Emo × Mod	3.75	3	.29	0.037
Var × Acc × Mod	6.08	1	.014	0.096
Emo × Acc × Mod	1.42	3	.70	< 0.001
Var × Emo × Acc × Mod	4.51	3	.21	0.053
log-RT	722.20	1	< .001	1.15
	New-identity GLMM			
Emo	0.96	3	.81	< 0.001
Acc	11.56	1	< .001	0.14
Mod	67.14	1	< .001	0.35
Emo × Acc	0.95	3	.81	< 0.001
Emo × Mod	4.14	3	.25	0.046
Acc × Mod	39.63	1	< .001	0.27
Emo × Acc × Mod	1.54	3	.67	< 0.001
log-RT	742.64	1	< .001	1.16

Abbreviation: Var = variability, Emo = *Uni*-emotion, Mod = modality, Acc = accuracy.

S.Table 3-6

Descriptive statistics of two cosine similarity indices of physical features of speech and face stimuli.

Stimulus Emotion	**Voice (x 10^{-3})**		**Face (x 10^{-1})**	
	Within-emotion similarity	***Emo-to-Neu identity distinctiveness***	***Within-emotion similarity***	***Emo-to-Neu identity distinctiveness***
Neutral	998.88 (0.29)	-	0.47 (0.98)	-
Angry	998.24 (0.60)	0.67 (0.54)	0.92 (1.20)	0.45 (0.35)
Fearful	998.05 (0.55)	0.86 (0.56)	0.78 (1.15)	0.46 (0.38)
Happy	998.46 (0.57)	0.80 (0.52)	1.07 (1.19)	0.42 (0.35)
Sad	998.30 (0.64)	0.70 (0.45)	0.79 (1.10)	0.44 (0.38)

S.Table 3-7

Post-hoc pairwise comparisons on the emotional expression main effect of the LMM on stimulus-based cosine similarity of physical features.

Post-hoc pairs	Voice					Face				
	b (x 10^{-4})	SE (x 10^{-4})	|*t*| *(df=92)*	*p*	Cohen's *d*	*b* (x 10^{-2})	SE (x 10^{-2})	|*t*| *(df=188)*	*p*	Cohen's *d*
neu - hap	3.63	1.37	2.66	.093	0.77	-6.01	1.98	3.04	.027	0.62
neu - sad	5.27	1.37	3.86	.002	1.11	-3.19	1.98	1.61	.99	0.33
neu - ang	5.91	1.37	4.32	< .001	1.25	-4.51	1.98	2.28	.24	0.46
neu - fea	7.81	1.37	5.71	< .001	1.65	-3.16	1.98	1.60	.99	0.33
hap - sad	1.64	1.37	1.20	.99	0.35	2.82	1.98	1.43	.99	0.29
hap - ang	2.28	1.37	1.67	.99	0.48	1.50	1.98	0.76	.99	0.15
hap - fea	4.18	1.37	3.06	.029	0.88	2.85	1.98	1.44	.99	0.29
sad - ang	6.39	1.37	0.47	.99	0.14	-1.32	1.98	0.67	.99	0.14
sad - fea	2.54	1.37	1.86	.67	0.54	0.03	1.98	0.02	.99	0.003
ang - fea	1.90	1.37	1.39	.99	0.40	1.35	1.98	0.68	.99	0.14

Abbreviation: neu = neutral, hap = happy, ang = angry, fea = fearful.

S.Table 3-8

Fixed factor effects from the stimulus-level arousal rating model.

Fixed effects	Df1, Df2	*F*	*p*	S
Arousal	1, 269	22.15	<.001	0.27
Emo	3, 269	1.09	.36	0.020
Mod	1, 269	17.73	<.001	0.22
Arousal × Emo	3, 269	0.23	.87	< 0.001
Arousal × Mod	1, 269	0.007	.94	< 0.001
Emo × Mod	3, 269	1.98	.12	0.11
Arousal × Emo × Mod	3, 269	0.74	.53	< 0.001

Abbreviation: Emo = stimulus emotional expression, Mod = modality.

S.Table 3-9

Fixed factor effects from (G)LMM results on recongition accuracy and drift rates from the Follow-up Study 2.

Fixed Effects	Accuracy				Fixed Effects	Drift Rate (*v*)			
	χ^2	Df	*p*	S		(Df1,	*F*	*p*	S

						Df2)			
Var	20.28	1	<.001	0.33	**Var**	1, 502.1	20.42	<.001	0.33
Emo	2.74	1	.10	0.10	**Emo**	1, 167.1	4.24	.041	0.14
Mod	8.46	1	.006	0.21	**Mod**	1, 503.4	127.12	<.001	0.85
Var × Emo	0.98	1	.33	<0.001	**Var × Emo**	1, 502.1	0.84	.36	<0.001
Var × Mod	1.30	1	.25	0.042	**Var × Mod**	1, 502.1	1.95	.16	.074
Emo × Mod	0.86	1	.35	<0.001	**Emo × Mod**	1, 503.4	1.62	.20	.059
Var × Emo × Mod	0.71	1	.40	<0.001	**Var × Emo × Mod**	1, 502.1	1.73	.19	.064

Abbreviation: Var = variability, Emo = *Uni*-emotion, Mod = modality.

References

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech*, *5*, 1517-1520.

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion, 22*(6), 1094-1118. https://doi.org/10.1080/02699930701626582

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lyons, M., Kamachi, M., & Gyoba, J. (1998). The Japanese Female Facial Expression (JAFFE) Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3451524

Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance ratings for facial stimuli. *Quarterly Journal of Experimental Psychology, 72*(8), 2046–2055. https://doi.org/10.1177/1747021819829012

Connecting Chapters 3 to 4

In Chapter 3, a classic encoding-recognition two-stage paradigm was implemented to examine the effect of encoding exemplar variance on identity memory (e.g., Ritchie & Burton, 2017). By comparing behavioral performance between two encoding condition, we observed a strong exemplar variance advantage on identity recognition across modalities in three groups of participants. In addition, we uncovered significant associations between implicit measures of identity memory in encoding and explicit measure of identity recognition. Specifically, the more consistent and faster participants responded to the same identity in encoding, the more likely the identity would be recognized in recognition. As we argued, this implicit-explicit memory relation (Turk-Browne, Yi, & Chun, 2006; Gagnepain et al., 2008) provided indirect support for an identity-specific representation was being formed in encoding, when processing the same identity with variable emotional expressions. In Chapter 4, this claim would be examined directly. To explicitly test whether participants are able to recognize a previously encountered individual with a different emotional expression while learning, a modification of the experimental paradigm became imperative. Accordingly, we adapted the original two-stage recognition test into a single-stage continuous recognition paradigm, which required participants to make old/new judgements since the beginning. In this way, people performed the recognition task while learning the presented identities at the same time. Additionally, we withdrew the *Uni* condition in Chapter 3 from the paradigm, since the current research interest focused particularly on the *Multi* condition. Lastly, a repetition of each novel exemplar was added into the experimental sequence, in order to (1) provide a reference of same-stimulus recognition performance, and (2) serve as a complementary way to boost participants' familiarity of the identities (e.g., Bonner, Burton & Bruce, 2003). With the modified paradigm, Study 3 was able to examine whether explicit recognition and identity integration take place upon encountering novel or repeated emotional exemplars of unfamiliar faces and voices, not only behaviorally, but also at a neural level.

Chapter 4. Cross-emotional-expression recognition in unfamiliar faces and voices – an fMRI study

(Study 3)

Hanjian Xu[1,2,3], Jorge L. Armony[1,2,4]

[1]Douglas Mental Health University Institute, Verdun, Canada;
[2]BRAMS Laboratory, Centre for Research on Brain, Language and Music, Montreal, Canada;
[3]Integrated Program in Neuroscience, McGill University, Montreal, Canada;
[4]Department of Psychiatry, McGill University, Montreal, Canada

In prep.

4.1 Abstract

Recognition of unfamiliar faces and voices is challenging and error-prone. Recent findings indicate that learning identities with more within-person variance, even from emotional expressions alone, can, to some extent, reduce the susceptibility to perceptual variance in face and voice, hence improve recognition and generalization of the learned identities. Yet, little attention has been directed towards the learning process, particularly regarding how people explicitly perceive face and voice identities when encountering novel emotional exemplars. In this fMRI study, we examined the behavioral and neural responses towards novel and repeated emotional exemplars of same identities, in a continuous recognition task, separately for both visual and auditory modalities. For faces, behavioral results revealed a poor recognition of the second novel exemplar, but an improved cross-expression recognition on the third novel exemplar. Bilateral anterior insula and supplementary motor area/anterior cingulate cortex, as part of the Salience Network, exhibited a smaller activation during successful recognition of the third novel exemplar. This may reflect a less effortful, or easier cross-expression recognition of the third novel exemplar for faces. However, no significant behavioral improvement or neural activity difference was found in either the second or third cross-expression recognition condition in voice. This may indicate a difficulty of associating speaker identities from multiple novel emotional exemplars.

Key words: cross-emotion recognition; face recognition; voice recognition; emotional expression; fMRI.

4.2 Introduction

It is well acknowledged that people recognize familiar faces with ease and fewer mistakes in daily life and laboratory experiments (e.g., Bruce, 1982; Burton, Bruce, & Hancock, 1999; Johnston & Edmonds, 2009). This powerful ability to recognize known faces stands in stark contrast to our comparatively weak capacity to identify relatively unfamiliar faces (Hancock, Bruce, & Burton, 2000). Past studies have consistently demonstrated the susceptibility of unfamiliar face recognition, in the cases of changes in viewpoint (e.g., Hill, Schyns, & Akamatsu, 1997; Bruce et al., 1999), emotional expression (e.g., Bruce, 1982; Liu, Chen & Ward, 2014), or even image/video quality (e.g., Burton, Wilson, Cowan, & Bruce, 1999; O'Toole et al., 2010). Such a vulnerability is also shown in voice recognition, that recognition (or generalization) of an unfamiliar voice onto novel exemplars deemed difficult (e.g. Saslove & Yarmey, 1980; Zäske et al., 2014).

Furthermore, recent studies indicated that exposure to within-person variance can mitigate the vulnerability of identity perception from exemplar changes, thereby enhancing recognition and generalization of learned identities (Murphy et al., 2015; Ritchie & Burton, 2017 for faces; Lavan et al., 2019 for voices). Our recent work (Xu & Armony, 2021; Xu & Armony, under review) reported similar recognition benefits in the case of encoding high within-person variance from diverse emotional expressions, in both faces and voices. From the view of the classic face (Bruce & Young, 1986), the key to a successful recognition is activation of the constructed face recognition units, which store mental representations of specific individuals upon learning. Learning within-person variance is regarded pivotal for creating or refining a stable face representation through either the averaging or storage of such variant instances (Burton, 2013). This premise is echoed in voice learning (Lavan, Knight, & McGettigan, 2019). Currently, the majority of research that tested the influence of learning variance, has focused on assessing post-learning recognition performance (e.g., Ritchie & Burton, 2017; Lavan et al., 2019), or directly compared behavioral (e.g., Jenkins et al., 2011) and neural differences (e.g., Eger et al., 2005; Rossion et al., 2001; Gobbini & Haxby, 2006 for faces; von Kriegstein & Giraud, 2004; von Kriegstein et al., 2005 for voices) in recognition tasks between unfamiliar and familiarized identities. One intriguing question that remains less investigated, is the learning or encoding stage. More specifically, little is explicitly investigated, for instance, what cognitive processes and neural correlates are involved in processing (multiple) novel exemplars from a previously

encountered identity, and at what point a robust face or voice representation starts to form. Building on our previous work (Chapter 3), which was centered on emotional expression induced within-person variance, this chapter seeks to elucidate how participants explicitly learn/recognize new emotional exemplars of previously encountered faces and voices, at both behavioral and neural levels.

Past behavioral studies have tackled part of the question, through methods such as the classical encoding/recognition task, and identity matching task, where an immediate comparison between two co-presented or sequentially-presented images or voices is made (e.g., Bruce, 1982). Identity matching and identity recognition are two related processes that sometimes displayed a performance correlation (e.g., Fysh, 2018; Robertson et al., 2017), however, different cognitive processes are involved: the matching task is designed to investigate the perception of two unfamiliar identities with no need of long-term memory recruitment, which is essential in identity recognition tasks. Moreover, a matching decision commonly involves a comparison between two stimuli, which makes it practically difficult to investigate participants' explicit identity perception for more than two exemplars. Another approach that previous research employed, was to assess performance of certain tasks towards the end of the learning phase, such as asking participants to estimate total identity numbers (Murphy et al., 2015), accuracy in a forced-choice recognition task after each study phase (e.g., Lavan et al., 2019; Lavan, Knight, & McGettingan, 2019). Yet, these strategies did not precisely delve into the explicit identity upon encountering novel exemplars of old identities during the early stages of face or voice learning.

Past neuroimaging research has also intended to identify neural correlates of feature-invariant face or voice identity processing (i.e., when exemplar change is involved). Studies in the early 2000s provided the important foundation of the contemporary neural models for face processing (Haxby, Holffman & Gobbini, 2000) and voice processing (Belin, Fecteau & Bédard, 2004). The face model proposes that invariant features of face such as face identity, are processed in fusiform face area (FFA), yet changeable aspects of faces are processed in superior temporal sulcus (STS). Voice processing mainly takes place around the bilateral middle and anterior superior temporal gyri/sulci (STG/STS), often referred to as Temporal Voice Areas (TVAs; Pernet et al., 2015). Later studies often use a functional magnetic resonance imaging-adaptation (fMRI-a) paradigm to probe brain regions that are sensitive to identity repetition (or change). For instance, the FFA has shown sensitivity to both changes in identity and expression, while the

occipital face area (OFA) was sensitive to only identity change, in a standard fMRI-a experiment (Xu & Biederman, 2010). Two other fMRI-a experiments with pair-presented or a block-presented voice stimuli with manipulations of voice identity consistency, suggested that vocal identity processing would involve a network of regions, including right mid-STS/STG, superior temporal pole, and inferior frontal cortex, in acoustic-based or identity-based representations of unfamiliar and familiar voices (Belin & Zatorre, 2003; Latinus, Crabbe & Belin, 2011). However, these neuroimaging studies bore the same limitations as in behavioral ones. Specifically, adaptation paradigms are essentially an identity matching task (either an explicit matching, or passive matching without explicit instruction and/or response) that involves an immediate comparison between two perceived faces or voices, without accessing short-term or long-term memory of face or voice identities. The nature of the adaptation design also makes it difficult to pinpoint changes in neural activation that take place in specific exemplars and importantly, over the course of more than two exemplars.

Hence, we aimed to investigate, in this study, how participants explicitly learn unfamiliar identities through novel emotional exemplars, at both behavioral and neural levels. Considering the two concerns mentioned above, we made modifications based on our previous behavioral paradigm (Xu & Armony, under review). Specifically, we modified the original paradigm, a traditional encoding-recognition two-phase memory task, to a single phase continuous recognition task (e.g., Ferris et al., 1980; Buchsbaum et al., 2015), that requires participants to explicitly identify the presented actor's novelty since the beginning. This change allowed us to directly examine the explicit recognition performance when people learn novel emotional exemplars. In addition to multiple same-identity novel exemplars being presented, we included a repetition for each novel exemplar, in order to provide a "reference" of a solid recognition and learning of presented identities, given that simple stimulus repetitions often produce superior recognition and are shown to improve variance-independent identity learning to some extent (e.g., Memon, Hope, & Bull, 2003; Roark et al., 2006). We hypothesized that the explicit recognition for repeated exemplars would be consistently higher than that for novel exemplars. Among novel exemplars, we expected an improvement in performance in the last novel exemplars due to continuously identity learning, which may be reflected in different activities in previously proposed FFA and right mid-STG/S in each modality, respectively. Given the

identical design in both faces and voices, and the similarities lied in face and voice processing, we also expected a similar performance in both modalities.

4.3 Methods

4.3.1 Participants

Thirty-four volunteers were recruited from the Greater Montreal Area to take part in the functional MRI experiment at the Montreal Neurological Institute. Four participants were excluded from both behavioral and fMRI analyses. Three of the exclusions were due to excessive motions and/or falling asleep inside the MRI scanner, while the other one participant was unable to complete the MRI session due to personal reasons. Ultimately, the final analysis included thirty participants (17 female, age: 18-35, M = 22.6, SD = 4.3). All participants were right-handed, and had normal hearing and (corrected-to-) normal vision. None had previously been diagnosed with or treated for mental or neurological disorders. All of them were fluent in English, sixteen of which identifying as native English speakers.

4.3.2 Stimuli

4.3.2.1 Face

The face stimuli used in the study comprised 72 gray-scale photos of 24 Caucasian individuals (half female) depicting fearful, happy, and angry facial expressions in the full-face view (see Figure 4-1), from the A-series of the Karolinska Directed Emotional Faces database (KDEF, Lundqvist, Flykt & Öhman, 1998). The individuals were selected based on the highest hit rates of emotion categorization across the three expressions (Goeleven et al., 2008). The images were preprocessed in Adobe Photoshop 7.0 (Adobe Systems, San Jose, CA) to ensure a uniform face size, contrast and resolution. In addition, exterior parts of faces (e.g., hair, ear, neck) were removed from the images (Sergerie, Lepage & Armony, 2006; 2007) to minimize the influences of external features on recognition (e.g., Ellis, Shepherd & Davies, 1979; Latif & Moulson, 2022; see Johnston & Edmonds, 2009 for a review).

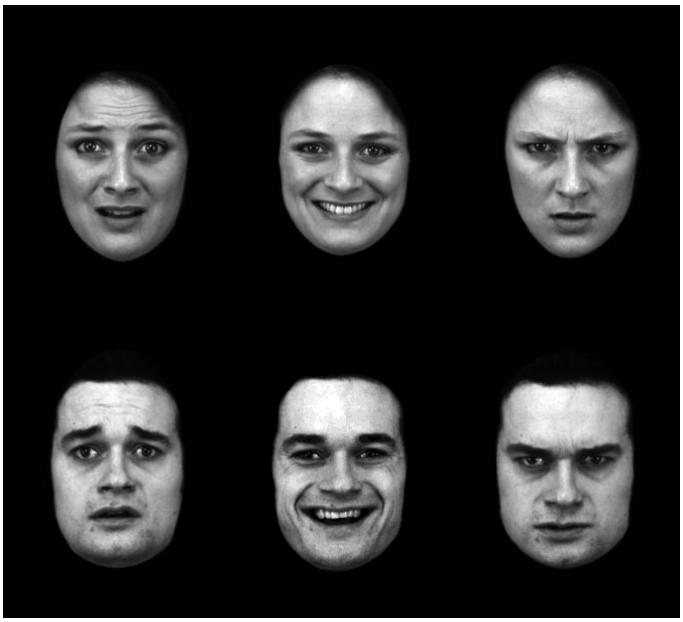

Figure 4-1. Image samples of a female and a male actor, displaying a fearful, happy, and angry expression (from left to right), from the KDEF database.

4.3.2.2 Speech

The auditory stimuli were chosen from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018), same as the stimuli used in Chapter 3. Audio-only recordings of 24 speakers (12 female) uttering one sample sentence with a neutral contents ("Kids are talking by the door"), in distinct emotional expressions of strongly fear, happiness and anger were used (a total of 72 speech stimuli). Silence at the beginning and end of each recording was removed using Praat v6.1.04 (Boersma & Weenink, 2019). Subsequently, the loudness of speech stimuli was normalized with the Loudness Toolbox (Genesis S.A.) in Matlab 2017b.

4.3.3 Experimental protocol

Participants completed four 13-minute runs of a continuous identity recognition task in the MRI scanner, comprising two face runs and two voice runs, alternating between the two modalities (i.e., Face-Voice-Face-Voice, or vice versa, with the starting modality counterbalanced across subjects). Each run contained only its corresponding type of stimuli. In each run, participants were asked to make explicit Old/New judgements on the identity of the presented stimuli (i.e.,

face images in face runs; speech clips in voice runs), regardless of their emotional expressions.. Specifically, each run included exemplars of twelve individuals (half female, target identities) and three additional individuals (1 or 2 female, counterbalanced across subjects, filler identities). Each target identity was presented six times, including two repetitions of three distinct exemplars that display angry, fearful, and happy expressions. Each filler identity was only presented three times within the last 30 trials of each run, with one single presentation of three emotional exemplars (angry, fearful, and happy). All stimuli were pseudo-randomized at the identity level, with equal frequencies of first-order transition of exemplar expressions. Post-hoc checks on the created sequences confirmed that no identity (face or voice) was repeated more than two consecutive trials, and the probability of two consecutive trials presenting the same identity back to back was 0.75% and 1.30% across all subjects. In the second face or voice runs, half of the target identities were randomly chosen from its respective first run, while the other half and the filler identities were new. Due to the complexity in design and fewer number trials when split by condition in the second runs, only data from the first run of each modality was analyzed and reported in this chapter. A simple briefing was given beforehand to participants that the face and voice identities were unrelated, since stimuli from both modalities were never present within the same run.

The experiment was run via Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007), and speech stimuli were delivered through MRI-compatible headphones (Model S14, Sensimetrics). Each face image was presented for 1500 ms. The audio clips had a mean presentation time of 1833 ms (SD = 30.9 ms). Intertrial interval was jittered in the range of 5 to 12 s (M = 6.0 s, SD = 2.1 s). A short quality check was conducted prior to the scan session, to ensure the audibility of voice stimuli in the presence of background scanner noise, and the visibility of face images on the in-scanner projector screen.

4.3.4 Behavioral analysis

As detailed in Section 4.3.3, each target identity was presented six times, comprising two repetitions of three emotional exemplars. We designated the first presentation of these emotional exemplars as E1P1, E2P1, and E3P1, ordered by their appearances, regardless of the exact expression category. Their corresponding repetitions were denoted as E1P2, E2P2, and E3P2. Consequently, these two factors, exemplar and repetition, yielded six trial types. Notedly, the

special nature of E2P1 and E3P1 trials is that, successful recognition in these cases involves a transfer of emotional expression to novel exemplars, distinct from same-image repetitions. Hence, E2P1 and E3P1 trials were the primary trial-conditions of interest. To address a range of questions, from general to more specific ones, we employed multiple trial-based generalized linear mixed models (GLMMs) to analyze recognition accuracy. To start, the overall recognition accuracy[1] in both modalities was tested in a simple one-way (modality) GLMM, and against chance level, to provide a comprehensive overview of general performance in this continuous recognition paradigm. Subsequent four GLMMs were constructed for each modality separately. In all the models, the random structure consisted of a participant intercept and a stimulus identity intercept. More specifically, four models were constructed to address the following questions: (1) to test the effects of previously described exemplar (i.e., E1, E2, & E3) and repetition (P1 & P2) factors in target identity trials on response accuracy:

$$accuracy \sim exemplar * repetition + (1|subject) + (1|id)$$

(2) to compare the detection accuracy of identities among "novel identity" trials (i.e., E1P1, and first exemplars of filler identities, referred as FL1 below). This serves as a complementary approach to inspect any bias or strategy that was developed in responses between early and late stages of the task:

$$accuracy \sim condition + (1|subject) + (1|id)$$

(3) to test if an actor's E2P1 response influenced the same actor's E3P1 accuracy, as our primary trials of interest fell in these two conditions:

$$accuracy(E3P1) \sim accuracy(E2P1) + (1|subject) + (1|id)$$

(4) to test the vulnerability of recognition in E2P1 and E3P1 trials. Specifically, we calculated, for each stimulus identity, the distance between its E2P1 (or E3P1) and the identity's preceding presentation and examined if the presentation distance affects subsequent memory:

$$accuracy \sim condition * distance + (1|subject) + (1|id)$$

Additionally, we ran drift diffusion models to incorporate both response accuracy and response time (RT) data into the analysis to estimate condition-specific drift rates (*v*), which is

[1] Responses to E1P1 trials (and the first exemplar trials of filler identities), involve the detection of new identities, rather than the recognition of old identities. However, for simplicity and conciseness in the description, both the overall performance and performance of all trial types in the Model (1) below were termed "recognition".

proposed to capture the quality or strength of a response being made (see Ratcliff et al., 2016 for a review). Here, we implemented the same method as in Chapter 3, namely the D*M approach (Verdonck & Tuerlinckx, 2016), to estimate key parameters, including drift rates for each trial condition, decision boundaries (*a)*, starting point (*z*), and within-subject variability of *v* (*sv*) and *z* (*sz*), separately for each modality and each participant. Parameters except drift rates were pre-defined constant across conditions, since there was no experimental manipulation involved that was hypothesized to influence them. Then, we constructed a linear mixed model (LMM) on the subject-condition-specific drift rates, in the same structure as the accuracy model in (1):

$$drift\ rate \sim exemplar * repetition + (1|subject) + (1|id)$$

4.3.5 FMRI acquisition and preprocessing

Functional images were acquired using a multiband sequence with a slice acceleration factor of 12 (Setsompop et al., 2012). Each run contained 1050 volumes (72 slices per volume, interleaved acquisition; FOV = 208 x 208 mm^2, matrix = 104 x 104, voxel size = 2 x 2 x 2 mm; TR = 515 ms, TE = 35 ms). The first 10 scans were discarded to avoid artifacts due to potential T1 saturation. In addition, a high-resolution 3D T_1-weighted whole brain image (1mm isotropic) was collected using a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (TR = 2.3 s, TE = 3 ms, 192 slices) for anatomical co-registration. Image preprocessing was performed in SPM12 (Wellcome Department of Imaging Neuroscience, London, UK; http://www.fil.ion.ucl.ac.uk/spm), where functional images were spatially realigned to the first volume, co-registered to the T1 image per participant, and then normalized to the MNI152 template. Finally, images were smoothed using a 6mm FWHM isotropic Gaussian kernel.

4.3.6 Univariate analysis and ROI definition

Univariate analysis was performed on each subject, using a General Linear Model (GLM). In the model, categories of interest were entered as boxcars with the length equal to the stimulus duration, convolved with the canonical hemodynamic response function (HRF). Here, the categories of interest included four types, namely each exemplar's presentation (novel - P1, vs. repetition - P2) by response accuracy (correct/incorrect). In addition, filler trials, as well as the six motion parameters were included as conditions of no interest. Then, an *F*-contrast, correct vs. incorrect responses in P2 trials, was examined separately for each modality. This contrast of

interest was used specifically to pinpoint brain regions that are sensitive to identities perceived as novel. We expected an already formed stable memory for encountered identities in repeated trials. Therefore, whenever participants made a "New" choice, it would most likely that they indeed perceived the stimulus as novel. A probabilistic threshold-free cluster enhancement (pTFCE) approach (Spisák et al., 2019) was used as the statistical inference for the tested contrast. Significant clusters were then defined as regions of interest (ROIs), from which parameter estimates were extracted in the following single trial analysis described below.

4.3.7 FMRI single trial analysis

A single trial analysis (e.g., Visser et al., 2016) was carried out in SPM12, using another GLM for each run per subject, for the trial-based ROI analysis. Specifically, we remodeled each run in a single GLM for the whole-brain analysis, with each trial as a separate regressor, in addition to the six motion parameters as regressors of no interest. Parameter estimates within the defined ROIs from Section 4.3.6 were extracted for each trial, and submitted for a trial-level LMM analysis, examining neural activities in E2P1 and E3P1 trials, where cross-emotion recognition was achieved.

4.4 Results

Overall performance

Overall recognition accuracy estimated in the simple GLMM yielded a significant modality effect ($\chi^2[1] = 28.71$, $p < .001$), confirming an overall better accuracy for faces than voices. Post-hoc tests on the estimated marginal means of accuracy further confirmed an above-chance overall accuracy in both runs (face: M = 0.77, SE = 0.018, $z = 12.15$, $p < .001$; voice: M = 0.69, SE = 0.021, $z = 8.14$, $p < .001$).

4.4.1 Faces

4.4.1.1 Behavioral results

Recognition Accuracy

The 3 (ordered exemplar) by 2 (repetition) GLMM on recognition accuracy revealed significant main effects of exemplar ($\chi^2[2] = 12.11$, $p = .002$), repetition ($\chi^2[1] = 125.14$, $p < .001$), and an exemplar-repetition interaction ($\chi^2[2] = 15.88$, $p < .001$). The interaction was driven by a

significant accuracy difference in P1 trials, but not P2 trials (see Figure 4-2a for a visualization): E2P1 accuracy was significantly lower than that of E1P1 and E3P1 trials (|z|'s > 5.0, p's < .001), while no difference was found among the three emotional exemplars in P2 trials (|z|'s < 0.9, p's > .8).

The second model tested the detection accuracy of novel identity trials (E1P1 vs. FL1). Results showed no accuracy difference between E1P1 and FL1 trials ($\chi^2[1] = 0.74$, p = .39). Importantly, accuracy in both trial conditions was above chance as shown in post-hoc tests (E1P1: b = 0.78, SE = 0.03, z = 6.32, p < .001; FL1: b = 0.73, SE = 0.06, z = 3.54, p < .001).

The next model testing the relationship of the identity-corresponding E2P1 and E3P1 accuracy, revealed a strong E2P1-accuracy effect ($\chi^2[1] = 8.93$, p = .003). This indicated that if a facial identity was correctly recognized at E2P1, it would be more accurately recognized at E3P1.

Lastly, we retested the recognition performance of E2P1 and E3P1 trials, with a distance covariate between the current trial and the same-actor's preceding trial. The model revealed a significant trial-condition effect ($\chi^2[1] = 18.49$, p < .001), confirming that E3P1 accuracy remained higher than E2P1, even when taking the distance factor into account. Moreover, there was a distance-by-condition interaction ($\chi^2[1] = 4.19$, p = .041), due to a marginal negative linear trend of distance on E2P1 accuracy (b = -0.017, SE = 0.009, z = -1.90, p = .058), but not E3P1 accuracy (b = 0.009, SE = 0.009, z = 1.00, p = .32). In other words, accuracy in E2P1 trials demonstrated a recency effect, that the more distant an E2P1 trial was away from the previous presentation of the same actor, the more likely it would be responded incorrectly (i.e., as "new").

Drift Rates

As described in the Methods, six drift rates (3 exemplars by 2 repetitions) were estimated per subject. A positive drift rate represents an "old" response, and the larger the rate is, the shorter time it takes to reach the final response (and vice versa). Here, we constructed the same exemplar by repetition structure for the LMM on DDM-derived drift rates. The model yielded strong main effects of both exemplar (F[2,145] = 17.54, p < .001) and repetition (F[1,145] = 196.79, p < .001), as well as the interaction (F[2,145] = 9.49, p < .001). Pairwise post-hoc tests on the interaction confirmed different patterns for P1 and P2 trials: in P1 trials, E2P1 drift rates were significantly larger than E1P1 (b = 0.1.10, SE = 0.24, t[145] = 4.51, p < .001), but smaller than

E3P1 (b = -0.65, SE = 0.24, t[145] = -2.65, p =.027). No difference was found among P2 trials (t's < 1.2, p's > .7; see Figure 4-2b for visualization). Furthermore, only drift rates in E2P1 trials were not different from 0 (b = -0.07, SE = 0.19, t[155] = -0.35, p = .73), suggesting that participants had ambiguous responses for E2P1 trials. The drift rate results displayed a highly consistent pattern as the accuracy results, while taking into account both response accuracy and speed. It indicated that participants experienced more difficulty (i.e., taking a longer time to respond, with less confidence, resulting in chance-level accuracy) to make decisions in E2P1, and were more confident and faster in "new" response to E1P1 and "old" response to E3P1.

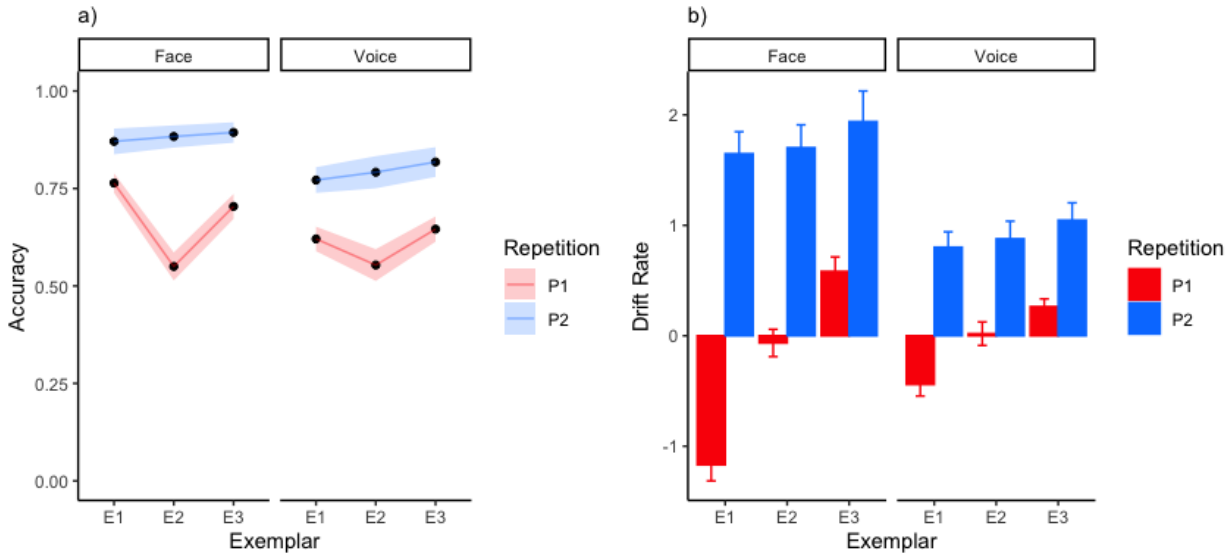

Figure 4-2. Averaged recognition accuracy (*a*) and DDM-derived drift rates (*b*) by exemplar by order, and repetition conditions, of faces and voices.

4.4.1.2 fMRI results

Univariate analysis

The P2 trial correct vs. incorrect *F*-contrast, was used to pinpoint brain regions that exhibit sensitivity to (perceived-as-) novel identities, as participants were expected to get familiarized with the encoded identities by the repetition trials. Indeed, this was supported by the behavioral results above, showing P2 trials maintained a stable high level of accuracy in repeated exemplars. Significant clusters of activity revealed under this contrast are listed in Table 4-1, with the corresponding peak coordinates, *z*-scores, and cluster extents. These regions included bilateral supplementary motor area (SMA)/superior frontal gyrus (SFG) that extended to

mid/anterior cingulate cortex (M/ACC), bilateral anterior insula (AI), posterior cingulate cortex (PCC) extending to precuneus, and left angular gyrus (AnG) (Figure 4-3a).

Table 4-1

Group-level significant activation under the *F*-contrast of interest, from the univariate analysis of faces.

***F*-contrast Correct vs. Incorrect in P2**					
Anatomical Location	***MNI coordinates***			***z-score***	K_E
	x	***y***	***z***	***(peak voxel)***	
R/L Supplementary Motor Area,	6	24	44	5.59	1693
R/L Superior Frontal Gyrus,	-4	20	44	5.50	
Mid/Anterior Cingulate Cortex	6	10	54	4.79	
R Anterior Insula	44	20	-8	4.97	560
	52	20	0	4.16	
	46	24	6	4.03	
R/L Posterior Cingulate Cortex,	-2	-26	42	4.52	788
R/L Precuneus	-2	-56	36	4.07	
	4	-48	30	4.05	
L Anterior Insula	-30	22	-10	4.50	424
	-36	16	-12	4.22	
	-46	16	0	3.60	
L Angular Gyrus	-38	-72	38	4.00	365
	-42	-60	36	3.68	
	-44	-58	22	3.50	

Post-hoc single-trial ROI analysis

As described in the Methods, a single trial analysis was carried out on the fMRI data for each subject, in order to obtain trial-level parameter estimates of the whole brain. A post-hoc linear mixed model was built on the single trial level parameter estimates of the conditions of interest, E2P1 and E3P1, which were the only novel exemplar trials after the initial presentation and required cross-expression recognition in the task. Briefly, trial condition, response accuracy, and ROI were included as fixed factors, and random effects remained the same structure as in behavioral analyses, including a participant intercept and a stimulus identity intercept.

Results from the LMM showed a significant accuracy effect (F[1,3452.8] = 7.82, p = .005) and ROI effect (F[4, 103.1] = 52.18, p < .001), with condition-by-accuracy (F[1,3304.0] = 7.13, p = .008), accuracy-by-ROI (F[4,2817.1] = 4.56, p = .001), and condition-by-accuracy-by-ROI (F[4,2994.2] = 2.53, p = .040) interactions. We directly dissected the triple interaction effect in each ROI. Post-hoc tests revealed different activation patterns between three ROIs (bilateral AI and SMC/ACC) and the other two (left AnG and PCC/precuneus) (see Figure 4-3b). In the former three ROIs, a significant activation difference between correct and incorrect trials was shown in E3P1 (z's > 3, p's < .002), but not E2P1 trials (z's < 0.7, p's >.5). The latter two ROIs did not show such a significant activity difference between accuracy in either trial type (z's < 1.4, p's > .15).

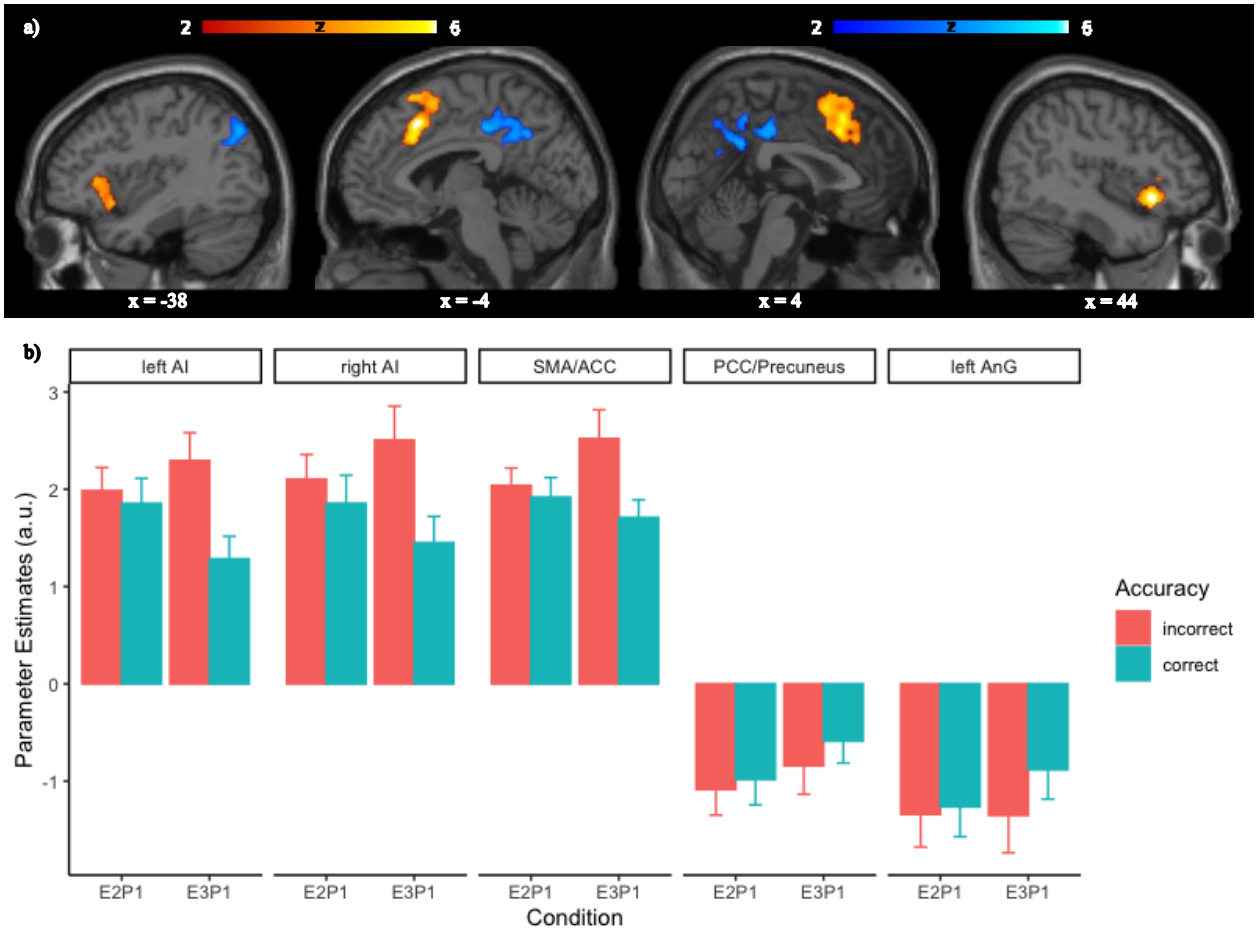

Figure 4-3. a) 2D renderings of the clusters of significant activity difference (red-scale corresponded to the incorrect > correct direction; blue-scale corresponded to the correct > incorrect direction) in response to the contrast Correct vs. Incorrect P2, under the pTFCE threshold. b) Parameter estimates in E2P1 and E3P1 trials from the ROIs,

activation in three of which (on the left) resulted in a condition-by-accuracy interaction.

4.4.2 Voices

4.4.2.1 Behavioral results

Recognition Accuracy

The 3 (ordered exemplar) by 2 (repetition) GLMM on voice recognition accuracy yielded both significant main effects of exemplar ($\chi^2[2] = 6.56$, $p = .038$) and repetition ($\chi^2[1] = 90.31$, $p < .001$), without a significant exemplar-by-repetition interaction ($\chi^2[1] = 1.70$, $p = .43$) (see Figure 4-2a for a visualization). Post-hoc pairwise tests on the exemplar factor revealed that only E2 accuracy was significantly lower than E3 ($b = -0.32$, SE = 0.12, $z = -2.56$, $p = .031$), while no significant difference was found between E1 and E2, or E1 and E3 ($|z|$'s < 1.3, $p > .5$). Similar to the results in face trials, the repetition effect confirmed a better recognition for all repeated over novel exemplar trials.

The second model on performance of novel identity trials (E1P1 vs. FL1), revealed a trending condition effect ($\chi^2[1] = 3.57$, $p = .058$), suggesting a marginally lower accuracy in FL1 than E1P1 trials. Planned *t*-tests against chance level confirmed an above chance detection accuracy in E1P1 trials ($b = 0.63$, SE = 0.04, $z = 3.18$, $p = .003$), but not in FL1 trials ($b = 0.52$, SE = 0.06, $z = 0.23$, $p > .99$).

To test the relationship between the identity-corresponding E2P1 and E3P1 accuracy, a one factor (E2P1 accuracy) GLMM was estimated on E3P1 accuracy. The results showed that, the same speaker's E2P1 accuracy did not affect its E3P1 performance significantly ($\chi^2[1] = 1.77$, $p = .18$).

Finally, recognition performance in E2P1 and E3P1 trials were tested with the distance to the same-speaker's preceding trial as covariate. The model did not yield any significant effects (distance: $\chi^2[1] = 2.19$, $p = .14$; condition: $\chi^2[1] = 0.91$, $p = .34$), nor the condition-by-distance interaction ($\chi^2[1] = 0.02$, $p = .90$). Unlike results in face, results here suggested that recognition was not affected by how distant speakers were away from their previous presentations, regardless of trial condition.

Drift Rates

Drift rates for all the six trial types were estimated for each subject in the same manner as in face trials. An exemplar-by-repetition LMM on these drift rates revealed both main effects (exemplar: $F[2,145] = 10.65$, $p < .001$; repetition: $F[1,145] = 130.35$, $p < .001$), and a trending interaction effect ($F[2,145] = 2.85$, $p = .061$). Post-hoc tests were conducted in P1 and P2 trials separately to dissect the trending interaction. In novel exemplar trials, E1P1 drift rates were significantly smaller than those in E2P1 and E3P1 trials ($|t[145]|$'s > 3, $p < .006$), but no significant difference was found between E2P1 and E3P1 conditions ($b = 0.25$, SE = 0.14, $t[145] = 1.67$, $p = .29$). No difference was found among repeated exemplar trials ($|t[145]|$'s < 1.7, p's > .25; see Figure 4-2b for visualization). Moreover, in planned post-hoc t-tests, drift rates in E2P1 and E3P1 were not significantly different from 0 ($t[110]$'s < 2.1, p's > .12), suggesting that participants cannot make a definitive response for either E2P1 or E3P1 trials, as information extracted from the stimuli accumulated over time. In contrast, drift rates were shown significantly negative in E1P1 ($b = -0.44$, SE = 0.13, $t[110] = -3.47$, $p = .002$), and positive in all P2 conditions ($t[110]$'s > 6, $p < .001$), similar to the face results.

4.4.2.2 fMRI results

Univariate analysis

Only one cluster was identified under the F-contrast of correct vs. incorrect responses in P2 voice trials. It was located in the supplementary motor area (SMA) (peak location: [4,12,62], $z_E = 3.82$, $K_E = 345$) (see Figure 4-4a). Hence, only this SMA mask was used for the subsequent ROI analysis.

Post-hoc single-trial ROI analysis

Same as the analysis approach on face-trial data, trial-level parameter estimates from the SMA ROI was extracted from the remodeled single-trial analysis, and submitted to a linear mixed model comprising two fixed factors, trial condition (E2P1 vs. E3P1) and accuracy. Random effects included intercepts for both participant and stimulus identity, same as in prior analysis. Neither main effects, nor the interaction were shown significant (p's > .19, see Figure 4-4b for a visualization).

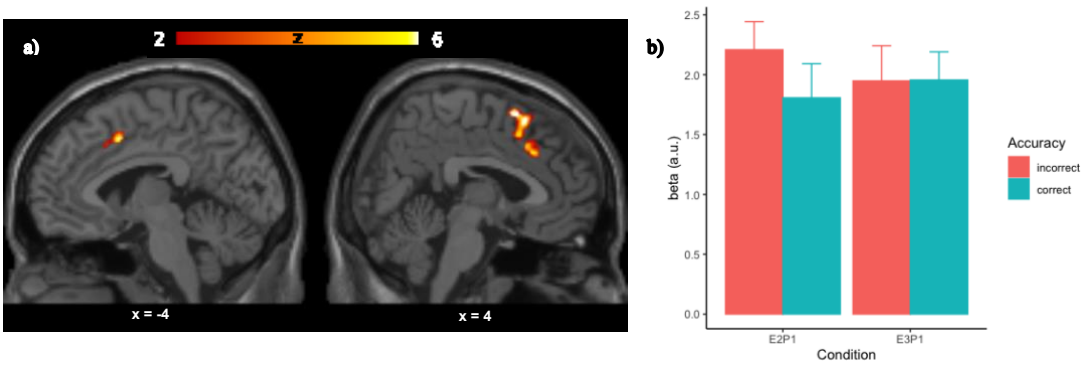

Figure 4-4. a) 2D renderings of the SMA cluster, showing significant activity difference in response to the contrast Correct vs. Incorrect P2, under the pTFCE threshold. b) Parameter estimates in E2P1 and E3P1 trials from the SMA cluster.

4.4.3 Individual differences – an Exploratory Analysis

Results of the voice trials showed a lack of difference in both recognition performance and ROI activity between E2P1 and E3P1 conditions. We intended to further explore possible factors that may have contributed to such null findings. As prior studies suggested that language familiarity/expertise (e.g., native vs. non-native language participants) could greatly influence speech and voice recognition and discrimination (see Perrachione, 2018 for a review), we speculated that individual difference of language expertise in English in our participant sample, might pose influences on the null findings of the voice run. Hence, we separated the participants into two groups based on whether their native language is English (Section 4.3.1), aiming to examine if language familiarity cast any influence on cross-expression voice recognition, as a preliminary exploratory analysis. The separation resulted in a group of 14 non-native English speakers, and the other group of 16 English native speakers. We constructed a GLMM on behavioral recognition accuracy of E2P1 and E3P1 trials, with one within-subjects (trial condition) and one between-subjects (participant group) factors.

The model yielded both significant main effects of trial condition ($\chi^2[1] = 5.40$, $p = .020$) and participant group ($\chi^2[1] = 90.31$, $p < .001$), and a significant two-way interaction ($\chi^2[1] = 10.53$, $p = .001$). Post-hoc tests revealed that this was due to a significantly higher E2P1 accuracy for non-native than native English-speaking participants ($b = 1.12$, SE = 0.27, $z = 4.23$,

$p < .001$), with no difference in E3P1 accuracy between the two groups ($b = 0.09$, SE = 0.27, $z = 0.35$, $p = .72$).

4.5 Discussion

This study aimed to examine explicit identity recognition and its corresponding neural correlates of novel emotional exemplars during the early learning stage, in a continuous identity recognition task, where both novel and repeated emotional exemplars were presented. Behavioral results suggested a strong and consistent superior recognition (i.e., recognized as "old" identity) on repeated exemplars than novel stimuli across modalities. This same-stimulus recognition advantage has been shown in a large body of identity matching and recognition studies (e.g., see Bruce, 1982; Longmore, Liu, & Young, 2008; Liu, Chen & Ward, 2014 for faces; see Saslove & Yarmey, 1980; Stevenage & Neil, 2014; Zäske et al., 2014 for voices). Worse recognition on novel exemplars indicated that changes in facial or vocal emotional expressions, similar to prior findings that, various changes in stimuli, such as viewpoint for face or speech for voice, could impair perception and recognition for unfamiliar or newly familiarized identities (e.g., Bruce, 1982; Saslove & Yarmey, 1980). This is consistent with the notion of unfamiliar identity perception or generalization relies more on image (for face) and excerpt-based (for voice) features (Longmore, Liu & Young, 2008; Stevenage, 2018). Differences in behavioral and neuroimaging results between modalities are further discussed in detail below.

4.5.1 Face representation built upon multiple exemplar exposures

For faces, we observed a recognition decrease in the second novel exemplar (i.e., E2P1, the first novel emotional exemplar after a new identity was introduced), which fit our hypothesis and replicated previous findings suggesting the vulnerability of unfamiliar face recognition (e.g., Bruce, 1982; Liu, Chen & Ward, 2014). Interestingly, recognition was significantly improved for the third emotional exemplar (i.e., E3P1), and that an above-chance identity recognition was achieved in this condition. Such results remained the same in diffusion model derived drift rates, when considering both response accuracy and speed. Drift rates, which is usually considered as a representation of the strength of memory trace, and the quality of the match between the probe and stored memory information (Ratcliff, Thapar, & McKoon, 2004; Ratcliff & Starns, 2009), revealed the same impairment of memory matching quality in E2P1, compared to E1P1 and

E3P1. We propose that while generalizing a face is difficult from one expression to another, it becomes substantially easier if at least two exemplars had been studied, which is possibly due to a relatively stable face representation formed after encoding E1P1 and E2P1.

The proposed notion is further supported by two compelling pieces of evidence. Firstly, a pronounced recency effect (e.g., Braddeley & Hitch, 1993) was observed in the explicit recognition of E2P1 trials. Specifically, recognition deteriorated as the distance between E2P1 trials and the corresponding actor's preceding trial increased. This indicated that a successful face generalization of E2P1 relies on being closer to its predecessor. However, E3P1 recognition did not show such a reliance. Secondly, an actor-specific relationship was found between E2P1 and E3P1 recognition. This strengthened the notion that certain processes of identity learning and integration took place during E2P1 and benefited subsequent E3P1 recognition. Importantly, this phenomenon cannot simply be attributed to the inherent good recognizability of certain actors (e.g., face distinctiveness, Johnston & Ellis, 1995), given that actor-specific accuracy was already accounted for in our model's random effect. Nor can it be explained by an increased tendency to respond "old" as the experiment progressed into the later stage, where more repeated exemplar trials took place by nature. We offer two accounts to argue against this possibility. If an increase in response bias towards "old" was the sole contributor that drove the E3P1 recognition enhancement, we would not observe the modulation of the same-actor's E2P1 accuracy on E3P1 response. In addition, the complementary analysis on novel identity trials confirmed that E1P1 and FL1 accuracy was on a comparable level, and not affected by the presumed bias. Taken the comprehensive behavioral results of face trials together, it becomes clear that encoding two same-actor exemplars resulted in the formation of a more stable face representation that, in turn, facilitated recognizing a third novel exemplar.

4.5.2 Neural correlates of familiarized facial identities

The contrast we employed to detect ROIs was designed from the purpose of revealing brain regions that were sensitive to perceived-as-novel identities. As repeated exemplar (P2) trials indeed showed much higher recognition accuracy than novel exemplar (P1) trials, it supported the expectation that the encoded identities during repeated trials were well remembered, regardless of whether it's recognized through the same-stimulus strategy or the true identity. And when an incorrect response was made, they perceived the repeated exemplar as a novel identity.

Hence, we reasoned that this contrast in the repeated-exemplar trials should reveal regions that are different in processing identity familiarity or novelty, in a similar way as in other fMRI studies contrasting personally familiar, or perceptually familiarized faces with novel/unfamiliar ones (see Natu & O'Toole, 2011; Ramon & Gobbini, 2018 for reviews). Five clusters were identified in the face trials, including bilateral AI, SMA/ACC, PCC/precuneus, and left AnG.

Studies on insula functionality consistently indicate that the dorsal AI is involved in detection of novel stimuli across sensory modalities (e.g., Sridharan, Levitin & Menon, 2008; Cai et al., 2016), and demonstrate strong causal influences on other networks, such as the Default Mode Network (DMN) and Central Executive Network (CEN) in tasks that require more cognitive control (Baldo et al., 2011). Moreover, the bilateral AI and the SMA/ACC clusters unveiled in our planned contrast, are often collectively considered as part of the Salience Network (SN), in addition to amygdala and other subcortical regions such as hypothalamus and ventral striatum (e.g., Chiong et al., 2013; Seeley et al., 2007). The SN is sensitive to subjective salience elicited by a task or stimulus (Seeley et al., 2007; Menon, 2015; Menon & Uddin, 2010; Peters, Dunlop, & Downer, 2016). Lamichhane and Dhamala (2015) further reported a positive correlation between the activity in SN nodes and task difficulty. The other two clusters (PCC/precuneus and left AnG), are nodes commonly found in the DMN. The DMN has shown active at rest, and been often involved in self-related social cognitive processes, mentalizing and theory of mind (Mars et al., 2012). Here, we found correctly recognized repeated-exemplar trials elicited less activity in SN clusters and less deactivation in DMN clusters, indicating less salient these repeated faces became when they were perceived "old", and more salient when they were perceived "new". DMN ROIs have also been found in prior neuroimaging studies when contrasting familiarized against strange or unfamiliar faces, such as precuneus (Gobbini et al., 2004; Gobbini et al., 2006) and PCC (Sugiura et al., 2001; Pierce et al., 2004; Gobbini et al., 2004). Taken together, our results showed that clusters in both SN and DMN were involved in perceiving newly familiarized facial identities through repeated exemplar presentations, in line with previous work examining the functionality of these networks.

Results from the post-hoc single trial analysis revealed no activity difference in E2P1 trials, but a significantly smaller activation for perceived-as-old than perceived-as-new E3P1 trials, in three SN clusters. Based on the functionality research of the SN mentioned above, such an activation pattern may indicate that E2P1 trials were equally salient (or difficult), regardless of

being perceived as old or new. On the other hand, E3P1 exemplars had a lower salience level when participants perceived them as old. This assumption aligns with the behavioral findings demonstrating improved recognition in E3P1 trials (i.e., the correct recognition became easier), and a difficult chance-level recognition in E2P1 trials. This may resemble, in a very simplified way, the process of a face identity being learned and becoming (perceptually) familiar. However, the DMN clusters did not show significant difference between response or trial types, even though the (de)activation seemed to follow a similar pattern as what the SN showed (Figure 4-3b). As a task-negative network, the DMN is constantly found active at rest (Mars et al., 2012), and its activity decreases when attention is directed towards external stimuli and/or the task becomes more attention demanding (Raichle et al., 2001; Buckner et al., 2008). The lack of the DMN difference may be due to that the overall task engagement is not as sensitive as deciding face novelty or salience. The DMN regions may not show significant difference between E2P1 and E3P1 trials, as the task requires constant attention and engagement for task responses. Altogether, the findings from the behavioral and the fMRI-ROI analyses showcased an impaired cross-expression recognition in E2P1, but an improved recognition in E3P1 trials. We posit that this E3P1 enhancement stems from the formation of face representations after being exposed to at least two distinct emotional exemplars. The activity in the clusters within the SN appeared reflective of this improvement.

One point worth mentioning is that, the ROI definition contrast did not reveal any face-specific processing regions that were reported in face processing studies, such as FFA, OFA, or temporal regions including the STS (see Natu & O'Toole, 2011 for a review). One possibility is that, with limited amount of stimuli exposures in our paradigm, the familiarity of individuals were not formed as strong as in some prior studies, where famous or personally familiar faces were used for contrast (Gobbini & Haxby, 2006). Another reason may be due to the experimental paradigm difference. In the current study, participants were completing the recognition task while learning novel exemplar variance at the same time, while other studies typically detected neural differences in a traditional old/new (or familiar/unfamiliar) recognition task after one or even multiple encoding/familiarization sessions. Moreover, studies that have investigated neural activity patterns between familiarized and unfamiliar faces, also showed inconsistent, yet sometimes conflicted results of corresponding brain regions. For example, some studies have found increased activity in familiar than unfamiliar faces in the FuG (e.g., Rossion, Schiltz, &

Crommelinck, 2003; Pierce et al., 2004), while others found the opposite activation pattern (e.g., Gobbini et al., 2004; Gobbini & Haxby, 2006). In addition, there were also studies unable to find differential activation patterns in these regions (e.g., Leveroni et al., 2000; Sugiura et al., 2001).

4.5.3 Voice as a weaker cue for identity information

Although we found cohesive behavioral and neuroimaging evidence of an improved cross-expression face recognition particularly in E3P1 trials, results remain underwhelming for voices. This contrasts with our hypothesis, which expected a similar behavioral and neural pattern as observed in the face task. Behaviorally, the emotion-by-repetition interaction on recognition accuracy was not qualified. Considering the drift rate results alongside, there was no significant improvement in recognition performance from E2P1 to E3P1 trials for voices. Furthermore, speaker-level accuracy in E2P1 could not predict E3P1 accuracy, implying a possible disconnection between speaker-wise E2P1 and E3P1 exemplars. In addition, the distance effect was absent in either E2P1 or E3P1 conditions. Two possibilities could be considered here. First, the cross-emotion recognition might be stable and independent of how distant its prior trial was, in the same case as the face-E3P1 condition. Second, the lack of the distance effect may be because participants had difficulty relating the current voice to its last appearance. Given that a lack of recognition improvement from E2P1 to E3P1, and a marginal decrease in novel identity detection for FL1 compared to E1P1, these behavioral results collectively suggested an overall confusing task performance in voice recognition. Hence, the second possibility seems more plausible to explain the absence of the distance effect, and indicates a difficulty in associating different emotional exemplars to the same speaker. Lastly, granted that cross-expression voice recognition was rather poor, a superior recognition of repeated voice stimuli was still shown clearly and consistently. Overall, the findings of voice recognition were in line with numerous studies in the past, suggesting that voice recognition is generally worse, less accurate and reliable than face recognition (e.g., Stevenage & Neil, 2014; Barsics, 2014).

At the neural level, only one cluster in SMA revealed a larger activity for P2 incorrect trials. The ROI was located around the same area, albeit with a smaller cluster size as the face ROI counterpart, suggesting some cross-modality similarity of the SN function under this novelty/saliency-sensitive contrast (e.g., Seeley et al., 2007; Menon, 2015; Lamichhane & Dhamala, 2015). However, no other region from either SN or DMN was revealed. This might be

linked back to the difficulty of voice recognition. Recognition of repeated exemplars turned out to be already difficult, not as easy as faces (e.g., Barsics, 2014), let alone cross-expression recognition. In light of the possible account of a difficulty in registering novel exemplars to the same speaker, it is thus not surprising to discover a lack of neural difference between responses in E2P1 and E3P1 trials in the single ROI within the SN.

In addition to the nature of voice being a weaker identity cue, we explored and proposed a possible factor, namely participants' language familiarity/expertise, that may influence voice learning and recognition, given that the speech stimuli used in the experiment were all North American neutral accented English speech samples. Indeed, the language-familiarity effect has been reported in earlier studies with a variety of tasks (e.g., Hollien, Majewski, & Doherty, 1982; Perrachione & Wong, 2007), showing listeners are more accurate at identifying or discriminating voices in their native language than a second or foreign language. One exploratory analysis on the behavioral results of E2P1 and E3P1 was carried out, aiming to offer some preliminary insights into the potential impact of the language factor. From the results, E2P1 accuracy seemed heavily influenced by participants' native language (as English or not). However, in contrast to the advantage proposed by the language familiarity effect (see Perrachione, 2018 for a review), the native language effect from our results reflected a disadvantage of cross-expression recognition in E2P1 trials. Two potential reasons could reconcile the contradictory effects. Firstly, native English speakers might be less engaged or concentrated in the task overall due to the simplicity of the speech stimuli and the task paradigm (i.e., only one sample sentence, uttered by different speakers), which may result in worse learning of the stimuli. The second possibility is that native speakers indeed had higher sensitivity to these speech stimuli than non-native speakers. As a result, they may have miscategorized certain within-speaker variance into between-speaker variance, establishing a new identity for E2P1, leading to more "new" responses (see a similar argument in Lavan, Burston, & Garrido, 2018). Nonetheless, the exploratory analysis and its results implied a possible influence from participants' language familiarity on cross-expression voice perception, specifically, in E2P1 trials, that warrants further investigation.

4.6 Limitations and Future Directions

The divergent results observed in face and voice recognition in the study, as we posit, could be influenced by the inherent difficulty of voice recognition. While individuals can associate multiple exemplars with one facial identity, leading to a facilitated cross-expression recognition, voice recognition may face challenges in associating different emotional exemplars with the same speaker. Further modifications shall be considered in the aim of strengthening voice encoding and identity co-registration, in order to improve performance in voice and potentially uncover neural discrepancies between different responses of voice recognition. One possible approach is by adding semantic information (e.g., name or occupation) along with voice exemplars, which have been used in other voice training tasks to strengthen voice encoding (e.g., von Kriegstein & Giraud, 2006). However, a relevant concern would be the difficulty to incorporate it into our continuous recognition task, since there was practically no training session beforehand. In the case of providing additional information throughout the course, participants would easily shift their response strategy, relying on such semantic information largely or even entirely. Pilot tests are needed to check if such a manipulation is applicable, for example, adding associated semantic information in certain amount of early trials, to only help facilitate voice identity formation or co-registration explicitly at the early stage of the task. Another approach is to increase the amount of distinct exemplars, as more exposures have been shown to improve familiarization of an individual (Bonner, Burton, & Bruce, 2003; Zäske et al., 2014). Following this approach, it is possible to have novel exemplar conditions such as E4P1 and E5P1. Given that voice is a less efficient cue for person identity than face, the same exemplar advantage observed in face may take more and/longer exemplar learning to occur in voice.

In addition, as discussed at the end of the last section, participants' familiarity/expertise on English should be controlled, either as an experimental factor, or in a more carefully designed manner during recruitment. Admittedly, we conducted this exploratory analysis, *post hoc*, with limited amount of demographical information collected from participants. It allowed us to examine the language effect, fortunately with a relatively equal, yet still small amount of participants per group. However, the group separation was simplified into a dichotomy, which raised several issues, such as the heterogeneity of mother tongues in non-English-native speaking participants, differences in English familiarity or expertise among them, especially given the substantial bilingual population within the Greater Montreal Area. Thus, a more rigorous

grouping method using more detailed language usage information, such as English proficiency (e.g., in numeric scales or multiple categories), may likely work better to capture the actual language familiarity, for testing group, or even individual differences on such cross-expression speaker recognition. Furthermore, the collected information in the study did not allow us to probe other candidate participant-specific factors that may affect voice recognition, although only a handful of studies focused on this avenue of research. To the best of our knowledge, gender differences (both of listeners and speakers) (e.g., Skuk & Schweinberger, 2013) and autistic traits (Skuk et al., 2019; Schelinski, Riedel, & von Kriegstein, 2014; Lin et al., 2015) have been shown to influence familiar and/or unfamiliar voice recognition. Hence, future studies with a larger sample size and such relevant psychological measures collected, will help provide a more comprehensive picture of how voice recognition varies at an individual level.

4.7 Conclusion

This study examined how people generalize identity information onto novel and repeated emotional exemplars in continuous face and voice recognition tests, without prior familiarization. In the case of face recognition, behavioral results indicated an impeded recognition when encountering the second novel exemplar of an previously learned identity, but an improved cross-expression recognition in the third novel exemplar. We proposed that a relatively stable face representation can be formed as early as after two distinct exemplars are learned. Bilateral anterior insula and supplementary motor area, part of the Salience Network, showed less activity when correctly recognizing the third novel exemplar, indicating an easier and improved recognition after encoding two distinct exemplars. However, such behavioral and neuroimaging findings were not observed in cross-expression voice recognition. This posed a possible need of a more effective voice learning approach, and a more carefully controlled participant sample, possibly based on language familiarity.

4.8 References

Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition, 21*(2), 146–155.

Baldo, J. V., Wilkins, D. P., Ogar, J., et al. (2011). Role of the precentral gyrus of the insula in complex articulation. *Cortex, 47*, 800–807.

Barsics, C. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244-254.

Belin, P., & Zatorre, R. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport, 14*(16), 2105-2109.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.04, retrieved 28 September 2019 from http://www.praat.org/

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition, 10*(5), 527–536.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Bruce, V., Henderson, Z., Greenwood, K., et al. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339–360.

Buchsbaum, B. R., Lemire-Rodger, S., Bondad, A., & Chepesiuk, A. (2015). Recency, repetition, and the multidimensional basis of recognition memory. *J Neurosci, 35*(8), 3544-54.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485.

Burton, A. M., Bruce, V., & Hancock, P.J.B. (1999). From pixels to people: A model of familiar

face recognition, *Cognitive Science, 23*(1), 1-31.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243–248.

Cai, W., Chen, T., Ryali, S., et al. (2016). Causal interactions within a frontal-cingulate-parietal network during cognitive control: convergent evidence from a multisite-multitask investigation. *Cerebral Cortex, 26*(5), 2140-2153.

Chiong, W., Wilson, S. M., D' Esposito, M., et al. (2013). The salience network causally influences default mode network activity during moral reasoning. *Brain, 136*(6), 1929–1941.

Eger, E., Schweinberger, S. R., Dolan, R. J., & Henson, R. N. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *NeuroImage, 26*(4), 1128-1139.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8*(4), 431–439.

Ferris, S. H., Crook, T., Clark, E., et al. (1980). Facial recognition memory deficits in normal aging and senile dementia. *Journal of Gerontology, 35*(5), 707-714.

Fysh, M.C. (2018). Individual differences in the detection, matching and memory of faces. *Cogn. Research, 3*(20).

Gobbini, M. I., & Haxby, J. V. (2006). Neural response to the visual familiarity of faces. *Brain Research Bulletin, 71*, 76–82.

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage, 22*, 1628–1635.

Goeleven, E., de Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion, 22*(6), 1094-1118.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in cognitive sciences*, *4*(9), 330–337.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223–233.

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62*, 201-222

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under

normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*(2), 139-148.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577–596.

Johnston, R. A., & Ellis, H. D. (1995). Age effects in the processing of typical and distinctive faces. *Quarterly Journal of Experimental Psychology, 48A*, 447 465

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1-16.

Lamichhane, B., & Dhamala, M. (2015). The Salience Network and Its Functional Architecture in a Perceptual Decision: An Effective Connectivity Study. *Brain connectivity*, *5*(6), 362–370.

Latif, M., & Moulson, M. C. (2022). The importance of internal and external features in recognizing faces that vary in familiarity and race. *Perception*, *51*(11), 820–840.

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex, 21*(12), 2820-2828.

Lavan, N., Burston, L. F. K., Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593.

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications, 10*, 2404.

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition, 193*, 104026.

Leveroni, C. L., Seidenberg, M., Mayer, A. R., et al. (2000). Neural systems underlying the recognition of familiar and newly learned faces. *Journal of Neuroscience, 20*, 878–886

Lin, I. F., Yamada, T., Komine, Y., et al. (2015). Vocal identity recognition in autism spectrum disorder. *PLoS ONE, 10*, e0129451.

Liu, C. H., Chen, W. F., & Ward, J. (2014). Remembering faces with emotional expressions. *Frontiers in Psychology*, *5*, 1439.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in

North American English. *PLoS ONE, 13*(5), e0196391.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77–100.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Mars, R. B., Neubert, F. X., Noonan, M. P., et al. (2012). On the relationship between the "default mode network" and the "social brain". *Front Hum Neurosci, 6*, 189.

Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: effects on eyewitness accuracy and confidence. *Br J Psychol, 94*(3), 339-54.

Menon, V. (2015). Salience Network. In: A. W. Toga, editor. Brain Mapping: An Encyclopedic Reference, vol. 2, pp. 597-611. Academic Press: Elsevier.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain structure & function*, *214*(5-6), 655–667.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity.*Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577–581.

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: a review and synopsis. *British journal of psychology, 102*(4), 726–747.

O'Toole, A. J., Weimer, S., Dunlop, J., et al. (2011). Recognizing people from dynamic video: Dissecting identity with a fusion approach. *Vision Research, 51*(1), 74-83.

Pernet, C. R., McAleer, P., Latinus, M., et al. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage, 119*, 164-174.

Perrachione, T.K. (2018). Recognizing speakers across languages. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception*, Oxford: Oxford University Press.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*(8), 1899–1910.

Peters, S.K., Dunlop, K., & Downar, J. (2016). Cortico-striatal-thalamic loop circuits of th salience network: A central pathway in psychiatric disease and treatment. *Frontiers in*

Systems Neuroscience, 10, 1-23.

Pierce, K., Haist, F., Sedaghat, F., & Courchesne, E. (2004). The brain response to personally familiar faces in autism: Findings of fusiform activity and beyond. *Brain, 127*, 2703–2716.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z. Power, W.J., Gusnard, D.A., & Shulman, G.L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98,* 676-682.

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition, 26*(3), 179–195.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, *116*(1), 59–83.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences*, *20*(4), 260–281.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*(4), 408-424.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897-905.

Roark, D. A., O'Toole, A. J., Abdi, H., & Barrett, S. E. (2006). Learning the moves: The effect of familiarity and facial motion on person recognition across large changes in viewing format. *Perception, 35*(6), 761–773.

Robertson, D. J., Jenkins, R., & Burton, A. M. (2017). Face detection dissociates from face identification. *Visual Cognition*, *25*, 740–748

Rossion, B., Schiltz, C., & Crommelinck, M. (2003). The functionally defined right occipital and fusiform "face areas" discriminate from visually familiar faces. *NeuroImage, 19*, 877–883.

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: A PET study of face categorical perception. *Journal of Cognitive Neuroscience, 13*, 1019–1034.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65*(1), 111-6.

Schelinski, S., Riedel, P., & von Kriegstein, K. (2014). Visual abilities are important for auditory-only speech recognition: Evidence from autism spectrum disorder. *Neuropsychologia, 65*, 1–11.

Seeley, W. W., Menon, V., Schatzberg, A. F., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *The Journal of neuroscience*, *27*(9), 2349–2356.

Sergerie, K., Lepage, M., & Armony, J. L. (2006). A Process-Specific Functional Dissociation of the Amygdala in Emotional Memory. *Journal of Cognitive Neuroscience, 18*, 1359-1367.

Sergerie, K., Lepage, M., & Armony, J. L. (2007). Influence of emotional expression on memory recognition bias: A functional magnetic resonance imaging study. *Biological Psychiatry, 62*(10), 1126–1133.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic resonance in medicine*, *67*(5), 1210-1224

Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research, 296*, 131-140.

Skuk, V. G., Palermo, R., Broemer, L., & Schweinberger, S. R. (2019). Autistic Traits are Linked to Individual Differences in Familiar Voice Identification. *Journal of autism and developmental disorders*, *49*(7), 2747–2767.

Spisák, T., Spisák, Z., Zunhammer, M., et al. (2019). Probabilistic TFCE: a generalised combination of cluster size and voxel intensity to increase statistical power. *Neuroimage, 185,* 12-26.

Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(34), 12569–12574.

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia, 116,* 162-178.

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*(3), 266–281.

Sugiura, M., Kawashima, R., Nakamura, K., et al. (2001). Activation reduction in anterior temporal cortices during repeated recognition of faces of personal acquaintances. *NeuroImage, 13*, 877–890.

Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time

data: Theory and implications. *Psychological Review, 123*(2), 208–218.

Visser, R. M., de Haan, M. I. C., Beemsterboer, T., et al. (2016). Quantifying learning-dependent changes in the brain: Single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology, 53*(8), 1117-1127.

von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*(2), 948-955.

von Kriegstein, K., & Giraud, A. L. (2006) Implicit multisensory associations influence voice recognition. *PLoS Biol 4*(10).

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*(3), 367-376.

Xu, H., & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology*, *74*(7), 1185–1201.

Xu, X., & Biederman, I. (2010). Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint. *Journal of Vision*, *10*(14), 36-36.

Zäske, R., Volberg, G., Kovacs, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience, 34*(33), 10821-31.

5. General Discussion

The intent of the dissertation is to study how within-person variance in emotional expression of face and voice affects our ability of learning person identity and subsequent expression-independent identity recognition. Through two behavioral studies and one fMRI study, evidence collectively suggests an intricate influence of emotional expression variance in shaping identity representation and subsequent recognition. The work also provides deeper insights and contribution to the ongoing discussion of the relationship between the processing pathways of emotional expression and identity information in face and voice (e.g., Bruce & Young, 1986; Belin et al., 2011; Calder & Young, 2005).

5.1 Summary of the findings

The findings across three studies are summarized in three aspects, as guided by the three thesis objectives. First, we provide behavioral evidence across studies (Study 1 and 2), showcasing that exposure to exemplar variance of emotional expression is beneficial for subsequent identity recognition on novel exemplars. The advantage has been supported by both theoretical mechanisms of face and voice mental representations (e.g., Jenkins & Burton, 2011; Lavan, Knight & McGettigan, 2019) and also imperial evidence using face images with large unsystematically controlled exemplar variance (Murphy et al., 2015; Ritchie & Burton, 2017). Moreover, such a variance advantage was observed not only in recognition performance from a classic encoding-recognition paradigm (Studies 1 and 2), but also in a continuous recognition task, after two distinct exemplars learned (i.e., E3P1 trials in Study 3). Smaller neural activity in the Salience Network clusters was found in correct (perceived as old) than incorrect (perceived as new) E3P1 trials, reflecting an easy recognition of E3P1 trials compared to E2P1 trials, consistent with the behavioral E3P1 improvement.

Secondly, we tested whether the revealed exemplar variance advantage is driven by specific emotion category of exemplars in Study 2, by altering the emotion category used in the single exemplar encoding (*Uni*) condition as a between-subjects factor. With an additional correlation modeling of stimulus arousal ratings, and the complementary support of two supplementary experiments and Experiment 2 in Study 1, we illustrated that the exemplar variance advantage did not simply stem from the inclusion of certain type of exemplars. On the contrary, it was the

single exemplar encoding condition that exhibited an arousal-led effect on identity memory (see 5.3 for a detailed discussion).

Lastly, Studies 2 and 3 consisted of identical experimental procedures for visual and auditory modalities, providing a solid opportunity to compare behavioral performance (Studies 2 and 3) and neural activity (Study 3) between modalities in a well-controlled experimental design. The exemplar variance advantage, as well as the relationship between stimulus arousal level and subsequent recognition accuracy observed in Study 2, were present in both faces and voices. This is not surprising, due to the proposed similar hierarchical structure of face and voice processing (Bruce & Young, 1986; Belin et al., 2011), and a similar prototype account of mental representations of face (e.g., Jenkins & Burton, 2011) and voice (e.g., Lavan, Knight & McGettigan, 2019). However, divergent results were found in Study 3 between faces and voices, at both behavioral and neural levels. Despite a common superior recognition of repeated exemplars than first-time exemplars in both modalities, cross-expression recognition in second and third exemplars (i.e., E2P1 and E3P1) showed an improved accuracy in E3P1 for faces, but not for voices. Similarly, the Salience Network ROIs revealed different activity patterns between two modalities, indicating an easier recognition for E3P1 faces, while no trial-condition or response difference was found in voices.

5.2 Connections of divergent findings between Studies 2 and 3

As mentioned above, Study 3 revealed divergent findings in face and voice. Two possible accounts have been discussed in Section 4.5.3. Looking at both studies, however, the divergent results in Study 3 may cast concerns of whether and how we should re-assess the consistent behavioral results of the exemplar variance advantage across modalities and multiple participant groups in Study 2. The issue rose mostly in voice recognition findings. Here, I offer a few arguments to address such concerns. Firstly, it is clear that the experimental paradigm was different between two studies. Study 2 employed a classic recognition paradigm that comprised an encoding and a recognition stage, while Study 3 only consisted of a single recognition stage, where participants learned identities and performed the recognition task at the same time. In Study 2, the separate encoding stage could allow some retrospective registration of different exemplars onto the same speaker. For instance, it is possible where one perceived the second novel exemplar (E2P1) of a speaker as a new identity immediately, but after hearing one or more

additional exemplars, they may retrospectively realize that these exemplars were all from the same speaker. Such a scenario would still lead to a better identity recognition when the test was conducted afterwards, but would not be recorded in a continuous recognition task. In addition, the tested stimuli were noticeably different: neutral exemplars in Study 2, and emotional exemplars in Study 3.

Secondly, there was more substantial evidence in Study 2, supporting that a better identity representation was involved to achieve better recognition. As discussed in Chapter 3, two implicit encoding measures (response time slope across four exemplars, and age judgment consistency) both reflected, albeit indirectly, the priming effects of repeated identities, and furthermore predicted subsequent explicit recognition.

Additionally, it is worth noting that an exemplar repetition was added in Study 3, while the original *Multi* encoding condition in Study 2 (and 1) contained only one single presentation per exemplar. We may have underestimated the influence of simple stimulus repetition on identity generalization towards novel exemplars. Specifically, it is possible that presenting an identical repetition may provide a reference for the identity and restrict participants' ability to generalize, causing a misclassification of within-person variance as between-person variance (see Lavan, Burston & Garido, 2018 for a similar argument). Although there is no empirical evidence yet to support such an assumption, our results in Study 1 may provide some indirect but relevant insight. Particularly, in the *Fear*-group of Experiment 1, participants encoded fearful speech stimuli, and recognition was tested with both fearful and neutral stimuli. Hence, the encoded stimuli were also appeared in the recognition test. Results of response bias revealed not only an "old" bias for same-emotional stimuli, but also a significant "new" bias for different-emotional stimuli. Although we cannot draw any causal conclusions about the "new" bias and the encoded stimuli being tested, the results seem to fit our assumption regarding the changes in the threshold of classifying within- or between-person variance.

Altogether, we acknowledge the multiple differences in the experimental designs between two studies, most of which were modified intentionally to suit Study 3's research questions. Considering these differences and the supplementary support from implicit encoding measures from Study 2, we believe that the divergent results observed in Study 3, particularly the absence of voice recognition improvement and the potential difficulty of registering different exemplars onto the same speaker (see Section 4.5.3), are more likely to attribute to the experimental

modifications mentioned above, rather than challenging findings in Study 2 as contradictory evidence.

5.3 Implications of emotional arousal in identity memory research

Now we take a further look at the emotional exemplar variance effect on identity recognition memory across three studies. Following the thesis objectives, we in fact examined two levels of the potential influence: one being if any change of emotional expression affects identity memory, and the other being how learning emotional expression variance affects identity memory. Results from the Experiment 1 in Study 1, and parts of Study 3 findings (E2P1 accuracy), directly and/or indirectly answered the first aspect, confirming that a change in emotional expression impairs identity memory, especially during the early stage of learning identities. This was consistent with a range of studies on unfamiliar face and voice recognition (e.g., Bruce, 1982; Liu, Chen, & Ward, 2014; Stevenage & Neil, 2014).

Results from Study 2 revealed an exemplar variance advantage in both modalities. Taken the main and 2 follow-up studies together, we ruled out the possibility that the observed advantage is driven by sad exemplars alone. However, the lack of recognition difference between the two encoding conditions in the *Uni*-sad participant group (Study 2) was indeed due to an impaired recognition from the *Uni* condition using high arousal exemplars (i.e., *Uni*-Fear, *Uni*-Angry, and *Uni*-Happy groups), compared to the *Uni*-Sad group. Should we interpret such results simply as a "sad-face" advantage, akin to the previously reported "happy-face advantage" (e.g., D'Argembeau et al., 2003; D'Argembeau & Van der Linden, 2007)? Upon further analysis, our results actually indicated that the recognition impairment was associated with the arousal level of the stimuli, rather than a specific emotion (see discussion in Section 3.6.2). Furthermore, this arousal-based account can indeed provide new insights into some of the inconsistent or conflicting results reported in past face memory studies.

To start off, the prominent studies that first proposed the happy face advantage, used happy-, angry-, and neutral-faces as encoding material and the identity memory was tested in neutral exemplars (D'Argembeau et al., 2003; D'Argembeau & van der Linden, 2007; Savaskan et al., 2007). As there has been demonstrated a larger negative bias for high-arousal than low-arousal stimuli (Yuan et al., 2019), it is likely that happy-encoded faces exhibited a better recognition due to a relatively lower arousal level compared to angry faces (see Hagemann, Straube &

Schulz, 2016 for a similar discussion). This could result in attention resource competition, interfering with the identity processing of high-arousal expressed faces. In line with this, studies revealing the opposite – an angry face advantage (e.g., Jackson et al., 2009; Jackson, Linden, & Raymond, 2014) – typically investigated visual working memory for face identities, where rapid capture of attention and processing are crucial for optimal working memory performance. In such cases, high-arousal emotions (e.g., angry) could indeed be beneficial due to its enhanced short-term attention. A similar arousal-based account was proposed to provide a fitting explanation to results from multiple visual search experiments using emotional face stimuli (Lundqvist, Juth, & Öhman, 2014). It is worth mentioning that we do not propose this arousal-led attention competition account to be a one-size-fits-all explanation for all emotional-face-memory related results. Instead, it is a recommended perspective worth considering when encountering and trying to reconcile seemingly contradictory findings in research involving emotion, attention, and/or emotional memory.

5.4 Interactions between emotional expression and identity processing

Looking back at the three studies in the thesis, they all centered on manipulations of facial and/or vocal emotional expressions to probe resulting effects in identity recognition. The nature of interaction and/or independence between processing pathways of emotional expression and identity has been long debated (see Bruce & Young, 1986; Calder & Young, 2005), and there is various neuropsychological, cognitive, and neuroimaging evidence supporting either proposal. Hence, the results in the thesis are meaningful in contributing to understanding the processing manners of the two components.

The common ground of the key findings from three studies is that changes in emotional expression pose influences on identity recognition performance, which overall is in favor of an interactive view of the identity and expression processing pathways. Specifically, learned or encoded voices were worse recognized once changes of the emotional expression were introduced in speech (Study 1); a similar recognition decrease was also found, particularly in the early stage of face learning (i.e., E2P1 decreased performance in Study 3).

Findings in Study 2 however, seems to suggest a more complex relationship. On the one hand, the *Uni* condition of identity encoding (i.e., repeated presenting a single exemplar of the identity) resulted in discrepancies in recognition performance driven by the emotional arousal of

exemplars (see Section 3.4). Our proposal is that attentional resources are in competition (Mather & Sutherland, 2011; Lee et al., 2014) between emotional and identity analysis when processing a highly aroused exemplar. This is in line with the separate processing view (Bruce & Young, 1986). More attentional resources are allocated to the processing of emotional component, rather than identity information when processing high aroused faces (Hagemann, Straube & Schulz, 2016). On the other hand, the *Multi* condition of identity learning supports the previously reported exemplar variance advantage (e.g., Ritchie & Burton, 2017; Murphy et al., 2015; Andrews et al., 2015). If the two pathways are totally separate, identity processing should be unaffected by which emotional expression is carried in the face images, hence the same identity information would be extracted from multiple exemplars of the same individual. In that case, no advantage would be expected from encoding more variable exemplars. Conversely, a principle component analysis (PCA) framework has been proposed to offer a computational basis for the interaction view (Calder et al., 2001; Calder & Young, 2005), as an alternative to Bruce and Young's face processing model (1986) which started the independence vs. interaction debate. In principle, it is a stimulus-driven analysis that extracts and derives face features into principle components (PCs). Some components code identity and emotional expression separately, while others code both pieces of information. According to this framework, the PCs responsible for coding both information are crucial in developing a stable identity representation, as they carry different information from various exemplar faces which can lead to a stable representation of the identity, achieved by either storing all the information or averaging them.

Taken together, our results lend support to both independent and interactive manners of emotional expression and identity processing. Indeed, a similar implication of the co-existence of both relationships has been proposed by Fitousi & Wenger (2012), that independence of two processing pathways is expected at a single face level, but interaction occurs at the ensemble level. One point worth noting is that the current thesis focused specifically on the influence of emotional expression on identity memory. We acknowledge that this directional research topic (the influence of emotion on identity memory) can only reflect part of the dynamic relationship between the processing of both pathways. It is beyond the capability of presented three studies to argue, for instance, how identity information in turn influences emotional expression processing or perception/recognition (e.g., Ellamil, Susskind & Anderson, 2008).

5.5 Methodological Implications

The three studies in the thesis are closely connected in research rationale. The development and modification of the experimental designs underwent meticulous piloting and selection. These considerations are discussed below, with the aim of providing valuable insights for choosing experiment designs or testing environments in future studies.

5.5.1 Experimental paradigms

The research questions in Studies 1 and 2 dictated our choice of a traditional encoding-recognition task for use, where the exemplar variance needed to be experimentally manipulated during encoding. The decision to modify the paradigm into a continuous recognition task for Study 3 was aligned with its specific objectives, aiming to explicitly test participants' recognition of a set of novel exemplars from previously encountered identities. Here, we continued to use a full-balanced design, same as in Studies 1 and 2, so that each identity was presented six times (three exemplars by two repetitions). This choice was statistically beneficial, as each identity would have the same amount of conditions and essentially all of the trials are useable for analysis. However, the downside emerged with the accumulation of old-identity trials over the course of the experiment. To address this problem, we introduced three novel (filler) identities in the last 30 trials, effectively balancing the proportions of "old" and "new" identity trials. Results suggested that this manipulation was effective in the face run, as participants maintained a comparably high detection accuracy for first filler trials (FL1), as for the first exemplar trials (E1P1). An alternative we considered before was adding more novel identities as filler trials throughout the course, to balance out the old/new trial proportions. The drawback with this approach is also clear, that the experiment would have a large amount of filler identities but much fewer target identities to be used for the actual recognition analysis. Overall, the original encoding/recognition task proved to be reliable and yielded consistent results within and across modalities, in multiple experiments in Studies 1 and 2. This design has been employed in a number of studies that we mentioned throughout the thesis (e.g., Liu, Chen & Ward, 2015; Liu et al., 2016 for faces; Zäske et al., 2014; 2017 for voices).

Besides the classic encoding-recognition task, a novel face sorting task was developed (Jenkins et al., 2011) and has gradually garnered attention and more applications in both face and voice identity studies (e.g, Redfern & Benton, 2017; Gipson & Lampinen, 2020; Lavan, Burston & Garrido, 2018; Lavan et al., 2019b). The trade-off in this paradigm is that, it allows a closer and cleaner examination of which exemplars are exactly perceived as the same person, but usually is able to examine very limited number of actual identities (2 identities in most studies). Noticeably, there have been a few standardized tests for unfamiliar face and voice recognition, such as Cambridge Face Memory Test (CFMT) (Duchaine & Nakayama, 2006) and Glasgow Voice Memory Test (GVMT) (Aglieri et al., 2017). However, these tests mostly do not include stimuli with emotional expressions. Hence, our experimental paradigms seem appropriate option as of now, if one is interested in investigating emotional expression-affected identity memory.

5.5.2. Testing platforms

The two behavioral studies were conducted on different platforms. Study 1 followed the traditional approach of in-lab testing, prioritizing control and consistency in testing equipment and environment. However, due to the halt of in-person activities during COVID-19, we had to shift the testing protocol for Study 2 to an online platform. The pros and cons of online testing have been extensively reviewed (e.g., Finley & Penningroth, 2015; Dandurand, Shultz & Onishi, 2008). Despite the challenges, the shift from in-lab to online testing significantly accelerated the recruitment process, allowing us to achieve our target sample size much faster. As more studies have gradually embraced online testing to replace in-lab procedures, an increasing number of reports showcase comparable behavioral results from both testing methods (e.g., Dandurand, Shultz & Onishi, 2008; Buso et al., 2021; Schidelko et al., 2021; Nussenbaum et al.., 2020), as was the case in my studies.

I attribute the comparable results obtained from Studies 1 and 2 to two crucial factors: careful adaptation of the experimental implementation and a homogenous participant pool. We conducted extensive debugging and piloting to ensure an identical version of the online test, overcoming certain function differences between Matlab-based Psychtoolbox-3 and Javascript-based jsPsych environments. Additionally, as pointed out in a recent study (Uittenhove, Jeanneret & Vergauwe, 2023), the testing participant pool is more important than the testing platforms. The majority of the participants across my studies were college students, ensuring a

homogenous participant base. Based on our experiential insights gained from conducting experiments in both approaches, I am inclined to advocate for the utilization of online studies, especially for experiments encompassing less complex designs and/or fewer participant constraints. Moreover, it is always advisable to pilot in both platforms to ensure an identical paradigm, the same response measures recorded in the same manner, and optimally comparable preliminary data.

5.6 Limitations and Future Directions

5.6.1 Restricted use of stimulus database

The stimulus use across three studies are all from the same databases, namely, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018) for voice stimuli, and the Karolinska Directed Emotional Faces (KDEF, Lundqvist, Flykt & Öhman, 1998) for face stimuli. It provided good control in comparing results from multiple studies, but also raised concerns of potential stimulus or database-specific effects driving the experiment findings. We did make efforts to conduct stimulus-based feature analysis (in Studies 1 and 2), to offer a better understanding of what experimental results are related to or independent of perceptual features. Moreover, I have argued that the key findings of the exemplar variance advantage are not likely to be stimulus-database specific, due to its cross-modality presence. Nonetheless, future replication studies using different stimuli sets are welcomed, for example, Amsterdam Dynamic Facial Expression Set (van der Schalk, Hawk, Fischer & Doosje, 2011) and Berlin Database of Emotional Speech (Burkhardt et al., 2005), or unused modality (singing voices) of the RAVDESS itself.

5.6.2 Individual differences

As reported and discussed in Study 3, we have probed some influence of individual difference in language familiarity (English native speakers or not), particularly on cross-emotion voice recognition. Given that we did not have enough demographic data or adequate sample size, the significantly better cross-emotion recognition in E2P1 trials for non-native English speakers, warrant future studies to fully examine if language familiarity (e.g., Perrachione & Wong, 2007; Bregman & Creel, 2014; Xie & Myers, 2015) interferes with cross-expression voice learning, and if it further influences the observed exemplar variance advantage in Studies 1 and 2.

Following the broad avenue of examining individual differences in voice recognition, another individual trait we have indeed thought about earlier, that deems very relevant to both emotion perception and identity recognition, is autism symptoms, or autistic-like traits in healthy controls. Autism spectrum disorders (ASD) are characterized by deficits in social communication and/or interactions (DSM-5, American Psychiatric Association, 2013). The two core components of the thesis, identity recognition of face and voice, and emotional expressions, both play important roles in effective social interactions. Not surprisingly, a large body of behavioral studies have demonstrated that face and voice memory can be impaired in people with ASD (e.g., Ipser et al., 2016; Weigelt, Koldewyn, & Kanwisher, 2012 for a review on face; Schelinski, Roswandowitz & von Kriegstein, 2017), or affected by autistic-like traits in typically developed cohorts (Rhodes et al., 2013; Davis et al., 2017; Skuk et al., 2019). On the other hand, disruptions in emotion processing and recognition are also commonly found in individuals with ASD or high autistic traits (e.g., Pazhoohi, Forby, & Kingstone, 2021). This raises the question of whether people with high autistic traits would experience a greater difficulty in identity recognition involving changes in emotional expression. Future experiments on this research question will advance our understanding in autistic-trait related disruptions in processing emotional expression and identity.

Another relevant factor that our three studies were not able to fully address is sex-based influences on voice and face recognition. Studies 1 and 3 were relatively balanced in participant sex, but both did not contain a sample size large enough in each participant gender group to probe potential performance differences. Study 2, as mentioned in its own limitations, suffered from a heavy participant imbalance with a majority of participants being female. Past research indeed intensively examined such influences, of participant gender, stimulus gender, or combination of both genders (own-gender or other-gender effects, e.g., McKelvie, 1987; Herlitz & Lovén, 2013 for a review), in memory (e.g., Armony & Sergerie, 2007; Skuk & Schweinberger, 2013; Patel, Fredborg, & Girard, 2023; Herlitz & Lovén, 2013; Lovén, Herlitz, & Rehnman, 2011; Mukudi & Hills, 2019) and emotion studies (e.g., Montagne et al., 2005; Filkowski et al., 2017; Kret & de Gelder, 2012). Among the large body of studies in this avenue, some tested on memory of emotional faces (e.g., Armony & Sergerie, 2007; Patel, Fredborg, & Girard, 2023), specifically same-stimulus recognition, leaving out the core behavioral performance we have assessed throughout this thesis: cross-expression (or expression-

independent) recognition. Hence, it is important and useful for future studies to fill in this gap, to disentangle such sex-based influences on expression-independent face, and moreover, voice memory, which limited research has examined as of now.

5.6.3 Neural difference in recognizing high- and low-variability learned identities

Study 3 focused on investigating the explicit identity perception and neural correlates when encountering/learning a novel emotional exemplar, with the aim of gaining a better understanding of the encoding process that leads to the exemplar variance advantage in Study 2. The ultimate goal remains an investigation of neural differences between high- and low-variability encoded identities. Our hypothesis based on the strong and consistent behavioral advantage of high-variability encoding would be, high-variability encoded identities elicits and resemble a more familiar(isc) face/voice activation pattern (see Natu & O'Toole, 2011; Ramon & Gobbini, 2018), compared to low-variability encoded identities. Hence, a follow-up imaging study should resemble the experimental design of Study 2, allowing us to directly examine neural correlates of the exemplar variance advantage, particularly during the recognition process.

5.6.4 Learning identities through multi-modal inputs

Throughout all three studies, face and voice recognition were tested separately. I chose to design and conduct unimodal experiments in the thesis, intending to compare recognition performance between auditory and visual modalities. In reality, however, the most common way individuals encounter and become familiar with others is through concurrent audiovisual signals. Thus, it is crucial and with realistic implications to pursue similar research questions in the context of multi-modal encoding (learning) of individual identities: What are the benefits (or disadvantages) of multimodal learning against unimodal learning; Whether the observed advantage of emotional exemplar variance remains present after learning with multiple audiovisual exemplars (excerpts), as opposed to learning with single repeated exemplars; Whether exemplar arousal level continues to interfere with the recognition of identities learned through repeated audiovisual exemplars.

Indeed, there has been a gradual increase in recent research in examining the learning and recognition of identities encoded in multiple modalities, and the differences between multi- and single-modal learning and their subsequent recognition. Give that face serves as a stronger and

more accessible identity cue than voice (e.g., Stevenage & Neil, 2014; Barsics, 2014), studies have unsurprisingly found that multimodal learning would only significantly influence voice recognition, but not face recognition (e.g., McAllister et al., 1993; Stevenage, Howland & Tippelt, 2011). However, the exact influence on voice recognition varies across studies. Some behavioral and neuroimaging studies supported a benefit of audio-visual encoding (e.g., Maguinness, Schall, & von Kriegstein, 2021; von Kriegstein et al., 2008; Zäske, Mühl, & Schweinberger, 2015), others indicated an impairment (e.g., Lavan et al., 2023; Cook & Wilding, 2001; Tomlin, Stevenage, & Hammond, 2017). As of now, reconciling these contradictory impacts on voice recognition remains challenging, despite noticeable differences in for example, experimental methodology (e.g., task variations in encoding and/or recognition, various delays in between) and stimulus use. One possible mechanism is based on attention reallocation (e.g., Lavan et al., 2023; Zäske, Mühl, & Schweinberger, 2015), suggesting that audiovisual co-presentation may initially shift attentional resources towards face in an automatic fashion, hence interfere with encoding of the voice (Cook & Wilding, 1997, 2001). Once a better learning or audio-visual association is built (through longer or more exposures), attention may be reallocated from face back to voice stimuli, which in return facilitated subsequent memory. This theoretical hypothesis was supported by empirical findings from Zäske and colleagues (2015).

If this is indeed the case, the usage of emotional materials would provide substantial help to confirm (or contradict) this mechanism, given the tight relation between emotion (arousal) and attention, and a similar attention-based mechanism we proposed to interpret results in the thesis (mainly in Study 2). Following this notion, we would expect, for instance, the *Uni*-high-arousal encoding condition could be further impaired with more attentional resources being directed to co-presented face stimulus, leading to an even larger advantage of multiple exemplar multimodal learning. Nonetheless, to the best of our knowledge, current research on audiovisual identity learning has yet to extensively use emotional materials or explicitly test emotional related effects. Hence, this line of research holds immense potential for advancing our understanding of the mechanisms underlying the divergent results particularly on voice recognition between multimodal and unimodal learning. Additionally, it will contribute significantly to expand our knowledge of the exemplar variance advantage reported here (emotional) and previously (general) in a multimodal learning scenario.

5.7 Conclusions

This thesis highlights the complex influences of changes in emotional expressions, on identity learning and recognition across modalities. A single change in emotional expression between learning and test can interfere with newly learned identities, especially in the case of learning high-arousal exemplars. However, learning through more exemplar variances can compensate for this interference and lead to improved cross-emotion recognition of both face and voice. In face particular, the advantage of learning variance occurs both behaviorally and neurally, as early as after learning two distinct exemplars of an individual. The findings of this thesis put into perspective our understanding of the complicated interactive/independent relationships in identity and emotional expression processing, and furthermore, provide an arousal-based account in explaining some inconsistent effects on emotional-expression-independent recognition reported in our work and previous studies.

General Reference List

Aglieri, V., Watson, R., Pernet, C. et al. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behav Res 49*, 97–110.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author.

Andics, A., McQueen, J. M., Petersson, K. M., et al. (2010). Neural mechanisms for voice recognition. *NeuroImage, 52*, 1528-40.

Andrews, T., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*(10), 2041-2050.

Armony, J. L., & Sergerie, K. (2007). Own-sex effects in emotional memory for faces. *Neuroscience Letters*, *426*(1), 1-5.

Armony, J. L., Chochol, C., Fecteau, S., & Belin, P. (2007). Laugh (or cry) and you will be remembered: Influence of Emotional Expression on Memory for Vocalizations. *Psychological Science, 18*(12), 1027–1029.

Armony, J. L., Vuilleumier, P., Drive, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron, 30*(3), 829-841.

Aubé, W., Peretz, I., & Armony, J. L. (2013). The effects of emotion on memory for music and vocalizations. *Memory, 21*(8), 981-990.

Baddeley, A., & Woodhead, M. (1983). Improving face recognition. In S. M. A. Lloyd-Bostock, & B. R. Clifford (eds). *Evaluating eyewitness evidence*. UK: John Wiley and Sons.

Barrett, L. F., & Russell, J. A. (1999). The Structure of Current Affect: Controversies and Emerging Consensus. *Current Directions in Psychological Science*, *8*(1), 10–14.

Barsics, C. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244-254.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF, 74*, 110-120.

Belin, P., & Zatorre, R. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport 14*(16), 2105-2109.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711–725.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135.

Belin, P., Zatorre, R., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*(1), 17-26.

Belin, P., Zatorre, R., Lafaille, P. et al. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*, 309–312.

Benson, P. J., & Perrett, D. I. (1993). Extracting prototypical facial images from exemplars. *Perception, 22*(3), 257–262.

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: merging evidence from patients and healthy individuals. *Neuroscience & Biobehavioral Reviews*, *47*, 717-734.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual cognition*, *10*(5), 527-536.

Braje, W. L., Kersten D., Tarr M. J., & Troje, N. F. (1998). Illumination effects in face recognition. *Psychobiology, 26*, 371-380.

Bregman, M. R., & creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition, 130*(1), 85-95.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116.

Bruce, V., & Burton, M. (2002). Learning new faces. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 317–334). MIT Press.

Bruce, V., & Valentine, T. (1985). Identity priming in the recognition of familiar faces. *British Journal of Psychology, 76*(3), 373–383.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Bruce, V., Healey, P., Burton, A. M., et al. (1991). Recognising facial surfaces. *Perception, 20*, 755-69.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied,*

7(3), 207–218.

Bruyer, R., Laterre, C., Seron, X., et al. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition*, *2*(3), 257–284.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8*(1-2), 135-148.

Bulthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are 3-dimensional objects represented in the brain. *Cerebral Cortex, 5*, 247-260

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech*, *5*, 1517-1520.

Burton, A. M. (1994). Learning new faces in an interactive activation and competition model. *Visual Cognition, 1*(2-3), 313-348.

Mike Burton, A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science, 23*(1), 1-31.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *Cognitive Science, 23*(1), 1-31.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*(4), 943-958.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science, 10*(3), 243–248.

Buso, I. M., Di Cagno, D., Ferrari, L., et al. (2021). Lab-like findings from online experiments. *Journal of the Economic Science Association*, *7*(2), 184-193.

Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: Extension and limits. *Mem Cogn*, *27*, 139–151.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis on facial expressions. *Vision research*, *41*(9), 1179-1208.

Calder, A., & Young, A. (2005) Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci,* 6, 641–651.

Chen, W., & Liu, C. H. (2009). Transfer between pose and expression training in face recognition. *Vision Research*, *49*(3), 368-373.

Cook, S., & Wilding, J. (1997). Earwitness testimony 2. Voices, faces and context. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(6), 527-541.

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92*(4), 617-629.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior research methods*, *40*(2), 428-434.

D'Argembeau, A., & van der Linden, M. (2007). Facial expressions of emotion influence memory for facial identity in an automatic way. *Emotion, 7*(3), 507–515.

D'Argembeau, A., van der Linden, M., Comblain, M., & Etienne, A. M. (2003). The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cognition and Emotion, 17*(4), 609-622.

Darwin, C. (1872). *The expression of the emotions in man and animals.* John Murray.

Davis, J., McKone, E., Zirnsak, M., et al. (2017). Social and attention-to-detail subclusters of autistic traits differentially predict looking at eyes and face identity recognition ability. *British Journal of Psychology, 108*(1), 191-219.

Duchaine, B., & Nakayama, K. (2006). The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44(4), 576-585.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), 169–200.

Ellamil, M., Susskind, J. M., & Anderson, A. K. (2008). Examinations of identity invariance in facial expression adaptation. *Cognitive, Affective, & Behavioral Neuroscience, 8*, 273-281.

Ellis, A. W., Young, A. W., Flude, B. M., & Hay, D. C. (1987). Repetition priming of face recognition. *Quarterly Journal of Experimental Psychology, 39*(2), 193–210.

Ellis, H. D., Quayle, A. H., & Young, A. W. (1999). The emotional impact of faces (but not names): Face specific changes in skin conductance responses to familiar and unfamiliar people. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 18*(1), 88–97.

Filkowski, M. M., Olsen, R. M., Duda, B., Wanger, T. J., & Sabatinelli, D. (2017). Sex

differences in emotional perception: Meta analysis of divergent activation. *NeuroImage, 147*, 925-933.

Finley, A., & Penningroth, S. (2015). Online versus in-lab: Pros and cons of an online prospective memory experiment. *Advances in psychology research*, *113*, 135-162

Fitousi, D., & Wenger, M. J. (2013). Variants of independence in the perception of facial identity and expression. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 133–155.

Garrido, L., Eisner, F., McGettigan, C., et al. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia, 47*(1), 123-131.

Gipson, N. I., & Lampinen, J. M. (2020). Within lab familiarity through ambient images alone. *Visual Cognition, 28*(3), 165–179.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448-458.

Gorno-Tempini, M. L., Price, C. J., Josephs, O., et al. (1998). The neural systems sustaining face and proper-name processing. *Brain, 121*, 2103-2118.

Hadj-Bouziane, F., Bell, A. H., Knusten, T. A., Ungerleider, L. G., & Tootell, R. B. (2008). Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(14), 5591–5596.

Hagemann, J., Straube, T., & Schulz, C. (2016). Too bad: Bias for angry faces in social anxiety interferes with identity processing. *Neuropsychologia, 84*, 136-149.

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warren, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia, 48*(4), 1104-1114.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223–233.

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences, 7*, 425-427

Herlitz, A. & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: a meta-analytic review. *Visual Cognition, 21*(9-10), 1306-1336

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of*

Experimental Psychology: Human Perception and Performance, 22, 986-1004.

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62*, 201-222

Ipser, A., Ring, M., Murphy, J., et al. (2016). Similar exemplar pooling processes underlie the learning of facial identity and handwriting style: Evidence from typical observers and individuals with Autism. *Neuropsychologia, 85*, 169-176.

Jackson, M. C., Linden, D. E. J., & Raymond, J. E. (2014). Angry expressions strengthen the encoding and maintenance of face identity representations in visual working memory. *Cognition and Emotion, 28*(2), 278–297.

Jackson, M. C., Wu, C. Y., Linden, D. E. J., & Raymond, J. E. (2009). Enhanced visual short-term memory for angry faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 363–374.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philos Trans R Soc Lond B Biol Sci, 366*(1571), 1671-83.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577–596.

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception*, *33*(4), 399-408

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, *15*(4), 241-252.

Kensinger, E. A. (2007). Negative emotion enhances memory accuracy: behavioral and neuroimaging evidence. *Current Directions in Psychological Science, 16*(4), 213-218.

Kensinger, E. A. (2009). Remembering the details: effects of emotion. *Emotion Review*, *1*(2), 99-113.

Kensinger, E. A., & Schacter, D. L. (2005). Retrieving accurate and distorted memories: Neuroimaging evidence for effects of emotion. *NeuroImage, 27*(1), 167-177.

Kerstholt, J. H., Jansen, N. J., van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology, 18*(3), 327-336.

Knapp, B. R., Nosofsky, R. M., & Busey, T. A. (2006). Recognizing distinctive faces: A hybrid-similarity exemplar model account. *Memory & Cognition, 34*(4), 877–889.

Kolers, P. A., Duchnicky, R. L., & Sundstroem, G. (1985). Size in the visual processing of faces and words. *Journal of Experimental Psychology: Human perception and performance*, *11*(6), 726–751.

Kottoor, T. M. (1989). Recognition of faces by adults. *Psychological Studies, 34*(2), 102–105.

Kreitewolf, J., Friederici, A. D., & von Kriegstein, K. (2014). Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *NeuroImage*, *102 Pt 2*, 332–344.

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Front. Psychol., 8*, 1584.

Kret, M. E., & de Gleder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia, 50*(7), 1211-1221.

Kringelbach, M. L., & Berridge, K. C. (2009). Towards a functional neuroanatomy of pleasure and happiness. *Trends in cognitive sciences*, *13*(11), 479–487.

Krouse, F. L. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology, 66*(5), 651–654.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54-64.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Front. Psychol. 2,* 175.

Latinus, M., Crabbe, F., & Belin, P. (2009). FMRI investigations of voice identity perception. *NeuroImage, 47,* S156.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*(12), 1075-1080.

Lavan, N., Burston, L. F. K., Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593.

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019b). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240–2248.

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications, 10*, 2404.

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019a). The effects of high variability training on voice identity learning. *Cognition, 193*, 104026.

Lavan, N., Ramanik Bamaniya, N., Muse, M. M., Price, R. L. M., & Mareschal, I. (2023). The effects of the presence of a face and direct eye gaze on voice identity learning. *British Journal of Psychology, 114*(3), 537-549.

Lee, T. H., Sakaki, M., Cheng, R., et al. (2014). Emotional arousal amplifies the effect of biased competition in the brain. *Social Cognitive and Affective Neuroscience, 9*(12), 2067-2077.

Leppänen, J. M., & Hietanen, J. K. (2007). Is there more in a happy face than just a big smile? *Visual Cognition,15(*4), 468-490.

Linden, D. E. J, Thornton, K., Kuswanto, C. N., et al. (2011). Brain's Voices: Comparing Nonclinical Auditory Hallucinations and Imagery. *Cerebral Cortex*, *21*(2), 330–337.

Liu, C. H., Chen, W. F., & Ward, J. (2014). Remembering faces with emotional expressions. *Frontiers in Psychology*, *5*, 1439.

Liu, C. H., Chen, W. F., & Ward, J. (2015). Effects of exposure to facial expression variation in face learning and recognition. *Psychological Research, 79*(6), 1042-53.

Liu, C. H., Chen, W. F., Ward, J., & Takahashi, N. (2016). Dynamic emotional faces generalize better to new expression but not to a new view. *Scientific Reports, 6*, 31001.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE, 13*(5).

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77–100.

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women's own-gender bias in face recognition memory. *Experimental Psychology, 58*(4), 333.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Lundqvist, D., Juth, P., & Öhman, A. (2014). Using facial emotional stimuli in visual search experiments: The arousal factor explains contradictory results. *Cognition and*

Emotion, 28(6), 1012-1029.

Maguinness, C. , Schall, S. , & von Kriegstein, K. (2021). Prior audio-visual learning facilitates auditory-only speech and voice-identity recognition in noisy listening conditions. *PsyArXiv*.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia, 116*(Part B), 179–193.

Malone, D. R., Morris, H. H., Kay, M. C., & Levin, H. S. (1982). Prosopagnosia: a double dissociation between the recognition of familiar and unfamiliar faces. *Journal of Neurology, Neurosurgery & Psychiatry, 45*, 820-822.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: Freeman

Mather, M., & Sutherland M. (2011). Arousal-biased competition in perception and memory. *Perspectives on Psychological Science*, *6*(2), 114-133.

Matthews, C. M., Davis, E. E., & Mondloch, C. J. (2018). Getting to know you: The development of mechanisms underlying face learning. *Journal of Experimental Child Psychology, 167,* 295–313.

McAllister, H. A., Dale, R. H., Bregman, N. J., McCabe, A., & Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology*, *14*(2), 161-170.

McKelvie, S. J. (1987). Sex differences, lateral reversal, and pose as factors in recognition memory for photographs of faces. *Journal of General Psychology*, *114*, 13–37.

Montagne, B., Kessels, R. P. C., Frigerio, E., de Haan, E. H. F., & Perrett, D. I. (2005). Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? *Cognitive Processing, 6,* 136-141

Mukudi, P. B. L., & Hills, P. J. (2019). The combined influence of the own-age, -gender, and -ethnicity biases on face recognition. *Acta Psychologica, 194*, 1-6.

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology, 25*(1), 29-34.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and*

Performance, 41(3), 577–581.

Nakamura, K., Kawashima, R., Sugiura, M., et al. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia, 39*(10), 1047-1054.

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: a review and synopsis. *British Journal of Psychology, 102*(4), 726–747.

Niedenthal, P. M., & Wood, A. (2019). Does emotion influence visual perception? Depends on how you look at it. *Cognition and Emotion, 33*(1), 77-84.

Nomi, J. S., Rhodes, M. G., & Cleary, A. M. (2013). Emotional facial expressions differentially influence predictions and performance for face recognition. *Cognition and Emotion, 27*(1), 141-149.

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., et al. (2020). Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Psychology, 6*(1), 17213.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376.

O'Toole, A. J., Edelman, S., & Bulthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 38*, 2351-2363.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness memory. *Journal of Investigative Psychology and Offender Profiling, 10*(1), 57–70.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition, 143,* 36–40.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85*, 913-925.

Patel, R., Fredborg, B. K., & Girard, T. A. (2023). Modulation of emotion-enhanced recollection by gender and task instructions. *Emotion, 23*(6), 1764–1772.

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: evidence from Chinese and British listeners. *Cognition & emotion*, *28*(2), 230–244.

Pazhoohi, F., Forby, L., & Kingstone, A. (2021). Facial masks affect emotion recognition in the general population and individuals with autistic traits. *PLoS One*, *16*(9), e0257740.

Pernet, C. R., McAleer, P., Latinus, M., et al. (2015). The human voice areas: Spatial

organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage, 119*, 164-174.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*(8), 1899–1910.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J Acoust Soc Am*, *130*(1), 461–472.

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1950–1960.

Perrachione, T.K. (2018). Recognizing speakers across languages. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception*, Oxford: Oxford University Press.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Human neurobiology*, *3*(4), 197–208.

Peynircioğlu, Z. F., Rabinovitz B. E., & Repice J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice, 31*(2), 256.e13-17.

Pichora-Fuller, M.K., Dupuis, K., & Smith, L. (2016). Effects of vocal emotion on memory in younger and older adults. *Experimental Aging Research, 42*(1), 14-30.

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition, 26*(3), 179–195.

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *I-perception, 8*(5).

Rhodes, G., Jeffery, L., Taylor, L., & Ewing, L. (2013). Autistic traits are linked to reduced adaptive coding of face identity and selectively poorer face recognition in men but not women. *Neuropsychologia, 51*(13), 2702-2708.

Righi, S., Marzi, T., Toscani, M., et al. (2012). Fearful expressions enhance recognition memory: Electrophysiological evidence. *Acta Psychologica, 139*(1), 7-18.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897-905.

Roark, D. A., O'Toole, A. J., & Abdi, H. (2003). Human recognition of familiar and unfamiliar people in naturalistic video. *IEEE International SOI Conference. Proceedings,* 36-41.

Robinson, M. D., Storbeck, J., Meier, B. P., & Kirkeby, B. S. (2004). Watch out! That could be dangerous: Valence-arousal interactions in evaluative processing. *Personality and Social Psychology Bulletin*, *30*(11), 1472-1484.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145–172.

Sanders, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*(4), 317-352.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65*(1), 111-6. https://doi.org/10.1037/0021-9010.65.1.111

Savaskan, E., Müller, S. E., Böhringer, A., et al. (2007). Age determines memory for face identity and expression. *Psychogeriatrics, 7*(2), 49-57.

Schelinski, S., Roswandowitz, C., & von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism research, 10*(1), 155–168.

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. *Front. Psychol, 12,* 703238.

Schimer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends Cogn Sci, 21*(3), 216-228.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research, 40*(2), 453-463.

Sergent, J., Ohta, S., Macdonald, B., & Zuck, E. (1994) Segregated processing of facial identity and emotion in the human brain: A PET study. *Visual Cognition, 1*:2-3, 349-369.

Sergerie, K., Lepage, M., & Armony, J. L. (2005). A face to remember: emotional expression modulates prefrontal activity during memory formation. *NeuroImage, 24*(2), 580-5.

Sidtis, D. V. L., & Zäske, R. (2021). Who we are: Signaling personal identity in speech. In J. S. Pardo, L. C. Nygaard, R. E. Remez, et al. (Eds.), *The handbook of speech perception* (pp. 365-397). John Wiley & Sons.

Skuk, V. G. & Schweinberger, S. R. (2013). Gender differences in familiar voice identification.

Hearing Research, 296, 131-140.

Skuk, V. G., Palermo, R., Broemer, L., & Schweinberger, S. R. (2019). Autistic Traits are Linked to Individual Differences in Familiar Voice Identification. *Journal of autism and developmental disorders*, *49*(7), 2747–2767.

Smith, H. M. J., Baguley, T. S., Robson, J., et al. (2018). Forensic voice discrimination: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272-287.

Soto F. A., Vucovich L., Musgrave R., Ashby F. G. (2015). General recognition theory with individual differences: a new method for examining perceptual and decisional interactions with an application to face perception. *Psychon. Bull. Rev,* 22, 88–111.

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*(3), 266–281.

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, *25*(1), 112-118.

Talmi, D., Anderson, A. K., Riggs, L., Caplan, J. B., & Moscovitch, M. (2008). Immediate memory consequences of the effect of emotion on attention to pictures. *Learning & Memory, 15*(3), 172–182.

Tomlin, R. J., Stevenage, S. V., & Hammond, S. (2017). Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition, 25*(4-6), 629–643.

Tranel, D., Damasio, A. R., & Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, *38*(5), 690–696.

Tyng, C. M., Amin, H. U., Saad, M. N. M., & Malik, A. S. (2017). The influences of emotion on learning and memory. *Front. Psychol, 8*, 1454.

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, *6*(1), 13.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 43A*(2).

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion, 11*(4), 907–920.

van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1*(2), 185-195.

von Kriegstein, K., Dogan, O., Grüter, M., et al. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6747–6752.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends Cogn. Sci.* 9, 585–594.

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: a review of behavioral studies. *Neuroscience and biobehavioral reviews*, *36*(3), 1060–1084.

Xie, X., & Myers, E. B. (2015). General Language Ability Predicts Talker Identification. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX.

Yarmey A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology, 8*(5), 453-464.

Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society, 31*(4), 421-428.

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in cognitive sciences*, *24*(5), 398-410

Yuan, J., Tian, Y., Huang, X., et al. (2019). Emotional bias varies with stimulus type, arousal, and task setting: Meta-analytic evidences. *Neuroscience & Biobehavioral Reviews, 107*, 461-472.

Zadra, J. R., & Clore, G. L. (2011). Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, *2*(6), 676-685.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and Exemplar Accounts of Category Learning and Attentional Allocation: A Reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1160–1173.

Zäske, R., Hasan, B. A. S., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex, 94*, 100-112.

Zäske, R., Mühl, C., & Schweinberger, S. R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLoS One, 10*(11), 1–12.

Zäske, R., Volberg, G., Kovacs, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience, 34*(33), 10821-31.

Zych, A. D., & Gogolla, N. (2021). Expressions of emotions across species. *Current opinion in neurobiology*, *68*, 57–66.