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

# Influence of emotional prosody, content and repetition on memory recognition of speaker identity

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

[1]Douglas Mental Health University Institute, Montreal, Quebec, Canada

[2]Integrated Program in Neuroscience, McGill University, Montreal, Canada

[3]Department of Psychiatry, McGill University, Montreal, Canada

## 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.

# Influence of emotional prosody, content and repetition on memory recognition of speaker identity

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).

## 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.

## Methods

### *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.

### *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.

### *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 $g_s = 2.65$).

***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 Table 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.

### *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.

### *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.

## Results

### *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.

### *Recognition*

Response accuracies for all conditions in each group are summarized in Table 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 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 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 Fig. 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, $z = 4.62$, $p < .001$).

[insert Figure 1.]

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. 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). 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 Fig. 1b.

**Table 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.

## 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 Supplementary Table 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 et al., 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.

## 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.

## Methods

### *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.

### *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.

### *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.

***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.

## Results

### *Encoding*

Changes in encoding RTs across the four presentations of speakers are illustrated in Fig. 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 Fig. 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.

[insert Figure 2.]

### *Recognition*

Response accuracy for each condition (overall, *Multi* and *Uni*) is shown in Table 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 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 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).

## 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.

## 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.

## 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.

## 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.

**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.

**Conflict of interest**

Authors declare no conflicts of interest.

## 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

**Figures**

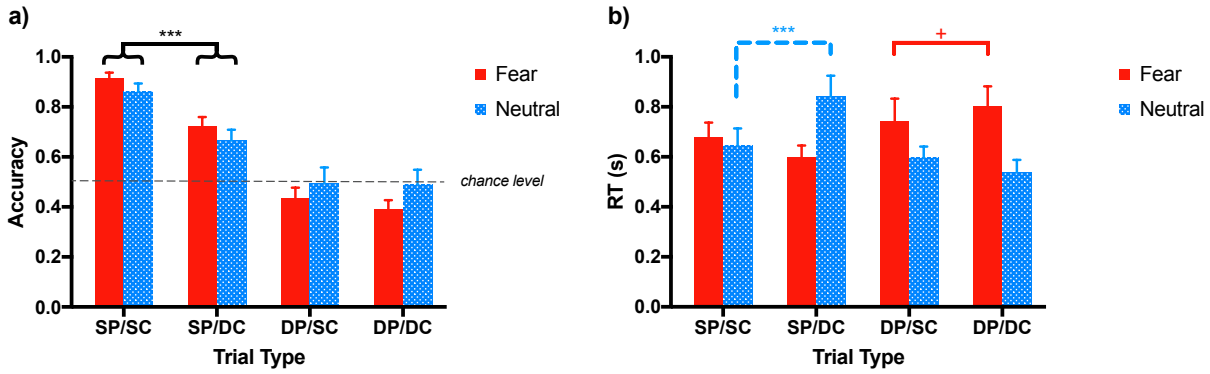


**Figure 1.** Recognition accuracy (a) and response times (b) in Experiment 1. Average (a) accuracy and (b) RTs for each trial type in each participant group. Horizontal lines show the significant differences between conditions in post-hoc tests. Horizontal dashed lines in (a) represents chance-level 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.

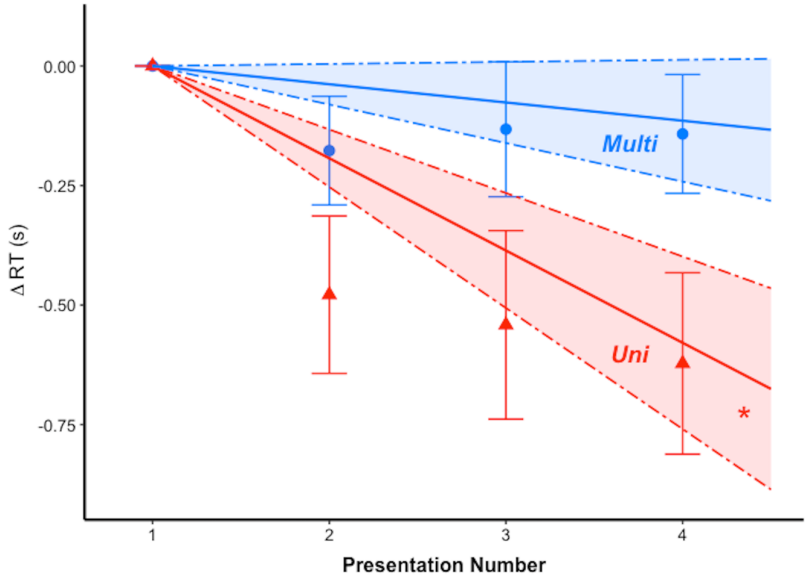


**Figure 2.** Changes in response times (RTs) during encoding in Experiment 2 for the *Uni* (red triangles) and *Multi* (blue circles) conditions (relative to the first presentation). The solid lines represent the subject- and speaker-averaged slopes obtained in the LMMs (see Methods for details). Dashed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

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

**Supplementary Table 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* | $\eta_p^2$ | F(1,11) | *p* | $\eta_p^2$ |
| 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)