The effects of frequency and type of feedback on L2 phoneme learning

Haruka Saito

School of Communication Sciences and Disorders

McGill University, Montreal

March 2021

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

Table of Contents

Abstract

In the research and practice of instructed second language (L2) acquisition, feedback has been primarily used to correct learners' errors in order to achieve higher linguistic attainment. Research has shown that corrective feedback (CF) is deemed effective on learning of many components of L2, including pronunciation, while teachers often fear that the provision of CF may negatively impact learners' motivation. Feedback has also been intensively investigated in different areas of research, such as (speech) motor skill learning, where the role of feedback as error correction has been challenged and the motivational effect of feedback has attracted attention. Inspired by studies in the two different domains of learning, the present thesis investigated two features to further understand the effects of feedback in L2 pronunciation training: frequency (frequent vs. reduced) and type (non-corrective vs. corrective) of feedback.

The first study examined whether the frequency of CF (i.e., how often errors are corrected) affected learners' improvement in pronunciation accuracy, motivation, and error detection ability. The results from the first study showed that when the frequency of CF was reduced to 50%, it was still effective on accuracy when the type of feedback given was recast (the provision type feedback), but not when it was prompt (the elicitation type feedback). Prompts increased participants' awareness for their own errors in pronunciation, but may negatively affect learners' motivation when given frequently. Recasts did not improve learners' error detection ability, but increased motivation when given frequently. These results led us to conclude that that the two types of CF may encourage learners to employ different learning strategies.

The second study investigated the effectiveness of non-corrective feedback (NCF) compared to CF. The results from the second study revealed that NCF was in fact effective much like CF was, but only when it was of the elicitation type (i.e., a NC version of prompts). NCF also increased learners' error detection ability and motivation, while the provision-type non-corrective feedback (i.e., NC repetition) presented a striking contrast: it did not positively affect pronunciation accuracy, error detection, or motivation. These results, in line with the first study, suggested the differential effects of the elicitation and provision types of feedback.

The third study is a complimentary analysis to examine the acoustics of participants' productions of the target sound of the present study, the Japanese geminate-singleton contrast. We found that participants in the present studies seemed correctly aware that consonant closure duration is the primary cue, and did not utilize other secondary acoustic cues. The study also revealed that participants seemed to face difficulty in producing the target sound when it occurred in a specific position in the word, which may be partially due to articulatory difficulty.

Taken together, the present thesis contributes to the further understanding of the effect of feedback in L2 speech learning. Practically, the papers revealed promising feedback conditions that optimize L2 pronunciation training in terms of frequency and types of feedback. From a research standpoint, the present thesis challenged the existing context where the role of feedback in instructed L2 acquisition is primarily considered to be error correction and further discussed the differential roles that different types of feedback may serve in L2 pronunciation learning.

Résumé

Dans la recherche et l'exercice de l'apprentissage de la langue seconde (L2), la rétroaction est principalement utilisée pour corriger les erreurs des apprenants afin d'atteindre un niveau linguistique plus élevé. La recherche montre que la rétroaction corrective (RC) est jugée efficace dans l'apprentissage de nombreuses composantes de la L2, y compris la prononciation, malgré le fait que les enseignants craignent souvent que son utilisation puisse avoir un impact négatif sur les aspects affectifs des apprenants, tels que la motivation. D'autre part, la rétroaction a également été étudiée de manière intensive dans différents domaines de recherche, tels que l'apprentissage des habiletés motrices (y compris la parole), où le rôle de la rétroaction en tant que correction d'erreur a été remis en question alors que son effet motivationnel a attiré l'attention. Inspirée par des études dans ces deux domaines d'apprentissage différents, la présente thèse a permis d'examiner deux caractéristiques qui contribuent à une meilleure compréhension de l'effet et du rôle de la rétroaction dans l'apprentissage de la prononciation en L2 : la fréquence (fréquente ou réduite) et le type (non correctif ou correctif) de la rétroaction.

La première étude visait à déterminer si la fréquence de la RC (c.-à-d., la fréquence à laquelle les erreurs sont corrigées) affectait l'amélioration de la précision de la prononciation, de la motivation et de la capacité de détection des erreurs des apprenants. Les résultats de la première étude ont montré que lorsque la fréquence de la RC était réduite à 50 %, elle était toujours efficace lorsque le type de rétroaction donnée était une reformulation (« *recasts* »), mais pas lorsqu'il s'agissait d'une incitation (« *prompt* »). Les incitations ont sensibilisé les participants à leurs propres erreurs de prononciation, mais elles peuvent avoir un effet négatif sur la motivation des apprenants lorsqu'elles sont fréquemment données. Les reformulations

n'affectaient pas la capacité de détection des erreurs des apprenants quelle que soit leur fréquence, et augmentaient la motivation des apprenants lorsqu'elles étaient données fréquemment. Ces résultats nous ont amenés à conclure que les deux types de RC peuvent encourager les apprenants à employer des stratégies d'apprentissage différentes.

La deuxième étude avait pour objectif d'examiner l'efficacité de la rétroaction non corrective (RNC) en comparaison avec la RC. Les résultats de la deuxième étude ont révélé que la RNC était aussi efficace que la RC, mais seulement lorsqu'elle était de type incitation. Les incitations non correctives ont également amélioré la capacité de détection des erreurs et la motivation des apprenants, tandis que les répétitions non correctives (c.-à-d., une version non corrective de reformulation) présentaient un contraste frappant : elles n'amélioraient pas la précision de la prononciation, la détection des erreurs ou la motivation. Ces résultats, conformément à la première étude, suggéraient que les types d'incitation et de répétition ont des effets différents sur la perception des erreurs et la motivation des apprenants.

La troisième étude est une analyse complémentaire visant à examiner l'acoustique des productions du son cible par les participants, soit les consonnes japonaises géminées et simples. Nous avons constaté que les participants semblaient conscients que le principal indice phonétique différenciant les consonnes géminées et simples est la durée de fermeture des consonnes, et qu'ils n'ont pas utilisé d'autres signaux acoustiques secondaires. L'étude a également révélé que les participants semblaient avoir des difficultés à produire le son cible lorsqu'il se produisait dans une position spécifique dans le mot, ce qui peut être en partie dû à une difficulté articulatoire.

Dans son ensemble, la présente thèse contribue à une meilleure compréhension de l'effet de la rétroaction dans l'apprentissage de la parole en L2. En pratique, les études ont révélé des

conditions de rétroaction prometteuses qui optimisent l'apprentissage de la prononciation en L2 en termes de fréquence et de types de rétroaction. Du point de vue de la recherche, la présente thèse a remis en question le contexte existant où le rôle de la rétroaction dans l'enseignement de la L2 est principalement considéré comme une correction d'erreur et a davantage discuté des différents rôles que peuvent jouer les différents types de rétroaction dans l'apprentissage de la prononciation en L2.

Acknowledgements and contributions

I would like to express my sincere gratitude to my supervisors, Dr. Shari Baum and Dr. Vincent Gracco, for their supervision throughout the entire duration of this thesis project. The proposal, data analyses, and manuscript of the present thesis were all shaped and revised based on their constructive input. Dr. Baum, who has been the most patient and supportive supervisor throughout my study at SCSD, has also provided a significant amount of editorial help in the process of writing. I would also like to thank Dr. Roy Lyster, who generously agreed to be my thesis committee member and provided invaluable feedback on the proposal, data analysis and manuscript of the thesis.

I would also like to extend my appreciation to my colleagues and friends who have helped me at various stages of this journey. I am thankful to my fellow PhD students who participated in the preliminary experiment that had shaped the initial proposal of this project. I would like to thank Dr. Sarah Colby, who helped me publish online advertisements for the participant recruitment, and the current and former members of Shari Baum lab, especially Dr. Annie Gilbert, who have always offered insightful feedback and encouragements whenever I presented my project. I am especially grateful to my friend and colleague in Japan, Dr. Keiko Hanzawa (Tokyo University of Science), who had conducted the recording of Japanese speakers on my behalf and have provided many insights and advises throughout the process.

Besides the immensely valuable assistance that I have described above, I, Haruka Saito, am solely responsible for all three papers, general introduction and general discussion in the present thesis, from conceiving the ideas, to preparing experimental materials, to collecting data, to analyses, and finally, to writing manuscripts.

Finally, I would like to express my tiny gratitude to my tiny cat, Neko. She did no contribution to the present thesis (as all cats do, she thinks such work is beneath her), but the presence of her helped me immeasurably in the process of writing during the almost year-long quarantine since the pandemic.

Acknowledgement of funding

Over the years working on the present thesis project, I have been supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) awarded to Dr. Baum, a Graduate Student Stipend from The Centre for Research on Brain, Language and Music (CRBLM), a Graduate Excellence Fellowship and a Travel Award from the School of Communication Sciences and Disorders, and a Doctoral Research Scholarship (bourse de doctorat en recherche) from the Fonds du Québec – Société et Culture (FRQSC).

General Introduction

The importance of L2 pronunciation and the role of training

Attaining accurate and natural L2 pronunciation has a significant impact on an L2 speaker's life. L2 speakers' divergence from the target phonetic features of their L2 often increases communication difficulty (Munro & Derwing, 1995; Adank et al., 2009; Floccia et al., 2006). In multiple studies, native listeners have been shown to spend more time and make more errors when comprehending sentences uttered by foreign-accented speakers than by native speakers with a standard accent (Munro & Derwing, 1995) and with a dialectal accent (Floccia et al., 2006), and disadvantages of foreign accent are further exacerbated in noisy conditions (Adank et al., 2009). These difficulties are perceived by foreign-accented speakers themselves: in surveys of non-native English speakers in Canada, over half of the respondents believed that their accent was a problem in communication (Derwing & Rossiter, 2002) and that they would be "more respected" if they spoke without a foreign accent (Derwing, 2003). Difficulty in communication has further impact on social behaviours and on the psychological status of L2 speakers (Derwing & Rossiter, 2002; Gluszek & Dovidio 2010a; 2010b). Foreign-accented speakers tend to enjoy conversation less, put more effort into communicating, and are more likely to avoid conversation than native speakers (Derwing & Rossiter, 2002). Furthermore, for L2 speakers, greater difficulty in communication is also associated with feeling less of a sense of belonging to the country that they live in, which is different from speakers with a dialectal accent whose difficulty in communication does not decrease their feeling of belonging (Gluszek & Dovidio, 2010a).

Although the attainment of high L2 pronunciation skills is arguably important, learning new phonetic features is not an easy task, especially for adult learners. People who arrived in a new country as adults not speaking its national language are less likely to attain native-like pronunciation in the new language than those who arrived as children (e.g. Flege, Munro & MacKay, 1995; Abrahamsson & Hyltenstam, 2009), which indicates that it is difficult for adults to learn new phonetic features from naturalistic exposure. This makes the importance of pronunciation training more essential for adult learners. For example, Derwing et al. (2014) demonstrated that pronunciation training at the workplace had a positive impact on adult immigrants' oral intelligibility.

An important task, therefore, is to establish effective training methods. The present thesis will focus on one of the important components of training—feedback. By feedback, we refer to external feedback, specifically a person's (often teacher's) reactions to a learner's performance. We do not refer to learners' internal sensory feedback unless specified otherwise. Feedback in training has been one of the major research interests in different disciplines in parallel, two of which are relevant in the present thesis: instructed L2 acquisition and motor skill learning.

Feedback in instructed L2 acquisition research

The definition of feedback, in its most general and simplistic way, is information about one's performance as a basis for improvement. In the research on L2 acquisition in formal instruction settings, however, feedback has been predominantly corrective: that is, learners receive feedback on the use of linguistic form only when it is erroneous. Another distinctive characteristic of feedback in L2 teaching is the potential presence of follow-up actions. Some types of feedback used in L2 teaching not only inform learners of the presence of an error, but also explicitly or implicitly expect learners to correct the error by themselves in response to

feedback (see Lyster & Ranta, 1997 for common types of CF). Note that whether learners are aware of the expectation for self-correction and whether learners in fact take such action depend on many factors including types of feedback (Panova & Lyster, 2002), but what emphasize here is that the potential presence (or absence) of follow-up actions has been a significant part of the feedback research in L2 instructional settings, which is not necessarily the case for feedback studies in other research areas, such as motor skill learning. This may be seen in the two major types of feedback in the L2 classroom: prompts and recasts (Lyster & Ranta, 1997; Brown, 2016). Prompts refer to responses that elicit students' self-correction (e.g. "I lead this book…" "Pardon? Say it again."), whereas recasts refer to teachers restating students' erroneous utterances with the corrected form (e.g. "I lead this book…" "I 'read' this book."). When prompting, teachers explicitly request (or in other cases, implicitly expect) that learners should correct themselves on their own. When recasting, teachers offer a corrected utterance and sometimes (but not always) expect students to pick up the implicit (i.e., unstated) request for learners to imitate the correct model. Both types of feedback are popular in practice: according to Brown's (2016) meta-analysis, prompts and recasts account for approximately 30% and 66% of classroom feedback, respectively.

Although both types of feedback have been proven to be beneficial for the acquisition of L2 syntax and morphology (Lyster & Saito, 2010; Lyster et al., 2013), it was not until recently that their effectiveness on phonological targets began to be investigated. Despite the popularity of these types of feedback, there are only a few recent studies that provide empirical evidence for the effectiveness of recasts and prompts on pronunciation (Saito & Lyster, 2012), and specifically, potential differential effects of recasts and prompts (Gooch et al., 2016). For example, a classroom intervention study by Saito and Lyster (2012) showed that Japanese

students who received recasts during a discussion class improved more in their pronunciation of English /r/, measured by acoustics (F3) and native listeners' judgement, compared to their counterparts in the control class, where the teacher did not provide any feedback on pronunciation. Gooch et al. (2016), who investigated the differential effects of recasts and prompts for Korean students learning English /r/, found that although both types of feedback showed positive effects compared to a no-feedback condition, prompts appeared to be more effective and generalizable than recasts: learners who received prompts improved in English /r/ in both controlled (i.e. reading a word list) and spontaneous (i.e. describing a picture) production, whereas learners who received recasts only showed improvement in the controlled production task. The authors argued that prompts should be more beneficial to learners' pronunciation skills involving untrained items, because "unlike recasts, which give students the option to mimic the instructor's pronunciation, prompts push students to use their own resources to try to produce a target-like utterance" (Gooch et al., 2016). One of the possible limitations of these studies (Saito & Lyster, 2012; Gooch et al., 2016) is that they did not control the amount of practice. Since both prompts and recasts may increase the amount of practice by eliciting additional productions, it is uncertain whether L2 phoneme learning was enhanced by the provision of recasts and prompts, or simply by the increased amount of practice due to the post-feedback productions. In the present papers (Study 1 and 2), measures were taken so that the amount of practice (including productions in response to prompts or recasts) for each participant was as close as possible to further demonstrate the effectiveness of feedback on L2 pronunciation.

The present thesis aims to expand such investigations on feedback in L2 pronunciation. To do so, we will refer to findings and theories from another line of research where feedback has been extensively investigated—motor skill learning. Since pronunciation is one of the few

components in language that directly involves motor skills, the motor learning literature provides relevant insights on feedback in L2 pronunciation. Several decades of motor learning research have already accumulated a significant amount of knowledge about how to provide feedback to optimize learning. To start our discussion, the following sections will introduce basic concepts, findings and theories from motor learning studies.

Augmented feedback and knowledge of results in motor learning

In studies of motor learning, what is called corrective feedback in L2 studies belongs to the category of "augmented" feedback because it is additional to intrinsic feedback, which comes naturally from the learner's own visual, auditory, or somatosensory perception. Augmented feedback arguably plays a crucial role for certain aspects of motor learning, especially when a learner cannot detect or take full advantage of intrinsic feedback (Magill, 2007, pp. 336-8).

There are two major types of augmented feedback: knowledge of results (KR) and knowledge of performance (KP) (Schmitt, 1982). KR refers to feedback about whether the outcome of the action reaches the environmental goal (e.g., "that was correct" or "that was /t/, not /tt/"), while KP refers to information about the movement pattern per se (e.g., "your tongue tip did not contact the alveolar ridge"). KR has long been one of the major topics in the motor learning literature (Salmoni et al., 1984) and is thought to be indispensable for learning to occur.

Corrective feedback such as prompts and recasts are essentially KR, not KP: they only inform learners if their production needs to be corrected, not including explicit information about the movement itself. Since many studies on feedback in motor learning have used KR, those previous studies should be relevant when we consider recasts and prompts. However, as discussed above, prompts and recasts potentially allow learners to take self-corrective actions,

which is different from most motor learning studies where only KR is provided. Therefore, while findings in motor skill learning research should inspire similar research in L2 speech acquisition, it is expected that there will be discrepancies in results. This is the gap that the present study aims to fill in.

What role does feedback play? (1) The guidance role

We have discussed that, in the research of instructed L2 acquisition, the role of feedback is primarily viewed as error correction. A similar take on feedback has prevailed in motor skill research as well. The classic, influential theory on the role of feedback in motor learning is the guidance hypothesis (Salmoni et al., 1984; Winstein & Schmidt, 1990; Schmidt, 1991), whose central ideas are: (1) the provision of feedback can tell learners what errors occurred and/or how to fix them, which is strong guidance for approaching the target movement, and hence learners perform well while receiving feedback; (2) on the other hand, the presence of guidance may also block other processing or development important to learning, such as how to use learners' internal feedback and establish their own criteria for errors (Salmoni et al., 1984). In other words, the guidance hypothesis posits that the primary role of feedback is guidance, but too frequent guidance may also be detrimental to learning.

Schmidt (1991) proposed two processing activities possibly blocked by feedback. First, feedback may block a learner's error-detection processing for the preceding trial. When learners are provided with feedback, their judgement of whether the preceding trial was successful may rely heavily on feedback, resulting in not processing their own intrinsic feedback, on which they must rely once feedback is withdrawn. Consequently, at retention, performance declines due to the lack of the learner's own error detection and self-correction ability (Salmoni et al., 1984; Schmidt, 1991). The second process that may blocked by feedback during training is the retrieval

process of the motor program for the next trial (Schmidt, 1991). If feedback provides information about the target movement, learners may not have to retrieve a motor program for the next action from memory. This notion has been discussed more intensively in the literature on the practice order effect (Magill & Hall, 1990, for a review). Studies have shown that learners learn and retain better when different targets are randomly practiced in training, compared to in conditions where the same target is practiced repeatedly in a block (Magill & Hall, 1990). Repeated practice for the same target is hypothesized to be detrimental to retention because learners may only retrieve a motor program for the target from memory on the first trial, then use it for the remaining trials with no or slight modification (Schmidt, 1991). The same problem may happen when feedback for the preceding trial practically provides the "answer" for the next trial.

What role does feedback play? (2) The motivator role

We have so far discussed how feedback changes learners' behaviours, but feedback also influences how learners perceive the target task and their own performance, which may increase or decrease their motivation. Feedback as a motivator has recently received much attention as one of the crucial factors in motor learning (for a review, Wulf et al., 2010). Recent studies found that learners who received feedback after accurate trials indeed performed better at retention than those who received feedback after poor trials (Chiviacowsky & Wulf, 2007; Badami et al., 2011; Saemi et al., 2012). In other words, positive (i.e., confirming desirable performance) feedback sometimes seems more effective than negative (i.e., indicating erroneous performance) feedback.

Why might positive feedback be more beneficial than negative feedback? Chiviacowsky and Wulf (2007) speculated that the primary factor was the motivational and affective effect of positive feedback. That is, positive feedback may have increased learners' perceived competence

on the target task, and hence benefitted their motivation and other cognitive/affective aspects. Badami et al. (2011) confirmed this speculation by finding that learners who received feedback after good trials indeed showed higher motivation scores than those who received feedback after poor trials, measured by the Intrinsic Motivation Inventory (McAuley et al. 1989). Likewise, Saemi et al. (2012) used a self-efficacy scale (Bandura, 2006) to show higher self-efficacy scores exhibited by learners who received positive feedback compared to those who received negative feedback. Moreover, recent studies revealed that there were certain brain areas that responded more strongly to positive feedback than negative feedback (Nieuwenhuis et al., 2005), which suggests that positive feedback may yield some direct influences on the processing of feedback itself.

Importantly, such findings seem to contradict the guidance hypothesis, which posits that the primary role of feedback is to "tell learners what errors were made and what to do next" (Salmoni et al., 1984). However, since positive feedback is given on successful trials, it does not provide learners with as much information about errors as negative feedback does. Yet a few studies found that learners who were given positive feedback made fewer errors than those who were given negative feedback both during training and at retention (Chiviacowsky & Wulf, 2007; Saemi et al., 2012).

How should we think of these two different views on the role of feedback in motor learning? As Schmidt (1991) has pointed out, feedback can simultaneously play different roles. That is, the role of feedback as error guidance and as a motivator—as much as the two view may seem contradictory—are not necessarily mutually exclusive. Negative feedback, for example, may yield a beneficial effect as guidance and a detrimental effect as a motivator at the same time, whereas positive feedback may provide less benefit as guidance but may better motivate learners.

Which role will manifest itself more prominently and will affect learning outcomes more significantly may all come down to what type of feedback, targets, and training settings are employed in that specific learning environment. This leads us to the broad research questions that motivate the present studies—what roles do feedback play in L2 pronunciation training? Is error correction the primary role? Or, alternatively, does the motivational aspect play a significant role? Do different types of feedback, such as prompts and recasts, play different roles?

Present studies

The present thesis consists of three studies, which all are concerned with how L2 learners acquire a foreign sound in an instructed setting, with the primary focus on the effect and role of feedback. These studies are inspired by the research on motor skill learning, where findings about feedback have been accumulated based on a similar view of feedback as in L2 research (i.e., feedback as a corrective/guidance role) and also on a new perspective (i.e., feedback as a motivator). The first study investigates the effect of feedback frequency, comparing two feedback conditions where participants receive feedback all the time or only half of the time. This is essentially a replication of a number of previous studies in motor skill learning (Nicholson & Schmidt, 1991; Sparrow & Summers, 1992; Vander Linden et al., 1993; Wulf et al., 1993; Wulf, Lee, & Schmidt, 1994; Weeks & Kordus 1998; Park, Shea, & Wright, 2000; Badets & Blandin, 2004; Sidaway et al., 2008; Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012; Austermann Hula et al., 2008; Maas et al., 2012; Adams et al. 2002; Katz et al., 2010; Van Stan et al., 2017), which repeatedly demonstrated that reduced frequency, in fact, may be more beneficial to learning than providing feedback constantly. The crucial difference between the current study and previous studies is that we investigated frequency with

respect to corrective feedback. As discussed above, feedback in L2 speech learning differs from feedback used in motor skill learning research because it is corrective and may require follow-up action from learners. We hypothesized that these differences in the characteristics of feedback may yield different results than the previous studies, which motivated us to conduct this study.

The second study aimed to investigate the more recent "feedback as a motivator" perspective in the context of L2 research. As previously mentioned, in motor learning research, positive (non-corrective) feedback has been shown to have a significant effect on learning, sometimes even more so than negative (corrective) feedback (Chiviacowsky & Wulf, 2007; Badami et al., 2011; Saemi et al., 2012). However, non-corrective feedback in L2 research has been studied with respect to its role in classroom discourse (Lyster, 1998; Waring, 2008), not in terms of its effectiveness on learners' linguistic attainment. To fill this gap, we defined a non-corrective counterpart of both prompts and recasts and examined whether non-corrective feedback may contribute to increase learners' accuracy in pronunciation just as corrective feedback does. Note that the data for this study partially overlapped with those for the first study, which may not be clear, as they are presented in separate manuscripts. The prompt, recast and control groups included the same participants, which means that only participants who received non-corrective feedback were newly recruited for the second study.

Both Study 1 and Study 2 served two objectives: practice-oriented and research-oriented. The practice-oriented objective was to compare feedback conditions, such as frequent vs. reduced in Study 1 and non-corrective vs. corrective in Study 2, in an attempt to draw implications on effective feedback methods. We used both prompts and recasts in both studies because we expected differential results for the two types of feedback, not because we anticipated that reduced feedback would always have an advantage over frequent feedback for

any type of feedback. Rather, we aimed to inform expectations for performance under varying conditions and types of feedback, to help teachers make educated decisions on the feedback they choose to provide. The research-oriented objective is related to the broad questions mentioned above—to discuss what roles feedback plays in L2 pronunciation learning. The effectiveness of feedback has been closely linked to error correction in the instructed L2 research, but both of the present studies challenge the notion from different perspectives: the first study asks whether frequent error correction may in fact be less effective than infrequent error correction, while the second study directly investigates whether non-corrective feedback may also be effective. These are intended to re-consider the notion of feedback as error correction, and further discuss what other significant roles feedback may serve in L2 speech learning.

The third study is an acoustic study derived from the previous two. In the first and second studies, we used the Japanese geminate-singleton consonant contrast as the target sound for participants to learn, yet did not discuss in detail what features learners were actually acquiring. In this third study, we utilized a portion of the same data collected for the previous two studies, but looked into the details of the acoustic nature of participants' production of Japanese geminates and singletons, to identify the major difficulty in mastering the contrast. Note that this study employed a cross-linguistic comparison approach, which means that the primary interest was the comparison between English and Japanese speakers; the different feedback conditions during training were not taken into consideration. However, the study still provides complementary information to the general discussion on the effects of feedback because, as discussed above, the effect and role of feedback may vary as a function of what learners have to learn in a certain training setting.

Finally, the general discussion provides an overview of the findings from the three studies, including differences from the existing literature, and draws conclusions concerning the effects and roles of feedback in L2 pronunciation training.

Preface to Paper 1

The first paper investigates feedback frequency—how frequently learners should receive feedback on their performance to achieve maximal performance. At first glance, frequent information on one's own performance may seem essential to learning and it is reasonable to provide feedback as frequently as possible. However, early studies in motor skill learning (e.g., Salmoni et al., 1984; Winstein & Schmidt, 1990) have proven otherwise, which is why feedback frequency has been one of the long-standing topics in motor learning. This classical issue has rarely been empirically investigated in the context of L2 learning. The question of how frequently learners should receive corrective feedback not only has practical value by yielding potential implications for teaching practice, but may also contribute to answering a more fundamental question: whether every error should be corrected, and furthermore, whether correcting errors is the primary purpose of feedback.

Paper 1: The effect of feedback frequency on L2 pronunciation training: the differential effects of prompts and recasts

1. Introduction

Over the decades, enormous effort has been dedicated to improving the outcomes of instructed second language (L2) acquisition in terms of various aspects of learners' performance, including linguistic accuracy. Corrective feedback (CF) has long been utilized in practice to help learners attain "correct" linguistic performance, and thus has been one of the long-studied topics in research (Lyster et al., 2013; Ellis, 2017 for overviews). Yet until recently, most research has concentrated on *whether* CF is effective at all (Russell & Spada, 2006; Li, 2010; Lyster & Saito, 2010); thus, there remains much to be learned concerning what features might improve its effectiveness. Once one decides to use CF, there are numerous variables to be specified that potentially have an impact on its effectiveness: for example, in what form, when, by whom, or in response to what kind of errors should CF be provided? Among those questions is its frequency—in other words, how frequently should learners receive CF?

In the following sections, we will first review how CF frequency has been discussed in L2 learning studies, then turn to the speech motor learning literature where feedback frequency has been extensively investigated, and finally discuss the current study's research question—the effect of frequent and reduced CF on L2 pronunciation learning, especially the possibility that frequency may have differential effects on different types of CF.

1.1. CF frequency in L2 learning studies

There are different, seemingly contradictory, beliefs and preferences regarding the ideal frequency of CF that have been revealed in the L2 learning literature. Teachers are more often than not hesitant to provide CF frequently, partly because one of their goals is to promote communicative interactions in the classroom, and they believe that CF would interfere with such communication (Long, 2006). Moreover, teachers are also concerned that CF may serve as a de-motivator, potentially embarrassing learners or lowering their confidence by highlighting errors (Lasagabaster & Sierra 2005: Vásquez & Harvey, 2010; Mori, 2011; Méndez & Cruz, 2012). On the other hand, in many surveys, learners show a clear preference for receiving CF more often than teachers use it; some even wish to be corrected "all the time" (Han & Jung, 2007; Jean & Simard, 2011; Lee, 2013; Zhang & Rahimi, 2014). It seems that the more learners find linguistic accuracy important, the more likely they are to prefer frequent CF (Lyster et al., 2013; Schulz, 2001; Loewen et al., 2009). Thus, beliefs and preferences about CF frequency seem to revolve around two opposing concerns: if given too frequently, it might do more harm than good to communication or learners' attitudes; if given too infrequently, it might neglect learners' need for error correction and potentially hinder accurate usage of language.

Although such beliefs and concerns exist, there have been few studies that have attempted to empirically manipulate CF frequency to investigate its effect on L2 learning. Previous studies on CF frequency were usually ones in which researchers observed and counted the number of instances of CF that occurred in classrooms or other learning settings, with the purpose of illustrating how CF was used in practice (Lyster & Ranta, 1997; Harvranek, 2002; Lochtman, 2002; Panova & Lyster, 2002; Sheen, 2004; Suzuki, 2004; Ahangari & Amirzadeh, 2011; Lee, 2013; Safari, 2013; Fu & Nassaji, 2016). These studies have documented that CF

frequency, as in the number of occurrences of CF, varies greatly depending on the type of CF, educational context, teacher's experience and learners' proficiency. With respect to type of CF, there are two major types of CF that have significantly different natures; given that the purpose of CF is to help learners attain the correct linguistic form, one type of CF is used to *elicit* the accurate target and the other is used to *provide* it. The former includes prompts, which refer to implicit or explicit requests such as "Pardon?" or "Try again" to elicit learner's self-correction, while the latter includes recasts, where teachers reform learners' erroneous output with a corrected language use (Lyster & Ranta, 1997; Sheen, 2011). Recasts are known to have higher occurrence than other types of CF. Brown's (2016) meta-analysis revealed that recasts accounted for 57% of all CFs that had been observed in primary studies in the past two decades, while prompts were seen less frequently, accounting for 29.5%. The reasons behind this difference in usage seem related to both teachers' beliefs and practical constraints: for example, Yoshida's (2008) survey revealed that teachers may believe that recasts pose less of a disruption in communication and also less of a threat to learners' motivation, while teachers may also choose recasts out of concern for learners' low proficiency and time management, since they feel that prompts are more difficult for learners and that learners may need more time to respond to prompts than recasts.

While the above discussion is concerning the occurrence of CF (i.e., the number of instances of CF that occurred), the present paper focuses more on the proportion of CF actually given and all errors that could potentially be followed by CF (hereafter by "CF frequency", we refer this proportion unless noted otherwise), which also varies significantly due to many factors. One of such factors is the setting of the study: laboratory and classroom. In laboratory studies, researchers usually make an effort to provide CF to every error regarding the target item that

occurred during the experimental sessions (e.g., Saito, 2013; Gooch et al., 2016; Yang & Lyster, 2010). In other words, CF frequency should ideally be close to 100%. Although it may not be achievable especially when the researchers themselves are not in charge of instruction, CF frequency in laboratory studies is expected to be high. For example, Yang and Lyster (2010) reported that 90.3% of all errors were followed by CF in one experimental group where the first author provided CF, while it was 74% in a group where another teacher was in charge. On the other hand, when studies are based on classroom observations, CF frequency tend not to be as high as in laboratory studies: for example, Lyster and Ranta's (1997) classroom observational data reported that 62% of total errors were followed by CF; in a smaller dataset analyzed by Kennedy (2010), only about 20% of total errors were followed by CF. While there has been discussions on the difference in outcomes between the laboratory and classroom research in L2 acquisition (ex., Gass, Mackey & Ross-Feldman, 2005), these discrepancies in CF frequency has not been discussed in relation to the effectiveness of CF. For example, if a certain type of CF is found to be effective in a laboratory setting where CF is given upon the occurrence of every error, would it still be effective in other situations where CF can only be given in response to some portion of errors?

Although these questions are yet to be answered in relation to CF in L2 acquisition, there is another line of research where the same question—how often learners should receive feedback—has been extensively investigated in empirical fashion: motor skill learning studies. The field has long accumulated findings on the feedback conditions that yield optimal learning outcomes in the acquisition of physical skills (see Schmidt & Lee, 2011 and Mcgill, 2007, for general overviews; Maas et al., 2008 and Bislick et al., 2012, for overviews in relation to applications in the speech domain), which could be of relevance to the L2 learning literature. The

current study aims to investigate if the findings on feedback frequency in motor skill learning would apply to CF, specifically to prompts and recasts, in one of the few aspects of language learning directly akin to motor skill training: pronunciation.

1.2. Feedback frequency in non-speech motor skill learning

In the early literature on motor skill learning, which focused primarily on limb movements, the most influential idea was that learning would not happen at all through practice without feedback and that feedback thus should be provided as frequently as possible (e.g., Bilodeau & Bilodeau, 1958). However, this idea was countered by subsequent studies that found a counter-intuitive result: frequent feedback (e.g., providing feedback after every trial) was, in fact, less effective than reduced feedback (e.g., providing feedback after every few trials), if the measurements were taken in a delayed post-test (e.g., Salmoni et al., 1984; Winstein & Schmidt, 1990). That is, although frequent feedback enhances learners' performance greatly during training (and thus at an immediate post-test), it is often followed by a significant decline (Winstein & Schmidt, 1990). This pattern, better training performance and poorer retention due to frequent feedback, has been proven to be robust and replicated in many studies (e.g., Nicholson & Schmidt, 1991; Sparrow & Summers, 1992; Vander Linden et al., 1993; Wulf et al., 1993; Wulf, Lee, & Schmidt, 1994; Weeks & Kordus 1998; Park, Shea, & Wright, 2000; Badets & Blandin, 2004; Sidaway et al., 2008). Several accounts have been offered to explain this phenomenon, among which are: (a) frequent feedback makes learners rely too much on feedback, which prevents them from noticing errors by themselves (the guidance hypothesis: Salmoni et al., 1984; Winstein & Schmidt, 1990); and (b) frequent feedback makes learners'

behavior unstable due to too frequent corrections (Schmidt, 1991); (we will return to this point below).

The majority of the studies cited above empirically compared two frequencies, “frequent” and “reduced”, and while the “reduced” frequency varied across studies (i.e., in most cases, feedback was given once every three trials [33%], every two trials [50%], or two times for every three trials [66%]), the “frequent” condition was always 100% (i.e., after every trial), which consistently led learners to poorer performance than the reduced condition at a delayed post-test. Therefore, when it comes to limb motor skill learning, a consensus has emerged that the optimal feedback frequency to maximize learning is “not 100%” (Magill, 2007, pp. 357).

Although the consensus underlines the disadvantage of frequent feedback to retention, it should be noted that there have been discrepancies in the literature. Some researchers claim that the effect of feedback frequency interacts with task complexity (Wulf & Shea, 2002, for a review). For example, Wulf et al. (1998) found that frequent feedback was more beneficial to retention than reduced feedback in a ski simulator task where participants engaged the whole body to learn a slalom movement, which was more complex and difficult than the motor tasks used in previous studies, such as lever rotation or line tracing with one hand (e.g., Winstein & Schmidt, 1990). Furthermore, other studies also found that 100% feedback was more effective than reduced feedback for children (Chiviacowsky et al., 2008; Sullivan et al., 2008). Proficiency also matters, with beginners benefiting more from frequent feedback and advanced learners from reduced feedback (Guadagnoli et al. 1996). It seems that the more complex and/or difficult the task is, due to either the task complexity itself, cognitive skills or the amount of experience of the learners, the more likely frequent feedback is to be beneficial. Wulf and Shea (2002) argued that complex tasks inherently demand high levels of attention, memory or control from learners, thus

reducing feedback frequency and imposing more demands on learners may not be beneficial to learning. One of the problems of this account, however, is that there is no clear definition or measurement of task complexity (Wulf & Shea 2002). Some studies have also found the advantage of reduced feedback in a seemingly complex task (for example, Kim et al. (2012), which will be reviewed in the next section, used an imitation task of complex sentences), which implies that there should be factors other than complexity that may explain the discrepant findings.

1.3. Feedback frequency in speech motor learning

It is only in recent years that researchers have started to investigate the application of the findings in limb motor learning to speech learning (Maas et al, 2008; Bislick et al., 2012). A handful of studies has investigated the effect of feedback frequency in speech learning tasks with normal subjects (Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012) and for those with communication disorders (Austermann Hula et al., 2008; Maas et al., 2012; Adams et al. 2002; Katz et al., 2010; Van Stan et al., 2017).

In normal subjects, a few previous studies (Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012) have supported the notion that reduced feedback yields better outcomes for speech learning, much like in limb motor learning (e.g., Nicholson & Schmidt, 1991; Sparrow & Summers, 1992; Vander Linden et al., 1993; Weeks & Kordus 1998; Park, et al., 2000; Badets & Blandin, 2004; Sidaway et al., 2008). For example, Adams & Page (2000) found that 20% feedback was more effective than 100% feedback in learning slow speech rate. In Steinhauer and Greyhack's (2000) task where participants learned to reach a certain vowel nasality level, 100% feedback was less effective than 50% feedback, and even less effective than

a no-feedback control condition, which suggests that feedback sometimes can be not only ineffective, but also detrimental to learning if it is too frequent. One of the few studies that directly used an L2 to investigate feedback frequency was that of Kim et al. (2012), where the researchers asked English-speaking participants to repeat short Korean sentences that were totally novel to them and provided them with positive or negative verbal feedback. Results showed that reduced feedback and a large amount of practice (i.e., 20% feedback with 100 trials) yielded the best outcome for learners, while frequent feedback with a large amount of practice (i.e., 100% feedback with 100 trials) consistently showed the poorest scores compared to the other conditions (including 20% and 100% feedback with 25 trials).

In studies on speech therapies for individuals with speech disorders, the advantage of reduced feedback has also been partly supported, although the results are more mixed (Austermann Hula et al., 2008; Maas et al., 2012; Katz et al., 2010; Van Stan et al., 2017). For example, Austermann Hula et al. (2008) examined the effect of feedback frequency on speech therapy for adults with apraxia of speech (AOS). Results showed that two of the four participants demonstrated enhanced performance for a reduced (60%) feedback condition compared to a frequent (100%) feedback condition, while no difference was found for the remaining two participants. Maas et al. (2012), who explored the effect of feedback frequency on children with AOS, found that two of the four children demonstrated an advantage of reduced (60%) feedback over frequent feedback, while one child showed the reverse, and the remaining one did not show any improvement in either condition. The authors concluded that the general advantage of reduced feedback could be partly supported in this disordered population, although various variables could have an impact on the effectiveness.

1.4. Factors influencing the effect of feedback frequency

There have been two major accounts to explain the apparent disadvantage of frequent feedback at retention. The first one is learners' ability to detect errors in their own performance (the guidance hypothesis: Salmoni et al., 1984; Winstein & Schmidt, 1990; Schmidt, 1991; Anderson et al., 2005). When learners are provided with feedback from the external source, their judgement concerning whether the preceding trial was successful may rely heavily on it, resulting in not using or refining their own criteria (i.e., *intrinsic* feedback), on which they must rely once feedback is withdrawn. Consequently, it is hypothesized that if learners are provided with feedback too frequently, even though they would apparently perform well while feedback is still present, performance declines significantly at retention due to the lack of the learner's own error detection and self-correction ability (Salmoni et al., 1984; Schmidt, 1991). It should be noted, however, that few previous studies have actually investigated learners' error detection ability in relation to feedback frequency. A few exceptions are Guadagnoli and Kohl (2001) and Silva et al. (2017), where learners who were instructed to estimate their own errors before receiving feedback performed better in a frequent feedback condition compared to a reduced feedback condition, but the third condition, which was frequent feedback without subjective error estimation, resulted in the poorest learning outcome. These studies suggest that frequent feedback may hinder learners' error detection ability, but an additional instruction or procedure that is designed to encourage error detection may compensate for or overcome that disadvantage.

Another account of the disadvantage of frequent feedback relates to motor variability (Lai & Shea, 1998; Schmidt & Bjork, 1992; Maas et al., 2012). Some early studies found that learners' performances were more stable (less variable) during training when they received reduced feedback compared to frequent feedback (Wulf & Schmidt, 1989; Wulf et al., 1993). It

was thus assumed that frequent feedback may be detrimental to retention because learners may attempt to correct deviations from the target every time they receive feedback, even when the deviations are relatively small and insignificant, which prevents learners from establishing a stable movement to be recalled at retention (Lai & Shea, 1998). This account may be relevant when feedback is given on a continuous scale (e.g., the performance time in milliseconds when the goal is to complete a task within a specified time), but we speculate that it may be less relevant when feedback is given in a categorical fashion such as CF in the context of L2 learning (which is only given when a teacher feels that the response is "incorrect"), since "small and insignificant" deviations that learners need not focus on are not expressed in this type of feedback. Rather, in a broader context of motor learning, feedback is generally thought to be able to reduce variability in learners' performance by correcting errors in the brain's motor planning before executing the target movement (e.g., Dhawale et al., 2017). In sum, although the previous studies on feedback frequency suggest that frequent feedback may increase variability in performance, which may be detrimental to retention, we assume that this may not be the case for CF in L2 learning; it is thus worth investigating how feedback frequency and variability in performance are related in the case of CF.

1.5. The present study

The present study aims to investigate the effect of CF frequency on L2 speech learning, specifically to compare the effectiveness of frequent and reduced feedback on laboratory pronunciation training. In doing so, we rely on two types of CF—prompts and recasts. Few studies have investigated the effect of CF frequency in the context of L2 learning, and although there has been extensive research on feedback frequency in motor learning, the type of feedback

such as prompts and recasts has not been explored in previous studies. While previous studies used feedback to inform learners of errors (or the lack thereof), prompts and recasts provide more functions: prompts request learners' post-feedback self-correction, while recasts provide learners with auditory models, which can potentially be followed by learners' imitation. We assume that these post-feedback corrective actions and provision of auditory models may change how learners learn the target, and consequently the effect of frequency may be different for the two types of CF.

Prompts indicate that there are errors to be corrected, and that learners must rely on their own judgement and resources to correct them. Therefore, prompts do not appear to be the type of feedback that prevents learners from detecting errors on their own, which was one of the possible reasons why frequent feedback could be detrimental (Salmoni et al., 1984; Winstein & Schmidt, 1990). On the contrary, prompts most likely encourage the error detection process by encouraging learners' self-correction. Another explanation that has been offered as to why frequent feedback may have a disadvantage at retention relates to variability (stability). As discussed above, frequent prompts may not increase learners' variability in production as previous studies have suggested (Wulf & Schmidt, 1989; Wulf et al., 1993; Lai & Shea, 1998). If this is true, giving frequent prompts will not yield a detrimental effect on retention as observed in previous studies; on the contrary, frequent prompts are predicted to be more beneficial to both training and retention than reduced prompts.

Recasts, on the other hand, provide correct auditory models that learners can directly imitate, which can be a strong guidance to approach the target. Therefore, it is expected that providing frequent recasts will improve learners' performance significantly during training. However, as we have seen, it has been argued that if feedback is provided too frequently,

learners will rely on feedback and withhold their own judgement to detect errors, which may lead to poorer retention. This might be the case for recasts, especially because providing a correct auditory model may distract learner's attention from their own erroneous production which occurred right before it, effectively blocking the error detection process. On the other hand, some researchers have argued that, with opportunities to listen to the correct auditory model right after learners' own erroneous production, learners who receive recasts might be able to compare the two and develop a better ability to detect errors ("cognitive comparison": Long, 1996). Given these different possibilities, we expect that there may be different results for recasts than prompts, especially if the "blocking error detection" account is the case; that is, learners who receive frequent recasts may excel at training but not retain well at a delayed post-test, while learners who receive recasts at reduced frequency may be better at retention.

In addition, we also measured learners' motivation in relation to the type and frequency of CF, given that the impact of CF on learners' attitudes has been one of the major factors in teachers' beliefs about CF frequency. As discussed above, teachers tend to believe that CF in general is discouraging to learners (Lasagabaster & Sierra 2005: Vásquez & Harvey, 2010; Mori, 2011; Méndez & Cruz, 2012), which is one of the reasons why they refrain from frequent CF, hoping that reducing frequency of CF may mitigate the negative effect. In addition, when choosing the type of feedback, teachers also tend to feel that recasts are less damaging to learners' motivation than prompts (Yoshida, 2008). To determine if these beliefs are warranted, we investigate learners' motivation during training and examine whether frequent CF is discouraging compared to reduced CF with respect to both prompts and recasts.

2. Methods

2.1. Participants

Participants were eighty young adult English speakers living in Montreal, who were recruited primarily from the McGill University and Concordia University student and alumni communities. Participants were compensated for their time upon completion of the experimental sessions. The requirements for participation were: (1) aged between 18 and 40 years old; (2) having no knowledge or learning experience of the target word or language (i.e., Japanese); (3) identifying English as their most dominant language if they spoke more than one language. Participants completed a questionnaire about their age (M = 21.6, SD = 3.5, max = 36), gender, language(s) that they spoke besides English (if applicable), and self-reported usage of each language in daily life. Many participants spoke one or two languages in addition to English, which they learned in school or from parents as a heritage language, although English usage was predominant in daily life (average self-reported usage > 90%). They were randomly divided into five groups according to the type and frequency of feedback that they were going to receive in the experimental sessions, namely prompt 100% (hereafter abbreviated as P100), prompt 50% (P50), recast 100% (R100), recast 50% (R50), and no-feedback (control). The number of participants[1], average age, gender, average number of languages they spoke in addition to English for each group are shown in Table 1.

[1] One participant in the control group was excluded due to a failure in the experimental procedures (there were originally eighty-one participants).

	Prompt 100% (P100)	Prompt 50% (P50)	Recast 100% (R100)	Recast 50% (R50)	Control
N	16	16	16	17	15
Age	21.1	21.2	20.1	23.9	22
Gender (female : male)	9 : 7	13 : 3	15 : 1	12 : 4[2]	12 : 3
Average N of L2 per person	1.4	1.8	1.8	1.8	1.8

Table 1. Demographic information of participants. N = number of participants; Age = average age at the time of experiment; Average N of L2 per person = the average number of languages that participants spoke besides English.

	singleton-singleton (SS)	geminate-singleton (SG)	singleton-geminate (GS)	geminate-geminate (GG)
trained item	atata ototo	attata ottoto	atatta ototto	attatta ottotto
untrained item	akaka okoko	akkaka okkoko	akakka okokko	akkakka okkokko
	akota otako	akotta okatto	attoka ottako	akkotta ottakko

Table 2. Target Japanese non-words. Eight trained items featured the /tt/-/t/ contrast. Untrained items featured either only the /kk-k/ contrast or both contrasts. Untrained items appeared in the pre- and post-test but not in the training phase.

[2] One participant identified neither.

2.2. Target words and materials

The experiment was designed to train participants to produce the Japanese geminate-singleton contrast, specifically the plosives /tt/-/t/. Japanese geminates and singletons are linguistically distinctive and discriminated primarily with the duration of stop closure (Hirata, 1990), although there are other secondary acoustic cues, such as the duration, fundamental frequency (F0) and intensity of the preceding and following syllables (Kawahara, 2015; Idemaru & Guion, 2008).

The target words were eight non-words, as displayed in Table 2. Non-words, rather than real words, were selected because unfamiliar linguistic elements in real words (e.g., pitch accent will differ even between geminate/singleton minimal pairs) may distract participants from the target element, considering that participants had no prior knowledge of Japanese. The target non-words consisted of three syllables, VCVCV, where Vs were either all /a/s or all /o/s and Cs were either geminate /tt/ or its singleton counterpart /t/. The geminate /tt/ appeared either at C1, C2, both C1 and C2, or not at all, to create four different word types; geminate-singleton (GS), singleton-geminate (SG), geminate-geminate (GG) and singleton-singleton (SS). These eight /tt/-/t/ words were used throughout the exposure phase, pre-test, training and post-test.

Another two sets of eight non-words were only used in the pre- and post-test (also see Table 2) to see if the effect of training could be generalized to an untrained geminate-singleton contrast. For this purpose, eight non-words that featured the Japanese /kk/-/k/ contrast instead of the /tt-t/ contrast, and eight non-words that contained both the /tt/-/t/ and /kk/-/k/ contrasts were included.

For recording materials for the experiment, we wanted to have a variety of speakers for the exposure phase (see Procedures below), because previous studies found that using multiple

speakers, rather than a single speaker, facilitates L2 phonetic learning because listening to different speakers helps learners rule out irrelevant cues and focus on the important ones (High Variability Phonetic Training, HVPT: Logan, Lively, & Pisoni, 1991; Barriuso & Hayes-Harb, 2018, for a review). Four native Japanese speakers, two younger speakers in their thirties (one female and one male) and two older speakers in their sixties (one female and one male), recorded these 24 words using a digital field recorder (22050 Hz, 16-bits). They were asked to produce the words at natural speed and as clearly as possible. Each word was recorded five times, from which only one token per speaker was used in the experiment: the third token was assumed to be the most stable and therefore used, unless there were some disfluencies or unclarities in the token (in that case, the fourth token was used).

2.3. Procedures

The experiment took place on three separate days over the course of approximately one week. Day 1 was reserved for preparation, including an exposure phase. The pre-test and training were performed on Day 2. Day 3 included the post-test. Days 1 and 2 took place on two consecutive days, while Day 3 occurred one week after Day 2. All participants underwent the same procedures in the exposure phase and the pre- and post-test: the only difference in procedures was participants' feedback condition in the training phase. All experimental procedures (except questionnaires) were implemented using OpenSesame software (an open-source experiment builder: Mathôt et al., 2012).

Day 1 (1): The exposure phase

On Day 1, participants filled out a demographic questionnaire (see 2.1) and proceeded to the exposure phase. The objective of this phase was (1) to give participants models of the target consonants because they had no prior knowledge of Japanese and (2) to measure participants'

initial perceptual sensitivity to geminates. Participants listened to each of the eight target words in separate trials and chose the word that they heard from four options with the same vowel (e.g., if the stimulus was /attata/, the four options would be /attata, atatta, attatta, atata/) (see Figure 1, left panel). If the answer was correct, a visual mark that indicated "correct" and the correct answer appeared on the screen (Figure 1, upper right panel), while the mark indicating “wrong” appeared when participants selected a wrong answer (Figure 1, lower right panel). The visual mark and correct answer were provided after every trial. The exposure phase lasted approximately 10 to 15 minutes, with 160 trials (8 words * 4 speakers * 5 repetitions).

Day 1 (2): Motivational questionnaire (baseline)

Participants also filled out the Intrinsic Motivation Inventory (IMI: McAuley et al. 1989; Self-Determination Theory Research Group, n.d.) to indicate their motivation toward learning pronunciation in general. There are several different subscales on the IMI, three of which were used in this study: Interest/Enjoyment, Perceived Competence, Tension/Pressure. Four items for each subscale, resulting in 12 items in total, had been selected from a pool of 16 items (Self-Determination Theory Research Group, n.d.), as shown in Table 3. Participants indicated how true the items were for them on a 1-to-7 Likert scale (1 = not at all true and 7 = very true). Note that higher scores were thought to represent higher motivation for Interest/Enjoyment and Perceived Competence, but lower scores represented higher motivation for Tension/Pressure.

Day 2 (1): Pre-test

Participants sat in a sound-treated room with a computer screen and a microphone. Throughout the pre-test, training and post-test, participants’ productions were recorded on the computer via a unidirectional microphone (22050 Hz, 16-bits). The microphone was placed on the desk, approximately 10 cm from the mouth.

Figure 1. The examples of the options and feedback for the exposure phase.

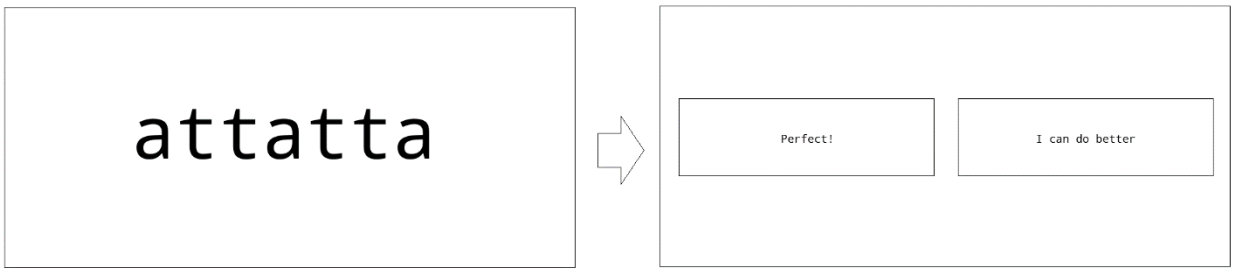

Figure 2. The examples of the pre-test. The left panel was presented to participants to read aloud the word. The right panel was for participants to judge their production was successful ("perfect!") or not ("I can do better").

Interest/Enjoyment
I enjoy pronunciation exercises very much.
I think pronunciation exercises are boring activities. (R)
I would describe pronunciation exercises as very interesting.
Pronunciation exercises do not hold my attention at all. (R)
Perceived Competence
I think I am pretty good at pronunciation in foreign languages.
After practicing pronunciation for awhile, I feel pretty competent.
I am satisfied with my performance at pronunciation in foreign languages.
Pronunciation exercises are an activity that I can't do very well. (R)
Tension/Pressure
I feel very tense while practicing pronunciation.
I am very relaxed in practicing pronunciation. (R)
I do not feel nervous at all while practicing pronunciation in foreign languages. (R)
I feel pressured while practicing pronunciation in foreign languages.

Table 3. Intrinsic Motivation Inventory (IMI), modified for the current study. The R after an item indicates a reverse item.

At the beginning of Day 2, participants performed a pre-test. Participants were instructed to read aloud the target words presented on the screen as quickly and accurately as possible (Figure 2, left panel). The target words were written in Roman alphabet since participants did not have a knowledge of Japanese orthographies. In addition, participants were asked to judge if their production was successful after every trial. They clicked one of two buttons on the screen: if they thought their production was successful, they clicked the "Perfect!" button, whereas they chose the "I can do better" button if they thought there were some errors in their production (Figure 2, right panel). The target words were the eight target words, to which participants had already been exposed in the exposure phase, and the sixteen untrained words, which they had not heard during the exposure phase. All 24 words were read aloud four times each in random order, resulting in 96 productions per participant.

The experimenter (the author), who sat outside the sound-treated room, listened to each production in real-time via headphones and judged if the production was either a "pass" or "fail". A production was judged as a "fail" if: (1) The target geminate /tt/ was heard as its single consonant counterpart /t/, or vice versa; (2) the target consonant was heard as another Japanese consonant (e.g. /tt/ heard as /tɕ/); otherwise the production was judged as a "pass". Participants, however, were not informed of the judgement: they had to rely solely on their own judgement.

Day 2 (2): Training phase

The training phase consisted of eight repetitions of a block. At the beginning of each block, participants were allowed to listen to native speakers' recordings of the target words twice each via headphones (the same recordings used in the exposure phase, but only the young female speaker's recordings were used) to refresh their memory (Figure 3). The recordings were played sequentially and automatically, with no control by the participants. Subsequently, they read

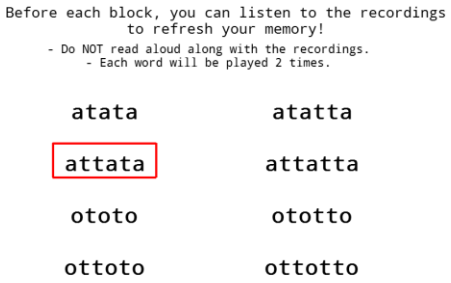

Figure 3. The screen from the part where participants were allowed to listen to model recordings.

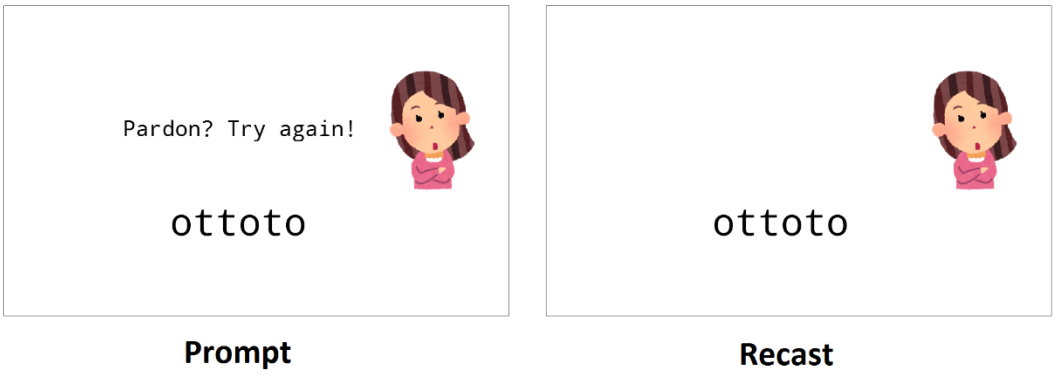

Figure 4. Examples of the screen presented to participants when they were provided with feedback. The left panel was presented to the Prompt groups (both 100% and 50%) when their production were judged as “fail”. The right panel was presented to the Recast groups (both 100% and 50%) along with a recording of native speaker producing the target word.

aloud the target words presented on the computer screen in random order. Participants' productions were listened to and judged by the experimenter in the same way as in the pre-test (except self-judgement was not required), but this time participants received feedback based on the judgement, whereas the control group continued to receive no feedback.

If a production was judged as a "pass", no feedback was provided: the participant simply proceeded to the next word. On the other hand, if a production was judged as a "fail", feedback was provided as follows.

For prompt groups (both P100 and P50, see below for frequency control), when a production was judged as a "fail", a character illustration with a frowning face (which reminds participants that they made errors) and a text ("Pardon? Try again!") appeared on the screen, in addition to the target word (Figure 4, left panel). Participants were required to produce the target again on their own: this second attempt is hereafter called "post-feedback trial", as opposed to the trials prior to feedback, which is called "normal trials". No feedback was provided for the post-feedback trials; even if the post-feedback trial was still not successful, they simply proceeded to the next word.

For recast groups, R100 and R50, when judged as a "fail", a recording of the target word was presented via headphones, along with the same illustration and the target word on the screen (Figure 4, right panel). Immediately after listening to the recording, the participant was required to produce the target again. As with the prompt groups, participants did not receive further feedback whether or not the post-feedback production was successful.

Feedback frequency was controlled automatically by the OpenSesame program. When a production was judged as a "fail", the program automatically decided if feedback should be given depending on the assigned frequency. For the 50% groups, the program provided a recast

or prompt after every two "fails". In other words, the program simply overlooked one in two "fails" and gave no feedback for the 50% groups, whereas it provided a recast or prompt after every "fail" for the 100% feedback groups.

Participants were explicitly informed of the feedback conditions to which they were assigned at the beginning of the training phase. That is, participants in the 50% groups were aware that they would only receive feedback on 50% of their errors, and that half of their errors were going to be ignored.

One block consisted of 32 productions (8 target words * 4 repetitions), resulting in 256 productions in a total of 8 blocks. The entire training phase lasted 20 to 30 minutes. Note that these 256 productions included the productions after feedback. In other words, the experiment was designed to have every participant produce the same number of productions regardless of how many times they received feedback: the experimental program controlled the total number of productions by adjusting the number of times the target words were presented on the screen; the more times a participant received prompts or recasts and consequently produced the target more often, the program presented the word less often, to keep the total number of productions consistent across all participants[3]. Without any experimental control, since prompts and recasts require participants to pronounce the target word again, participants who receive feedback 100% of the time consequently have more opportunities to produce the target, compared to participants in the 50% groups, or in the control group. If this difference in the number of productions is not

[3] However, each participant had a possibility that they may pronounce a target word one more time than expected, because participants were required to perform post-feedback trials when their productions were judged "fail". That is, if a participant failed at the last trial of the target word, he/she had to pronounce it one extra time in the post-feedback trial (naturally, this "extra" trial did not occur when he/she succeeded at the last trial of the target word, so it was largely by chance). As a result, each participant on average pronounced one extra word (raged from zero to four extra words), thus the average number of productions in fact was 257, instead of 256 (see Table 8).

controlled, it is difficult to exclude the possibility that any differences shown between 100% and 50% feedback might be attributed to the increase in the amount of practice, not to the frequency of feedback itself.

Similarly, the amount of auditory input was also controlled across participants. Since recast provides an auditory model upon feedback, learners who receive recasts have more auditory models than those who do not. Consequently, the effect of recasts on learning can be confounded with the increase in the amount of auditory input. To avoid this, we controlled the number of times that participants could listen to the model recordings: as described above, while the prompt groups and the control group were allowed to listen to the recordings of the target words twice each at the beginning of each block, the number of times was reduced for the recast groups, depending on how many times they received feedback in the previous block. For example, if a participant received one recast on a target word in the previous block, they could only listen to the recording of the word once before the next block. The objective of these controls was to focus on the effect of feedback itself, rather than potential confounds associated with the total amount of input and output.

Day 2 (3): Motivational questionnaire (task-specific)

Immediately after the training phase, participants filled out the Intrinsic Motivation Inventory as they did on Day 1. This time, however, the questions were slightly changed to measure participants' motivation toward this particular training session. For example, the question "I enjoy pronunciation exercises very much" was changed to "I enjoyed this session very much." The scores were calculated in the same way as at baseline, assuming higher enjoyment and confidence, and lower perceived pressure, represented higher motivation.

Day3: Post-test

The post-test, scheduled one week after training, was the same read-aloud task as the pre-test. The test used all 24 words (eight /tt/-/t/ target words, eight /kk/-/k/ words and eight /tt/-/t/ /kk/-/k/ mixed words) and consisted of 96 trials (24 words x 4 repetitions). Participants' productions were judged by the participants and the experimenter using the pass-or-fail judgement, identical to the pre-test. In addition, participants also performed a forced-choice perceptual task as they had during the exposure phase, as it was possible that there may have been a spillover effect on perceptual skills even though participants only performed production training during the experiment. Unlike the exposure phase, however, participants were not informed of whether their answer was correct. The perceptual task contained 160 trials (8 words [/tt/-/t/ contrast only] * 4 speakers * 5 repetitions).

2.4. Analysis

Reliability of the native listener judgement

We first analysed whether and to what extent the pass-or-fail judgements given by a single native Japanese listener (i.e., the author) were consistent and justifiable by objective measurements (i.e., acoustic data), since the binary judgements will be the primary indicator of the accuracy of participants' pronunciation in the following analyses. To examine this, we built Generalized Linear Mixed Effects Models (GLMMs) to predict listener's judgements using acoustic measurements to see how well the predictions from the model match the actual pass-or-fail judgements by the listener. All statistical analyses in the following analysis were performed in R (R Core Team, 2020) using the lme4 package (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017) for mixed effects models.

The primary acoustic correlate of geminate consonants is closure duration, that is, from the offset of the preceding vowel to the burst of the subsequent plosive (Kawahara, 2015): closure durations for Japanese geminates are usually two to three times longer than those for singletons (Kawahara, 2015). Although raw values of closure duration can be a direct indicator, they are known to be influenced by speech rate; for example, geminates in fast speech can be shorter than singletons in slow speech (Hirata & Whiton, 2005). Previous studies (Amano & Hirata, 2010; Hirata & Whiton, 2005; Idemaru & Guion, 2010) found that geminates and singletons were best distinguished by relational correlates, such as the durational ratio of closure to word, rather than by raw values of closure duration. Other secondary cues, such as the duration of the preceding and following vowels, fundamental frequency (F0) and intensity changes from the preceding to the following vowel, are also known to be acoustic correlates of Japanese geminates (Idemaru & Guion, 2008; Kawahara, 2015).

Given these findings, we used three categories of acoustic measurements: duration, F0 and intensity. First, the durations of two stop closures (abbreviated as C1 and C2) and of three syllables excluding the stop closures (i.e., voice onset time plus vowel, abbreviated as S1, S2 and S3) were measured using Praat (Version 5.4.19, Boersma & Weenink, 2020), as Figure 5 indicates. When words produced by participants consisted of any other sounds or noises than the above-mentioned five parts (e.g., omitting or inserting syllables, adding a release burst during closure, coughing or laughing), measurements were not taken and those trials were removed from all acoustic analyses. The durations were then all divided by the duration of the entire word. In addition, considering that there were two stop consonants in the target words, we assumed that the relative closure duration of C1 and C2 may also have an influence on the listener's judgement. Thus, the durational ratio was calculated by dividing C1 by C2, hereafter referred to

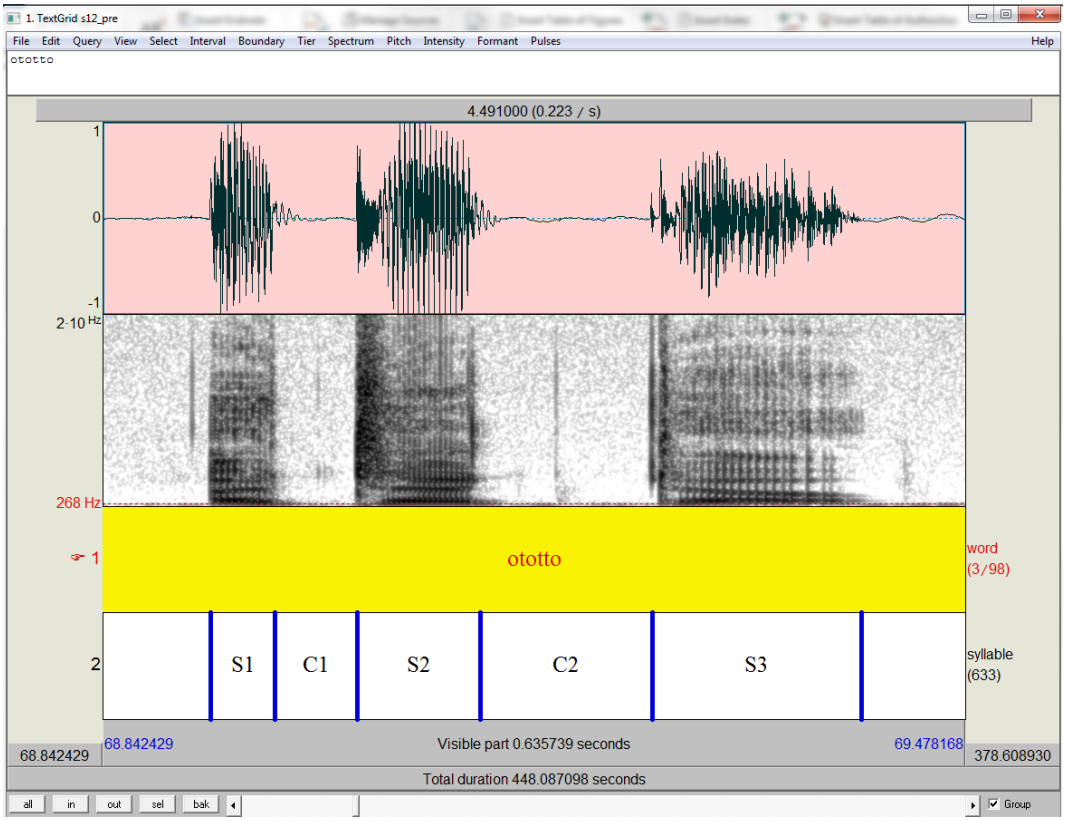

Figure 5. The examples of acoustic measurements. C1 and C2 represents closure durations for plosives, while S1, S2 and S3 are syllable durations excluding the stop closures.

as C1C2. Secondly, the F0 (in Hertz) ratio of S1 to S2, as well as S2 to S3, was measured by extracting F0 for all the voiced portions of S1, S2 and S3 using Praat's autocorrelation method (Boersma, 1993), calculating the mean for each syllable, and then dividing the mean of S1 by S2, as well as S2 by S3. Trials were removed from analysis when any of the syllables was devoiced, thus precluding F0 extraction. Finally, the intensity ratio of S1 to S2, as well as S2 to S3, was obtained by extracting intensity contours using Praat and computing the ratios as with F0.

The initial GLMMs were built using the pass-or-fail judgement by the native listener as the dependent variable and all the above-mentioned acoustic measurements as fixed effects, except for the duration of S3[4], and a random intercept for participants. Since there were four different word types (i.e., GS, SG, GG, SS) in the targets, an initial model was built for each of the four word type separately. After building the initial (full) models, the fixed effects were then selected using backward elimination (i.e., the effects were eliminated one by one, starting from the one with the largest p-value), until all the effects retained in the model were significant ($p>.05$). We only report the final model in the present paper.

The data of 25,681 productions in total, including the pre- and post-test and training[5], was analyzed. For all the following analyses, only the productions for the eight words with the /tt/-/t/

[4] The duration of S3 was not included because it was mathematically redundant: all other durations were converted into ratios to the word length, thus S3 = 100-(S1+C1+S2+C2). In addition, preliminary analyses that used raw durations instead of ratios showed that S3 was the only durational measurement that had little influence on listener's judgement in all models. The duration of S3 was not crucial to judgement in this case probably because participants produced target words without a carrier sentence, thus the last syllable (S3) was often lengthened due to phrase final lengthening, which probably made the duration of S3 less relevant to geminate-singleton contrast.

[5] The total number of 25,681was calculated as follows: originally, one participant was supposed to have 32 (8 words x 4 repetitions) from the pre-test + 256 (8 words x 4 repetitions x 8 blocks) + 32 (8 words x 4 repetitions) from the post-test = 320 productions in total, and therefore the total number should be 320 x 80 participants = 25,600 productions. As explained in a footnote above, participants produced (on average) one extra words during training due to experimental settings, which accounted for the remaining 81 productions.

contrast that participants practiced in the training phase were used, unless noted otherwise. After removing 357 productions from which acoustic measurements could not be properly taken due to omission/insertion of syllables or noises (1.4 % of all trials), 25,324 productions in total were used in this analysis. A separate model was built for each of the four word-types (M=6,331 trials for one word-type[6]).

Subsequently, from these models, we calculated the log odds, which represented how likely each production was judged as "pass" by the listener, given all acoustic measurements. Then the productions with log odds larger than zero (i.e., the probability of being judged as pass was larger than 50%) were labeled "predicted pass", while the rest of the trials were labeled "predicted fail". Finally, we obtained the prediction accuracy by calculating the percentages of productions whose predicted judgement and actual judgement matched.

Accuracy of participant production

The index of accuracy was the pass-or-fail judgement given by the native listener, which was verified in the previous analysis. GLMMs were used to examine (1) the pre-post improvement and (2) the performance during training, the latter only using normal trials (excluding post-feedback trials). Statistical analyses were performed to examine whether the four feedback methods were "effective", which was defined as yielding more improvement than the control (no-feedback) group. Therefore, the Group factor was treatment (dummy) coded using the control group as the reference level, which means that all the comparisons were made between the control group and each of the four feedback groups (this applies to other linear models in the following analyses).

[6] The number of trials for each word type were slightly different: 6343 for the GS, 6269 for the SG type, and 6327 for the GG type, 6385 for the SS type.

For the pre-post improvement analysis, GLMMs were fitted to the data of the pre- and post-test (8 target words x 4 repetitions x 2 tests [pre and post] x 80 participants = 5120 observations) to examine the changes between the two tests. To specify the random effects structure, we started with the maximal model, which contained six random effects that are theoretically possible given the current experimental design, based on recommendations by Barr et al (2013). This approach, including as many random effects as theoretically possible, is more conservative against Type I error (Barr et al, 2013). Both the random effects by-participant and by-item (word) were included (Baayen et al., 2008) in the initial full model, namely random intercepts by Participant (i.e., models assume that each of the eighty participant was different in terms of average performance) and by Word (i.e., models assume that each of the eight target words was different in terms of average difficulty), random slope for Test by Participant (i.e., models assume that each participant was different in terms of how much they improved between the pre- and post-test), random slope for Word Type by Participant (i.e., models assume that each participant performed differently with different word types[7]), and random slope for Test by Word (i.e., models assume that each target word was different in terms of how much participants improved on them between the pre- and post-test). We initially also allowed the correlations between random intercept and random slope, but since the initial full model (that included all the random and fixed effects) did not converge, we removed the interactions, which resolved the conversion issue. Other random effects were all retained in the final model.

[7] For example, the target words with the GG type are generally more difficult than words with the SG type (which was accounted for by the random intercept by Word), but some participants consistently succeed at the former and struggle at the latter. This systematic variance, above and beyond the variance that the random intercept by Word can account for, was accounted for by the random slope for Participant by Word Type.

The fixed effects of interest included Group (five levels), Test (the pre- and post-test), and the interaction between Group and Test (i.e., how well each group learned between the pre- and post-test), the last being of primary interest. To control for other possible covariates that may have an impact on the learning of a new foreign sound, we also included the following five variables. Participants' age was included since age has been shown to affect L2 speech learning (ex., Flege, 1999), although we were aware that age may not be a factor in the current dataset because our participants ranged narrowly in age (most of the participants were between 18 to 23). The number of languages spoken by each participant was also included because our participants had diverse language backgrounds and it has been argued that prior experience in learning an L2 may affect the learning of L3 phonology, sometimes in a beneficial way (Onishi, 2016; Wrembel, 2010). In addition to the above two demographic factors, participants' abilities or status prior to training may directly affect learning. Listening score at the exposure phase (the percentage of correct responses from the forced-choice perceptual task, hereafter abbreviated as Pre-Listening) was included because, although studies have shown that non-native speakers' perception and production skills for the same target L2 sound may not be closely correlated (Kartushina & Frauenfelder, 2014), there is nevertheless a complex relationship between perception and production in L2 speech acquisition (Hao & de Jong, 2016). In addition, as discussed above, we hypothesized that participants' error detection ability and motivation may affect learning outcomes, thus error detection score at the pre-test (see below, hereafter abbreviated as Pre-Error Detection) and baseline motivation scores (see below, hereafter abbreviated as Base-Motivation) were also included to control for possible individual differences in these scores prior to training. Finally, the interactions between each of the above five possible covariates (Age, Number of Language Spoken, Pre-Listening, Pre-Error Detection, Base-

motivation) and Test were included since we needed to control for the effect of these possible covariates on how well participants learned between the pre- and post-test. Because variables with different scales may result in convergence problems in model building, all continuous variables (except for Block in the training analysis, see below) used in this analysis, as well as the following analyses, were centered by subtracting the mean and then rescaled by dividing by the standard deviation unless described otherwise. As with the previous model, these fixed effects were selected using backward elimination and we only report the final model here.

For the training performance analysis regarding normal trials, the same random effects described above for the pre-post model were included in the initial full model. The interactions between random intercepts and random slopes were removed after a conversion issue, and then a random slope for Word was also removed due to a near-zero variance. Other random effects were all retained in the final model. The fixed effects of interest were also identical to what was used for the pre-post model, except that Block (8 blocks as a continuous variable, scaled from zero to one) replaced Test. In addition, to control for learners' individual differences in production ability prior to training, the pre-test score (referred to as Pre-Test Score, calculated as percentages of "pass" and then centered and rescaled as the other continuous variables) and the interaction between Pre-test and Block were also included. Again, the fixed effects that did not reach significant were removed using the backward elimination procedure and only the final model will be presented in the present paper.

Intra-individual acoustic variability

To examine how each participant varied their productions during training and whether there was a relationship between variability and learning outcomes (i.e., accuracy), our first step was to select an acoustic index to represent each production and then to calculate the standard

deviation within an individual. The most straightforward way was to use the closure duration, which was the primary acoustic cue for the geminate. However, after the listener judgement reliability analysis (see 2.4.1 and 3.1), it was revealed that not only the closure duration of the geminate but also the ratio of the two closures (C1 and C2) and other acoustic measurements were relevant to the geminate/singleton contrast in the current dataset. To represent how participants varied their productions on the continuum of the geminate/singleton contrast, we decided to use a composite index of all relevant acoustic measurements by utilizing the models built in 2.4.1 to calculate an index of variability: the standard deviation of the log-odds predictions for productions from the training phase using the above-mentioned models. In other words, we can treat the log-odds prediction as a composite variable of those acoustic measurements—higher values mean that the production's acoustic properties were overall more geminate-like, while lower values indicate that the production sounded more singleton-like.

The variability from trial to trial within a single participant during training was calculated as the standard deviation of the log-odds prediction of the participant's productions for the normal trials in the training phase (i.e., intra-individual standard deviation, ISD). An ISD was first calculated for each of the eight training blocks (32 trials for each), and then the eight ISDs were averaged to obtain an ISD to represent the variability during training. In addition, the ISDs for the pre- and post-test (also 32 trials for each) were calculated in the same fashion.

After obtaining ISDs (1 ISD x 3 phases x 80 participants = 240 ISDs), we built a Linear Mixed Effects Model to examine whether learners' variability in production changed between the pre-test, training, and the post-test. The fixed effects of interest included Group, Phase (three categorical levels: the pre-test, training, and post-test), and the interaction between the two. For the random effect, a random intercept for Participant was included.

We subsequently performed correlation analysis on ISD and accuracy, specifically to examine whether higher variability during training benefited the pre-post improvement in accuracy in each group. To adjust for the initial individual differences, we subtracted the ISDs at the pre-test from ISDs during the training phase. Similarly, the pre-post improvement in accuracy was calculated by subtracting the pre-test judgement score (i.e., the percentages of "pass") from the post-test judgement score. Pearson correlation coefficient was then calculated between the ISD and the pre-post score difference for each of the five groups and the alpha level was changed from 0.05 to 0.01 using Bonferroni correction.

Error detection ability

Participants' error detection ability was represented by the sensitivity index (d′) of signal detection theory (Green & Swets, 1966), using the pass-or-fail judgement by the native listener and participants' self-judgement during the pre- and post-test (see Table 4). For example, when a production was judged as a "fail" by the experimenter, if the participant also self-judged it as fail (i.e. clicked the "I can do better" button), the trial was counted as a "hit" (i.e. correctly detected an error). Likewise, if a participant self-judged a production as a "fail" when the experimenter judged the same production as a "pass," the trial was counted as a "false alarm" (i.e. the participant falsely detected a non-existent error). The error sensitivity index, d′, was calculated from the rates of hits and false alarms according to the formula below:

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$$

A positive d′ indicates that the participants were able to correctly detect errors, while a zero suggests that their judgement was at chance level. A negative d′ (less likely, but possible) indicates that the participant somehow got it reversed – they perceived an error as a correct answer, and vice versa. d′ was calculated separately for the pre- and post-tests.

		Participants' self-judgement	
		Fail ("I can do better")	Pass ("Perfect!")
Experimenter ’s judgement	Fail	**Hit**	**Miss**
	Pass	**False alarm**	**Correct rejection**

Table 4. Categories of the signal detection theory using the native listeners’ judgement and participants’ self-judgement

After calculating d′, a Linear Mixed Model was fitted to investigate whether error detection ability improved between the pre- and post-tests. We included Group (5 levels), Test (pre or post), and the interaction of the two (Group x Test) as fixed effects, with the interaction being of primary interest. As covariates, we also included Pre-Listening (assuming higher initial listening scores were correlated with higher error detection ability at the pre-test), and the interaction between Pre-Listening and Test (assuming higher initial listening scores would predict larger improvement in error detection scores between the pre- and post-test) as fixed effects. For the random structure, a random intercept by Participant was included.

Motivational Score (baseline and task-specific)

Finally, our last set of analyses focused on the role of motivation. In the scoring system for the seven-point Likert scale questionnaire described in 2.3.2 and 2.3.5, higher numbers indicate that participants felt more motivated toward pronunciation exercises in general (baseline motivation) or toward this training specifically (task-specific motivation). The term "motivated" here means that participants enjoyed the experience more, felt more confident, and felt less pressured or nervous during the activity in question.

Since the Likert scale was as an ordinal variable, we ran an ordinal logistic regression (using the *ordinal* package in R) with Group (5 levels), Questionnaire (baseline and task-specific, with the former being the reference level), and the interaction of the two (Group x Questionnaire) as fixed effects and an intercept by Participant and by Item (12 items in each questionnaire) as random effects.

3. Results

3.1. Reliability of the native listener judgement

Table 5 presents the fixed effect estimates from the final models for each word type. Note that not only the closure duration, but also the ratio of the two closures and some of the other secondary cues (the duration and intensity of the adjacent vowels) had significant effects on the listeners' judgement. The results of the prediction accuracy from the above models showed that the acoustic measurements predicted the listener judgement correctly for 90.4% of productions for the GS type, 94.5 % of the SG type, 88.4 % of the GG type, and 94.5 % of the SS type (*M* = 91.9 %). Cohen's kappa coefficients were .807, .870, .767, and .653, respectively[8]. Since these coefficients were deemed to be "substantial agreement (.61 to .80)" or "almost perfect agreement (> .81)" (Landis & Koch, 1977, p.165), we concluded that the listener's judgements were justifiable by acoustic data, therefore reliable enough to be used as the indicator of pronunciation accuracy in the following analyses.

3.2. Accuracy of participant production

To overview changes in accuracy by groups over the course of the experiment, Figure 6 presents the average percentages of "pass" that participants in each group received in the pre-

[8] The kappa coefficient for the SS type was lower than other word types despite its high agreement rate, because this word type was very easy for participants to pronounce (90.8% of all trials for this word type were judged as pass). Cohen's kappa mathematically cannot be high when classifications are unbalanced.

	GS					SG				
	β	Std. Err.	z value	p		β	Std. Err.	z value	p	
(Intercept)	-2.36	0.21	-11.28	0.000	*	-5.08	0.64	-7.98	0.000	*
C1 duration	1.51	0.10	14.48	0.000	*	-1.62	0.45	-3.61	0.000	*
C2 duration	-2.70	0.15	-17.96	0.000	*	1.46	0.33	4.43	0.000	*
C1/C2	-0.17	0.02	-7.74	0.000	*	-25.21	5.97	-4.23	0.000	*
V1 duration	-0.32	0.08	-4.09	0.000	*	-1.35	0.15	-8.88	0.000	*
V2 duration	-1.00	0.10	-9.62	0.000	*	-0.26	0.10	-2.67	0.008	*
Int. V1/V2	0.27	0.08	3.29	0.001	*	-0.38	0.09	-4.20	0.000	*
Int. V2/V3	-0.20	0.06	-3.16	0.002	*	0.17	0.08	2.06	0.040	*
F0 V1/V2										
F0 V2/V3										
	GG					SS				
	β	Std. Err.	z value	p		β	Std. Err.	z value	p	
(Intercept)	-4.35	0.27	-15.86	0.000	*	2.32	0.34	6.82	0.000	*
C1 duration	6.57	0.24	27.06	0.000	*	-1.69	0.20	-8.32	0.000	*
C2 duration	-3.29	0.22	-15.17	0.000	*	-2.77	0.21	-13.31	0.000	*
C1/C2	-54.71	2.33	-23.50	0.000	*	-6.12	1.13	-5.40	0.000	*
V1 duration	0.18	0.09	2.00	0.045	*	-0.92	0.11	-8.02	0.000	*
V2 duration	-0.36	0.09	-3.80	0.000	*					
Int. V1/V2	0.42	0.07	5.80	0.000	*	0.34	0.09	3.93	0.000	*
Int. V2/V3										
F0 V1/V2										
F0 V2/V3										

Table 5. The estimates for fixed effects from the final GLMMs with Japanese listener's perceptual judgement as the dependent variable and acoustic measurements as fixed effects. Blank rows indicate that those variables were eliminated during the backward elimination process.

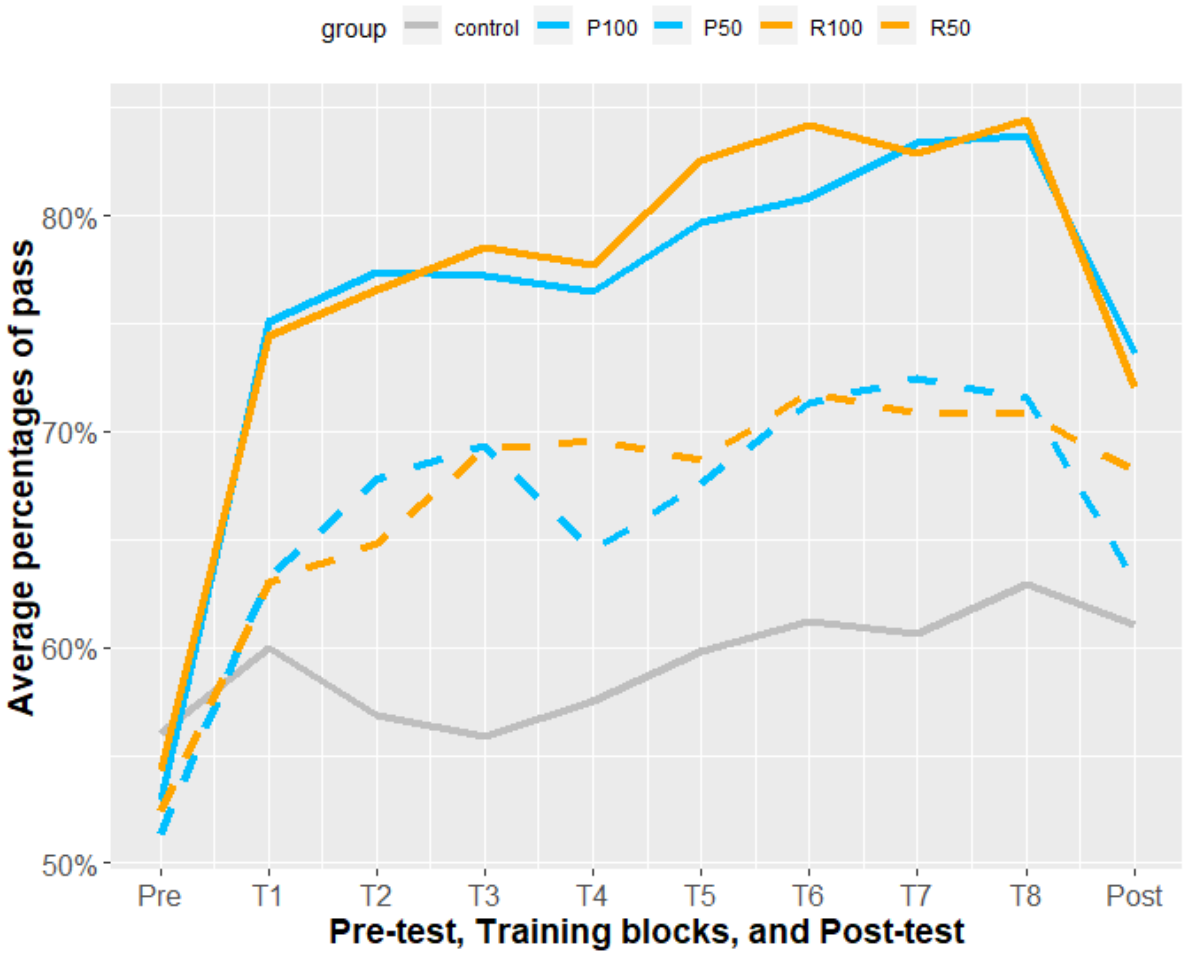

Figure 6. The average percentages of "pass" that participants in each group received in the pre-test, each of the eight blocks in the training phase, and the post-test. For the training phase, Figure 6 only presents the data for normal trials. Due to the unbalanced number of normal trials in the training phase, percentages were first calculated for each word within participant, and then averaged for each participant, then for each group.

test, each of the eight blocks in the training phase, and the post-test. For the training phase, Figure 6 only presents the data for normal trials (excluding post-feedback trials)[9].

Pre-post changes

After the initial full model did not converge, the interaction between random intercept and slope for Participant and for Word were removed, leaving all the remaining random intercepts and slopes in the final model. The fixed effect estimates from the final model are shown in Table 6. First let us look at the effects of the interaction between Group and Test, which is of primary interest for our analysis. We found that the P100 (β = 1.75, z = 4.26, p < .001), R100 (β = 1.07, z = 2.63, p = .008), and R50 (β = 0.91, z = 2.28, p = .023) groups showed significantly greater changes than the control group, while the difference in changes between the P50 group and the control group did not reach significance (β = 0.43, z = 1.04, p =. 297).

Other fixed effects were not of primary interest to the analysis, but they confirmed several characteristics of the data. The effect of Test with a positive β estimate indicates a positive change between the pre- and post-test for the control (reference) group (β = 0.78, z = 2.65, p = .008), suggesting that the target geminate can be learned to a certain degree even without any feedback. The Group effects confirmed that the four groups were at a similar level as the control group at the time of the pre-test (all p > .05). The positive significant estimate for Pre-Listening (β = 1.11, z = 5.64, p < .001) means that participants who initially had higher listening scores already did better in the pre-test, which suggests perception and production were

[9] Due to the unbalanced number of normal trials in the training phase (since the number of trials was controlled so that the sum of normal and post-feedback trials should result in the same number across words and participants; see Methods section for details), percentages in Figure 6 were first calculated for each word within participant, and then averaged for each participant, then for each group.

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.87	0.85	1.01	0.311	
Test (for the control group)	0.78	0.29	2.65	0.008	*
Group (P100 vs. control)	-0.63	0.59	-1.07	0.285	
Group (P50 vs. control)	-0.78	0.61	-1.29	0.199	
Group (R100 vs. control)	-0.53	0.60	-0.88	0.378	
Group (R50 vs. control)	-0.06	0.60	-0.10	0.922	
Pre-Listening	1.11	0.20	5.64	0.000	*
Pre-Listening x Test	0.59	0.14	4.31	0.000	*
Pre-Error Detection	0.62	0.19	3.22	0.001	*
Group x Test (P100 vs. control)	1.75	0.43	4.12	0.000	*
Group x Test (P50 vs. control)	0.43	0.41	1.04	0.297	
Group x Test (R100 vs. control)	1.07	0.40	2.63	0.008	*
Group x Test (R50 vs. control)	0.91	0.40	2.28	0.023	*

Table 6. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the pre-post change in productions of trained items (/tt-t/ contrast).

correlated in the current dataset. Similarly, the positive significant estimate for the interaction between Pre-Listening and Test ($\beta = 0.60$, $z = 4.31$, $p < .001$) also indicates that higher initial listening skills predicted greater pre-post improvements. Likewise, higher initial error detection ability predicted better pre-test performance ($\beta = 0.62$, $z = 3.22$, $p = .001$). Other variables, such as Age, Number of Language Spoken, Base-motivation did not have significant effects and excluded from the final model.

We also analyzed the data for productions that included the /kk/-/k/ contrast in the same way as described above to see whether the learning effects generalized to another geminate consonant that participants did not practice in the training phase. The final model (Table 7) shows that all groups (including the control group) improved in the /kk/-/k/ contrast between the pre- and post-test, but there were no significant group differences, and the estimates were generally smaller than those for the /tt/-/t/ contrast. This suggests that although the training effects on one particular geminate consonant can be generalized to another geminate, we did not observe clear effects of either type or frequency of feedback on participants' generalization ability.

Training performance

Table 8 shows the average number and accuracy of normal and post-feedback trials during the training phase per participant for each group. All groups (except for the control group who performed all 256 trials as normal trials) performed on average 217 normal trials during training. This means that participants on average received CF (hence performed post-feedback trials) 39 times. Note that the instances of CF received were relatively close between the 100 % groups (44 times for P100 and 42 times for R100) and the 50% groups (35 times for both the P50 and R50 group) despite the difference in feedback frequency. As may be seen in Figure 6, the

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.43	0.53	0.82	0.414	
Test (for the control group)	0.97	0.27	3.58	0.000	*
Group (P100 vs. control)	-0.20	0.43	-0.46	0.644	
Group (P50 vs. control)	-0.83	0.42	-1.98	0.047	*
Group (R100 vs. control)	-0.44	0.45	-0.97	0.331	
Group (R50 vs. control)	-0.37	0.42	-0.88	0.380	
Pre-Listening	0.79	0.15	5.33	0.000	*
Pre-Listening x Test	0.30	0.12	2.56	0.011	*
Number of Language Spoken	-0.08	0.14	-0.56	0.579	
Number of Language Spoken x Test	0.26	0.12	2.20	0.028	*
Pre-Error Detection	0.35	0.14	2.56	0.010	*
Group x Test (P100 vs. control)	0.42	0.37	1.13	0.260	
Group x Test (P50 vs. control)	-0.11	0.37	-0.29	0.773	
Group x Test (R100 vs. control)	0.56	0.37	1.52	0.128	
Group x Test (R50 vs. control)	0.37	0.36	1.01	0.313	

Table 7. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the pre-post change in productions of untrained items (/kk-k/ contrast).

	Normal trial		Post-feedback trial	
	n	Percent of "pass"	n	Percent of "pass"
Control	256	59.3%	-	-
P100	213	77.4%	44	56.1%
P50	222	66.3%	35	41.3%
R100	215	78.1%	42	86.6%
R50	222	67.5%	35	81.9%

Table 8. The number and average percentage of "pass" of normal and post-feedback trials during the training phase.

P100 and R100 group were faster in their improvement, which implies that they received CF at high frequency but only in the early few blocks, whereas the P50 and R50 group received CF at reduced frequency but they continued to receive CF throughout the training phase, resulting in the relatively close number of instances of CF received in the 100% and 50% groups.

The fixed effect estimates from the final model for normal trials during the training phase are shown in Table 9. First, the estimates for the Group effect suggest that the P100 (β = 0.60, z = 2.02, p = .043) and R100 (β = 0.87, z = 2.79, p = .005) groups already outperformed the control group at the first block of training, while both the effects for the P50 and R50 group indicated that their performances were still close to those of the control group at the first block (β = 0.20, z = 0.67, p = .502 and β = 0.27, z = 0.90, p = .368 respectively). Second, the interaction effect of Group and Block for the R100 group suggests that the R100 group improved more over the rest of the blocks than the control group (β = 0.84, z = 2.18, p = .030). The other three groups, P100, P50 and R50, also obtained positive β estimates, which implies that their improvement rates were numerically better than the control groups, but the differences were not significant (β = 0.51, z = 1.36, p = .174 for P100, β = 0.35, z = 0.94, p = .347 for P50, and β = 0.25, z = 0.68, p = .0498 for R50).

As for other covariates, the positive effects of Pre-Test Score and Pre-Listening suggest that participants with higher production and/or perception skills prior to training did better at the first block. Participants with higher pre-test scores also continued to improve faster over the course of training.

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.75	0.52	1.44	0.151	
Block (for the control group)	0.47	0.27	1.77	0.077	
Group (P100 vs. control)	0.60	0.30	2.02	0.043	*
Group (P50 vs. control)	0.20	0.30	0.67	0.503	
Group (R100 vs. control)	0.87	0.31	2.78	0.005	*
Group (R50 vs. control)	0.27	0.30	0.90	0.368	
Pre-Test Score	1.11	0.12	9.03	0.000	*
Pre-Test Score x Block	0.30	0.13	2.29	0.022	*
Pre-Listening	0.40	0.11	3.53	0.000	*
Group x Block (P100 vs. control)	0.51	0.38	1.36	0.175	
Group x Block (P50 vs. control)	0.35	0.38	0.94	0.347	
Group x Block (R100 vs. control)	0.84	0.38	2.17	0.030	*
Group x Block (R50 vs. control)	0.25	0.37	0.68	0.498	

Table 9. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the change during the training phase (normal trials only).

3.3. Intra-individual acoustic variability

Figure 7 presents the average ISDs for the pre-test, training, and post-test phase for each group. ISDs at the pre-test seemed somewhat different between groups (which, if true, was not ideal), but the estimates from the LMM (Table 10) indicated no significant differences between each of the four groups and the control group at the pre-test (β = 0.60, t = 1.77, p = .079 for P100, β = 0.06, t = 0.18, p = .857 for P50, β = 0.43, t = 1.28, p = .203 for R100, β = 0.52, t = 1.57, p = .120 for R50).

Participants in the P100 group significantly reduced their variability in the training phase compared to the reduction for the control group (β = -0.68, t = -2.05, p = .043). The R100 group also showed a marginally significant reduction at the training phase compared to the control group (β = -0.63, t = -1.90, p = .059), but both the P50 and R50 group did not (β = -0.19, t = -0.59, p = .560 for P50 and β = -0.04, t = -0.14, p = .892 for R50). At the time of the post-test, only the P100 group showed significantly reduced variability compared to the control group (β = -1.02, t = -3.07, p = .003).

Figure 8 shows scatter plots for ISDs during training (after adjusting for the initial variability at the pre-test) and the listener's judgement score difference between the pre- and post-test. The correlation analysis revealed that, among the P100 group, smaller variability during training was correlated with larger pre-post improvement in accuracy (R = -0.62, p = .010), whereas the same correlations did not reach significance for the other groups (R = -0.48, p = .066 for the controls, R = -0.35, p = .182 for P50, R = -0.27, p = .302 for R100, R = -0.36, p = .155 for R50).

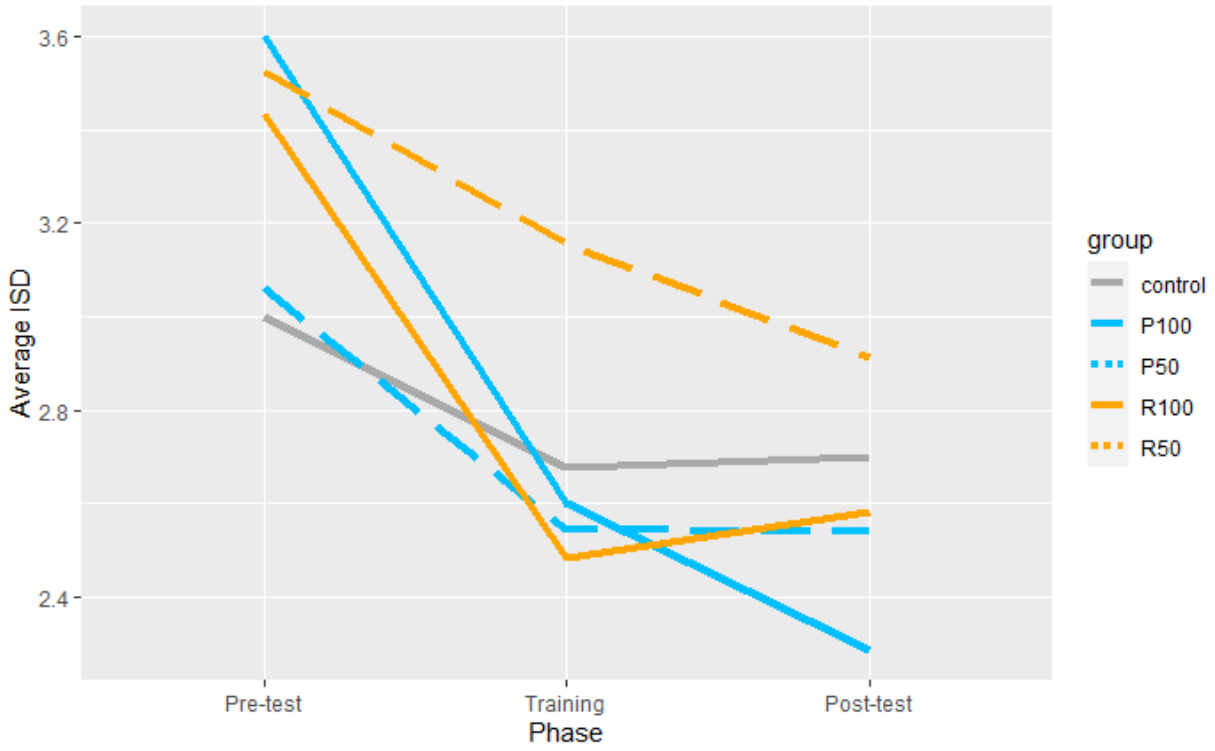

Figure 7. Average intra-individual acoustic variability (ISD) for each group at the pre-test, the training phase, and the post-test.

Fixed effects	**Estimate (β)**	**Std. Error**	**t value**	**p**	
Intercept	3.00	0.24	12.27	0.000	*
Group (P100 vs. control at Pre-test)	0.60	0.34	1.77	0.079	
Group (P50 vs. control at Pre-test)	0.06	0.34	0.18	0.857	
Group (R100 vs. control at Pre-test)	0.43	0.34	1.28	0.203	
Group (R50 vs. control at Pre-test)	0.52	0.34	1.57	0.120	
Phase (Pre-test vs. Training for the controls)	-0.32	0.24	-1.35	0.178	
Phase (Pre-test vs. Post-test for the controls)	-0.30	0.24	-1.25	0.212	
Group x Phase (P100 vs. control at Training)	-0.68	0.33	-2.05	0.043	*
Group x Phase (P50 vs. control at Training)	-0.19	0.33	-0.59	0.560	
Group x Phase (R100 vs. control at Training)	-0.63	0.33	-1.90	0.059	.
Group x Phase (R50 vs. control at Training)	-0.04	0.33	-0.14	0.892	
Group x Phase (P100 vs. control at Post-test)	-1.02	0.33	-3.07	0.003	**
Group x Phase (P50 vs. control at Post-test)	-0.22	0.33	-0.67	0.505	
Group x Phase (R100 vs. control at Post-test)	-0.55	0.33	-1.67	0.096	
Group x Phase (R50 vs. control at Post-test)	-0.35	0.33	-1.07	0.286	

Table 10. The fixed effect estimates from the LMM fitted to ISD (intra-individual acoustic variability).

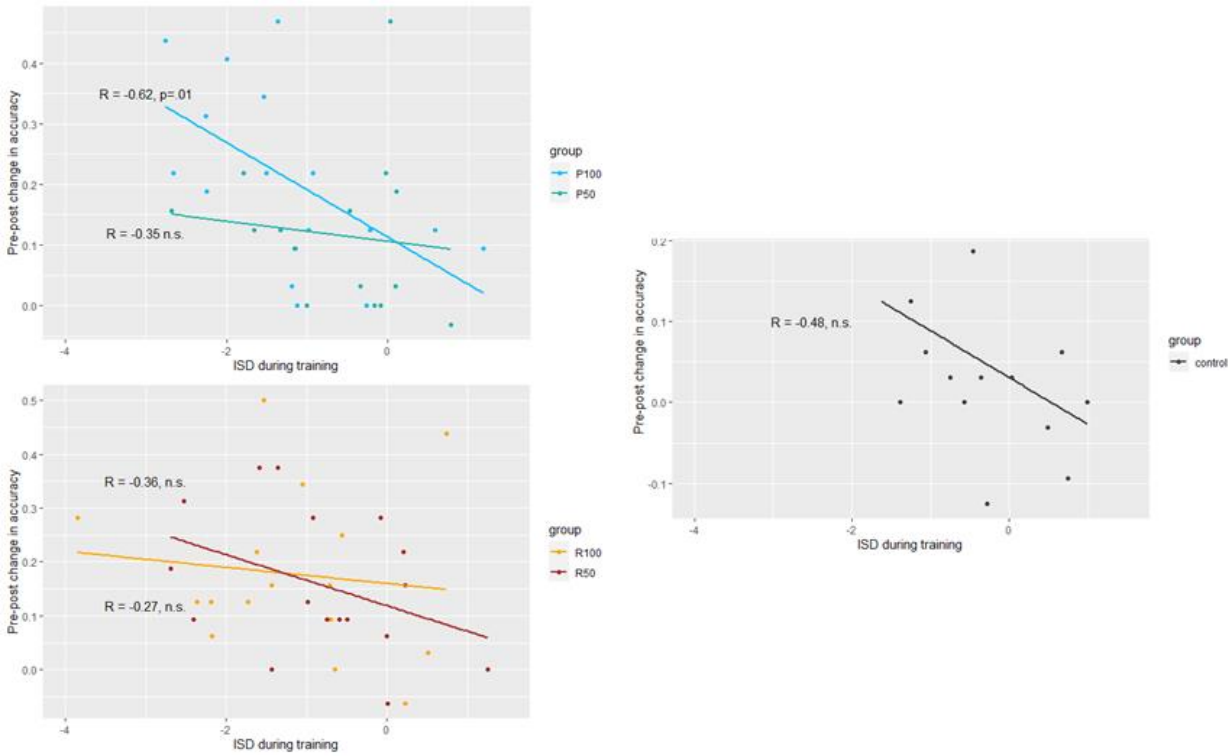

Figure 8. Scatter plots for the ISDs during training (after subtracting from the pre-test ISD, thus lower values indicate less variability during training) and the listener's judgement score difference between the pre- and post-test (higher values indicate larger improvements).

3.4. Error detection ability

Figure 9 presents average d-prime (d′) scores at the pre- and post-test for each group. Table 11 shows the estimates from the fitted Linear Mixed model. The results showed that the effect of Test for the control group did not reach significance (β = -0.15, t = -0.78, p = .437) and the estimate was negative, which suggests little (if anything, negative) change between the pre- and post-test for the control group. Compared with the control group, the P100 and P50 groups showed positive changes (β = 0.57, t = 2.19, p = .032 and β = 0.55, t = 2.09, p = .040, respectively), while the results for the R100 and R50 groups yielded non-significant effects with close to zero estimates (β = 0.01, t = 0.05, p = 965 for R100, β = 0.06, t = 0.23, p = .820 for R50), which suggests that the groups who received prompts showed some improvements in error detection, whereas participants who trained with recasts or no feedback did not show such improvement.

3.5. Motivation

Figure 10 shows average ratings on a seven-point Likert scale from the baseline and task-specific motivation questionnaires. The parameter estimates from the ordinal logistic model are shown in Table 12. Notice that the control group showed significantly higher task-specific scores than the baseline (β = 0.47, t = 2.54, p = .011), and only the R100 group showed positive β estimates for the Group x Questionnaire effect. As for the other three groups that showed negative estimates, it was not clear whether participants in those groups exhibited higher task-specific motivation than the baseline[10]. This led us to perform post-hoc pairwise comparisons

[10] That is, if the baseline vs. task-specific comparison (the Questionnaire variable) was significant in the control group and the β for Questionnaire x Group were positive, that means that the baseline vs. task-specific comparison is also

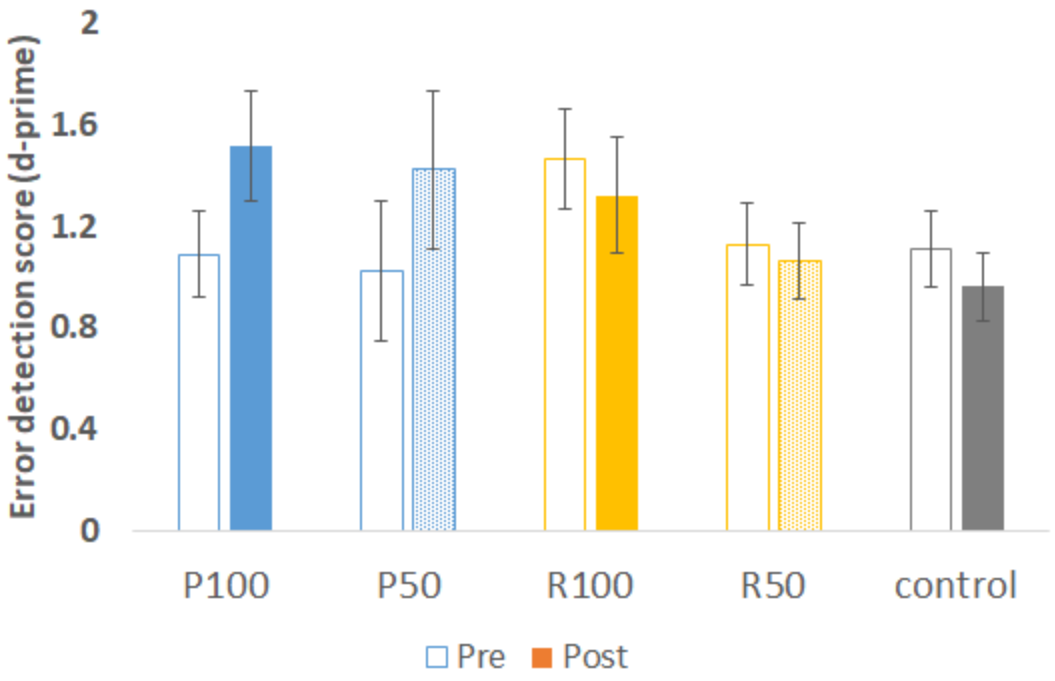

Figure 9. Average error detection ability (d-prime) for each group at the pre- and post-test. The error bars indicate the standard error of the mean.

significant because there is a larger difference between baseline and task-specific than for the control group. However, if the β estimates are negative, that just indicate that the difference between baseline and task-specific is smaller than for the control group, which does not indicate statistical significance of the baseline vs. task-specific difference.

Fixed effects	Estimate (β)	Std. Error	t value	p	
Intercept	1.11	0.21	5.24	0.000	*
Test	-0.15	0.19	-0.78	0.437	
Group (P100 vs. control)	-0.02	0.29	-0.07	0.948	
Group (P50 vs. control)	-0.09	0.29	-0.29	0.772	
Group (R100 vs. control)	0.29	0.29	1.00	0.320	
Group (R50 vs. control)	0.06	0.29	0.21	0.836	
Group x Test (P100 vs. control)	0.57	0.26	2.19	0.032	*
Group x Test (P50 vs. control)	0.55	0.26	2.09	0.040	*
Group x Test (R100 vs. control)	0.01	0.26	0.05	0.965	
Group x Test (R50 vs. control)	0.06	0.26	0.23	0.820	

Table 11. The fixed effect estimates from the LMM fitted to d-prime (error detection ability).

between the baseline and task-specific motivation scores for each of the five groups using the package *emmeans* (Russell Lenth, 2020). The alpha-level was changed from .05 to .01 (Bonferroni correction). The results confirmed that participants in the R100 group were more motivated by the current training setting than they normally would be in pronunciation training (β = -0.69, z = -4.57, $p < .001$), and also showed the same tendency for the P50 group at a marginal level (β = -0.69, z = -2.38, p = .017) and the control group (β = -0.39, z = -2.55, p = .011), whereas those in the P100 and R50 group were not (β = -0.13, z = -0.82, p = .410 and β = -0.16, z = -1.05, p = .293, respectively).

In an effort to determine which aspect of motivation was the most affected, Figure 11 breaks down the scores into the three subscales of the questionnaires, namely Interest/Enjoyment Perceived Competence, and Pressure/Tension. It presents the differences in average ratings between the baseline and task-specific motivation scores (for Pressure, higher scores mean that the participant felt less pressured). We observe that, compared to the other three feedback groups, the biggest advantage for the R100 group seemed to be in Pressure/Tension, suggesting that participants tended to feel more relaxed if they were given recasts 100% of the time.

4. Discussion

In this study, we investigated the effect of feedback frequency (100% or 50%) on CF by using laboratory pronunciation training. Specifically, we investigated the possibility that the effect of frequency may be different for two types of CF (prompts and recasts). We predicted that prompts could benefit from high frequency both during training and at retention (i.e., the post-test), whereas frequent recasts may hinder retention compared to reduced recasts, even though frequent recasts enhance performances greatly during training.

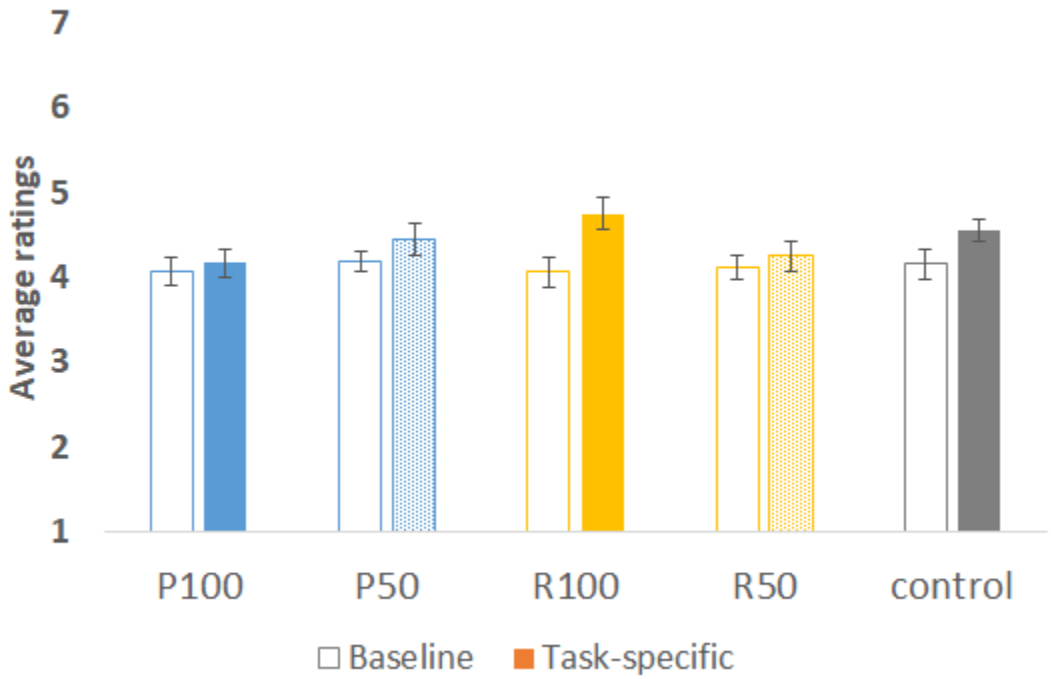

Figure 10. Average motivational scores (7-poing Likert scale) for each group at the pre- and post-test. The error bars indicate the standard error of the mean.

Fixed effects	**Estimate (β)**	**Std. Error**	**t value**	**p**	
Questionnaire	0.47	0.18	2.54	0.011	*
Group (P100 vs. control)	-0.11	0.26	-0.40	0.691	
Group (P50 vs. control)	-0.01	0.26	-0.04	0.970	
Group (R100 vs. control)	-0.20	0.27	-0.77	0.441	
Group (R50 vs. control)	-0.03	0.26	-0.13	0.899	
Group x Questionnaire (P100 vs. control)	-0.32	0.26	-1.23	0.218	
Group x Questionnaire (P50 vs. control)	-0.04	0.26	-0.14	0.890	
Group x Questionnaire (R100 vs. control)	0.36	0.26	1.40	0.162	
Group x Questionnaire (R50 vs. control)	-0.28	0.26	-1.11	0.266	

Post-hoc pairwise comparisons	**Estimate (β)**	**Std. Error**	**z value**	**p (α=.01)**	
Control (base vs. task-specific)	-0.39	0.15	-2.55	0.011	.
P100 (base vs. task-specific)	-0.13	0.15	-0.82	0.410	
P50 (base vs. task-specific)	-0.36	0.15	-2.38	0.017	.
R100 (base vs. task-specific)	-0.69	0.15	-4.57	0.000	*
R50 (base vs. task-specific)	-0.16	0.15	-1.05	0.293	

Table 12. The fixed effect estimates from the ordinal logistic regression model fitted to Likert-scale cores in the motivational questionnaires (base and task-specific).

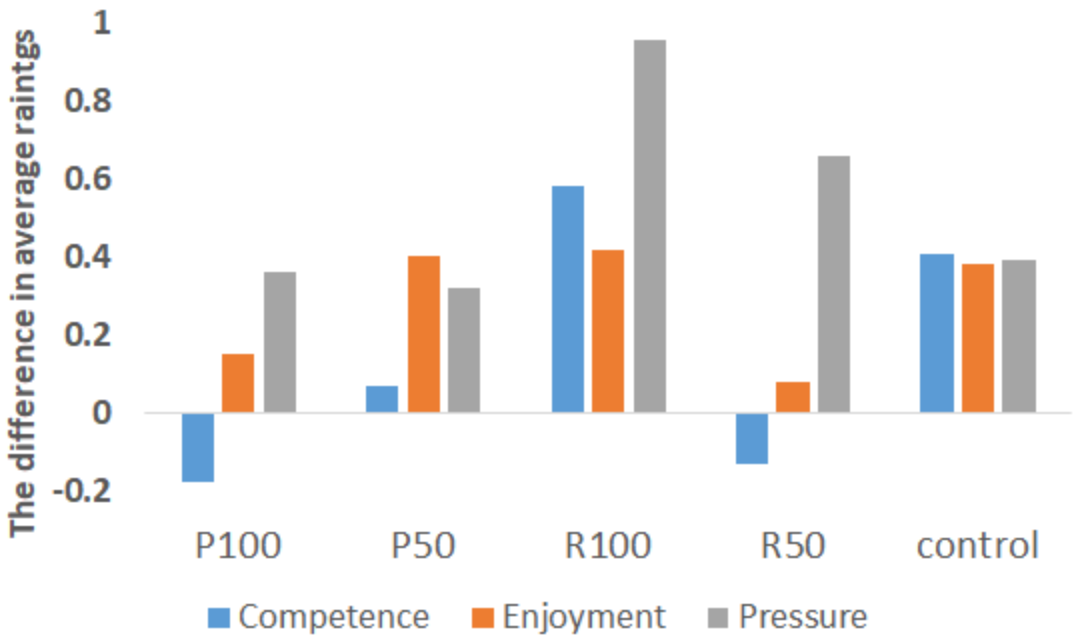

Figure 11. Differences in average scores between the baseline and task-specific motivation scores. Positive values indicate higher task-specific scores than the baseline. Note: for Pressure, higher scores mean that the participant felt *less* pressured.

4.1. The effect of frequency on prompts

In terms of prompts, we found that the P100 group improved more in accuracy between the pre- and post-test than the control (no-feedback) group while the P50 group failed to outperform the control group, which was in line with our predictions. It seemed that frequent prompts took effect quickly once given to learners, considering that the P100 group already outperformed the control group at the first block of training (within the first 32 trials), although they did not learn significantly faster than the control group after the first block. The P50 group, on the other hand, did not significantly outperform the control group at the first block of training, during the rest of the training, or at the post-test.

We investigated two underlying factors as to why high frequency could be beneficial to those who are provided with prompts. The first was because prompts may encourage error detection by requesting self-correction from learners and making them more aware of errors, although frequent feedback has been suggested to hinder error detection ability (Salmoni et al., 1984; Schmidt, 1991). Results showed that prompts did seem to enhance participants' error detection ability: both the P100 and P50 group improved in error detection ability between the pre- and post-test, while we observed virtually no change in the control group. We originally expected that we would see a greater improvement in the P100 group compared to the P50 prompt group, because the former group was more frequently asked to find errors in their productions, but this was not the case (although the P100 group showed a numerically better improvement). It is possible that our measurement for error detection ability, which was a pass-or-fail type of self-judgement, was not fine-grained enough to detect the small difference between the two groups. Nevertheless, this finding suggests that prompts' request for self-correction can effectively direct learners' attention to errors and relevant *intrinsic* feedback,

compensating for the potential disadvantage of frequent extrinsic feedback such as CF (Guadagnoli & Kohl, 2001; Silva et al., 2017 for similar results in non-speech tasks).

The second factor was variability in production. The previous studies on feedback frequency suggested that frequent feedback may increase variability in learners' performance because frequent feedback may have learners focused on small, insignificant errors (Lai & Shea, 1998; Schmidt & Bjork, 1992), but we suspected that, for CF such as prompts and recasts, this may not be the case. The variability analysis revealed that, contrary to the previous studies' suggestion, the P100 group showed significantly reduced variability during training compared to the control group, whereas the P50 group did not. This suggests that prompts can effectively reduce instability in learners' productions by correcting errors, and they need to be frequent to do so. Furthermore, only among the P100 group, smaller variability in productions during training was significantly correlated with better learning outcomes between the pre- and post-test. In other words, stable performance during training was retained in the post-test only in the P100 group. Together with the above finding about error detection ability, we speculate that participants who received prompts learned the new phoneme by figuring out what aspects of their productions were errors; those who successfully figured out the rules about errors (hence stable performance) during training tended to retain good performance at the post-test; those who received feedback less frequently may have tried the same learning strategy but may not have gathered enough information to establish their own criteria about errors, and they thus showed poorer learning outcomes.

4.2. The effect of frequency on recasts

In contrast with prompts, we expected that there may be a different pattern for recasts; 100% recast would be effective for training, while 50% would be more effective for retention –

because we assumed that recasts, by letting learners imitate the auditory models, would make it easy for learners to fix errors during training, but in turn may impede the error detection process on their own. These predictions were essentially supported, although evidence was somewhat indirect. First, it seemed that frequent recasts were effective during training compared to reduced recasts: the R100 group had a clear advantage during training compared to the control group. Its advantage was seen as early as the first block of the training and continued throughout the rest of the training blocks, while the R50 group failed to statistically outperform the control group both at the first block and for the rest of the training. This demonstrates that recasts can improve learners' performances quickly and consistently during training when provided frequently. Second, it seemed also that reduced recasts were more effective for retention – but this was shown only indirectly. The R50 group did not overperform the R100 group at the post-test as predicted, but both the R100 and R50 recast group outperformed the control group. The fact that both of the groups performed well at the post-test despite the disadvantage that the R50 group showed during training suggests that the R50 group retained better than the R100 group (which was also visually shown in Figure 6, as a smaller decline for the R50 group between the last block of training and the post-test compared to the R100 group). Although previous studies often found that the group with reduced feedback surpassed the frequent feedback group at the post-test due to superior retention (e.g., Winstein & Schmidt, 1990; Sparrow & Summers, 1992; Adams & Page 2000; Steinhauer & Greyhack, 2000; Sidaway et al., 2008; Kim et al., 2012), that was not the case in the current study. We think that this was probably because, in this study, the

advantage for frequent recasts during training was very powerful[11], therefore even if frequent recasts might lead to degraded retention, the remaining effect was still as good as reduced feedback. Although the present results did not replicate a clear advantage of reduced recasts, we observed the predicted pattern – the apparent advantage of frequent feedback over reduced feedback during training disappeared at the post-test – and it was different from the results that we observed for prompts.

In addition, we also found that recasts did not encourage error detection: neither 100% nor 50% recasts seemed to have an impact on error detection ability, just like the control condition. The fact that the 50% recast group, who were supposed to rely more on their own judgement due to the reduced frequency of feedback, did not show any improvement suggests that the presence of recasts, regardless of frequency, prevented learners from becoming independently aware of errors. These results are in line with the account that the provision of auditory model right after learners' erroneous production may effectively block learners' error detection process, and against the alternative account that recasts can offer the opportunity to compare the model and their own productions, which may increase error detection ability (Long, 1996). The finding that both the 100% and 50% recast group did improve in accuracy despite their lack of improvement in error detection implies that the underlying learning strategy used by those who received recasts might not involve the ability to detect errors in their own productions, whereas we observed the potential involvement of error detention ability for those who received prompts.

[11] One of the reasons may be because geminates are essentially a temporal feature, which may be especially easy to imitate immediately after listening to the model. Had we used other phonemes, such as the ones that have spectral features, the training effect of frequent recasts might have been less powerful.

We also observed significantly reduced variability during training for the R100 group, but not for the R50 group. This suggests that, together with the same finding for the P100 group, frequent CF reduces learners' variability in production by correcting errors, which again indicates that the concern that learners may attempt to correct their own productions unnecessarily when receiving too frequent feedback (Lai & Shea, 1998) is not warranted for CF such as prompts and recasts. What interests us more is that the correlation analysis revealed no significant relationships between variability in productions during training and better pre-post improvement for either the R100 group or the R50 group – whereas we observed a significant correlation for the P100 group. In other words, for participants who received recasts, being able to perform stably during training did not necessarily predict a better outcome at the post-test. We speculate that this may be because participants' stability in performance during training, when receiving recasts, was more correlated with immediate imitation rather than long-term learning. Together with the error detection analysis, these results imply that participants may have used different learning mechanisms according to the type of feedback with which they were provided.

4.3. Feedback frequency and motivation

We also examined whether CF with high frequency was more demotivating for learners than CF with reduced frequency, which is one of the concerns held by teachers (Lasagabaster & Sierra 2005: Vásquez & Harvey, 2010; Mori, 2011; Méndez & Cruz, 2012). We found that, for recasts, not only was this not the case, but the opposite was true: the R100 group was more motivated by the current training than at baseline, whereas the R50 group was not (considering that the control group, who did not receive any feedback, showed an increase in the task-specific motivation while the R50 group did not, reduced recasts may have in fact decreased participants' motivation). Among the three subscales of the questionnaire, the R100 group's advantage was

seen in the Pressure scale, which means that participants felt less pressured when provided with recasts frequently. This may suggest that learners perceive recasts as the source of the auditory models that they could imitate, and they did not focus on the aspect of recasts as CF. The current result is in line with teachers' perception that recasts are less damaging to learners' motivation (Yoshida, 2008), but it is interesting to observe that frequent recasts are not only less demotivating but actually increase motivation. On the other hand, for participants who received prompts, we found that the P50 group increased motivation following the current training at a marginally significant level, whereas the P100 group did not. Again, considering that the no-feedback condition showed better task-specific motivation, it is possible to assume that providing prompts frequently may actually decrease motivation, although evidence was statistically weak. These results may be in line with the notion that providing CF, specifically prompts, would decrease leaners' motivation but that reducing its frequency may mitigate the negative effect. Taken together, the findings highlight the differential effects of prompts and recasts on learners' motivation.

5. Conclusion, implications and limitations

The current results offer implications for both teaching practice and research. Practically, we are able to draw tentative suggestions as to how frequently prompts or recasts should be provided. First, the effect of prompts on learners' accuracy of the target phoneme seem to be weakened when provided less frequently, thus it would be reasonable to consider providing them as frequently as possible. One, then, may be concerned whether frequent prompts would lower learners' motivation, which is potentially possible, although evidence in the current results is not strong (we only found a marginally significant effect). Prompts may also encourage learners to

be aware of their own errors, which may be a plus if one would like learners to have a conscious understanding of the target sound. Secondly, recasts, in contrast, seem to retain their effectiveness on accuracy even when provided less frequently, thus we would suggest that recasts could be used at different frequencies depending on the needs and conditions, which might be convenient if one is concerned that providing feedback frequently would be distracting or time-consuming, but note that learners may find it more encouraging when they receive recasts as frequently as possible. One should also keep in mind that learners may appear to be performing very well during training when provided recasts frequently, but the performance will most likely decline after training.

The present results may also offer a new perspective as to why there were discrepancies in the previous motor learning literature on feedback frequency. It has been repeatedly found that reduced feedback is beneficial to retention (Sparrow & Summers, 1992; Badets & Blandin, 2004; Sidaway et al., 2008; Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012), while other researchers have found an overall advantage of frequent feedback (Guadagnoli et al. 1996; Wulf et al., 1998; Wulf & Shea 2002; Chiviacowsky et al., 2008; Sullivan et al., 2008) and some have argued that the advantage of reduced feedback may only be seen in simple tasks and that complex tasks benefit from frequent feedback. The current study found both results—an advantage of reduced feedback to retention in recasts and an overall advantage of frequent feedback in prompts—in the same task, which means that the complexity of the task did not differ. What differed was the type of feedback—and possibly the underlying learning mechanisms that each type of feedback may have encouraged learners to use. By taking a measure of error detection ability and variability in productions, we observed that error detection and stability in productions were facilitating factors for those who received prompts but not for

those who received recasts. These results seem to uphold our assumptions that prompts encourage learners to analyze errors and take a logical approach to the correct target. Recasts, on the other hand, did not encourage this type of approach, and we assume that the alternative approach that learners adopted was learning through imitation, although the current results did not offer clear evidence on how the approach exactly operated. We therefore suspect that the effect of feedback frequency may be modulated by *how* learners learn—rather than by how complex the targets are. In the present study, frequent feedback seemed beneficial when learners employed the "error analysis" type of learning mechanism, while reduced feedback seemed beneficial to retention when learners used other mechanisms including imitation. Complex tasks are more likely to require learners to utilize error analysis, compared to simple tasks where the target movement is clear and straightforward, which may explain why previous studies often found an advantage of frequent feedback in complex tasks; for example, in Kim et al. (2012), the task was to imitate complex foreign sentences orally, which naturally encourages learning by imitation. When the task itself promotes learning by imitation, even if it is a complex task, reduced feedback can be more effective to retention as recasts in the current study showed. Future studies on feedback frequency may need further analysis on *how* people learn the particular task, not only *how well* they learn.

In the L2 learning field, the effectiveness of CF, especially prompts and recasts, has been of significant interest (e.g., Lyster, 2004; Ammar & Spada, 2006; Yang & Lyster, 2010; Gooch et al., 2016). It would be important to take into consideration that the effectiveness of prompts and recasts may change depending on frequency, because it varies across studies. As mentioned, laboratory experiments usually make an effort to ensure all the errors that occurred are followed by CF, whereas studies conducted in classrooms may not be able to do so because it is up to

teachers' discretion. The difference in frequency potentially causes discrepancies in results: for example, prompts may appear to be more effective in laboratory studies (due to higher frequency) than in classroom studies (due to lower frequency). Among many factors and conditions that may have an impact on the effectiveness of CF, frequency has been so obscure in L2 learning that studies may not necessarily include the actual frequency implemented in their experiment (e.g., Saito, 2013; Gooch et al., 2016). The current study demonstrated that it may have a differential impact on prompts and recasts in pronunciation learning, thus including feedback frequency in the methods or interpretation of the study may well be of importance.

The current study has multiple limitations. First, the experiment involved a computer-based laboratory training, which ruled out significant factors that could have impact on the effectiveness of CF, such as affective factors of human interaction (e.g., from whom learners receive feedback seems a significant factor in learners' perception of, as well as the effectiveness of, CF: Sato, 2013; Sippel & Jackson, 2015). In addition, due to the need to control the number of trials, the methods used in this study did not exactly replicate prompts and recasts in reality: learners were explicitly instructed to perform post-feedback trials, which is usually up to individual learners in practice. Learner's proficiency (Ammar & Spada, 2006) and the presence of explicit meta-linguistic explanation of the target (Sheen, 2007) also matters, all of which were limited or controlled in this study (we only used complete novices without giving explanation). The current experiment was also completely decontextualized and form-focused, which involved no meaningful content, which, some researchers argue, reflects a less effective learning environment (Lightbown, 2008). Last, but not least important, we exclusively used either prompt or recast in this study, thus it is still unclear how the effect of feedback would change if they are mixed, which, we assume, is a more realistic situation in practice.

Preface to Paper 2

Motor learning research has traditionally been focused on specific, objective conditions for learning, such as the frequency, amount, or schedule of feedback (Maas et al., 2008), and thus our first paper investigated one of the factors, frequency, with respect to corrective feedback in L2 speech learning. However, recent research in motor learning has begun to focus on psychological and cognitive factors as significant features that affect the effectiveness of training, such as learners' motivation, attention, and expectancies (Wulf & Lewthwaite, 2016). In this context, the effect of positive feedback has attracted attention both from a practical and a theoretical standpoint. We believe that the effect of positive (non-corrective) feedback is worth investigating in the context of L2 learning, especially because L2 research has traditionally been focused on corrective feedback. Like the first study, this investigation offers a new perspective to the discussion of the role of feedback, such as whether feedback without any error correction can be effective for learners' linguistic attainment and, if so, whether the primary role of feedback may be more than error correction.

Paper 2: Non-corrective and corrective feedback in L2 pronunciation training

1. Introduction

Feedback is a significant part of various learning processes, including second language (L2) learning. The most well-investigated type of feedback in L2 research is corrective feedback (CF), which is given by a teacher or an interlocutor in response to learners' erroneous output, with the purpose of making learners aware of their own errors and, if possible, inducing correct output. While a number of studies have investigated the effectiveness of CF on acquisition of oral L2 skills (Russell & Spada, 2006; Li, 2010; Lyster & Saito, 2010), "non-corrective" types of feedback (NCF)—that is, a response that is given to learners' successful output to acknowledge its correctness—have received substantially less attention. Ellis (2017) commented that while L2 teachers are more likely to discuss both CF and NCF, L2 researchers have so far focused "exclusively" on CF, which suggests that NCF is under-investigated and how it affects L2 learning is still unclear despite its common usage in practice.

The present study aims to investigate the effectiveness of NCF (also referred to as *positive* feedback[12]) compared to CF (also referred to as *negative* feedback) on L2 pronunciation learning. In the following sections, we will first review the previous literature on NCF in L2 learning; then we will turn to another related field, motor skill learning, to consider previous findings that might

[12] Some researchers (e.g., Södergård, 2008) distinguish non-corrective and positive feedback: while the former refers to an affectively neutral response to help continue interactions (such as "Okay"), the latter refers to a praise or verbal reward (such as "Good!" or "Excellent!"). In this study we will use the two terms interchangeably, and both refer to any kind of response that is given to learner's error-free output, which implicitly or explicitly informs the learner of its correctness with or without affectively positive connotations.

inspire predictions about the potential effectiveness of NCF.

1.1. Non-corrective and corrective feedback in L2 learning

NCF has so far attracted much less attention than CF in L2 research, probably because error correction has been the most controversial and thus 'hot' topic in the field. The discussion surrounding whether correcting learners' errors is essential for L2 acquisition is fundamentally related to one's view of language acquisition and the role of instruction, such that it has sparked different theories and pedagogical approaches over many decades (e.g., El Tatawy, 2002; Russell, 2009; Rezaei et al., 2011, for reviews). The early behaviorist view maintained that all errors should be corrected because overlooked errors could lead to wrong habit formation (Brooks, 1960), while supporters of the nativist view (Krashen, 1981) claimed that error correction had no beneficial effect because language acquisition can only occur when learners are engaged in meaningful language processing, which error correction could hinder. These early contradictory views eventually converged into more accommodating theories (e.g., the interaction hypothesis: Long, 1991) and teaching approaches (e.g., Focus on Form or Task Based Language Teaching : Ellis, 2009; Ellis, 2016), where learning is supposed to occur in meaningful interactions while error correction is also encouraged to draw learners' attention to linguistic form. In this context, NCF is often only associated with the "meaningful interaction" aspect, not with linguistic form—that is, NCF is viewed as a discourse phenomenon to facilitate interactions, not as a beneficial factor in the acquisition of the target linguistic form, while CF is considered to have such an effect (Russell & Spada, 2006; Li, 2010; Lyster & Saito, 2010).

A small number of recent studies has examined NCF in L2 learning, and most of them focused on how NCF functions in classroom discourse. For example, Lyster (1998) found that teachers used non-corrective repetitions (i.e., teachers repeat students' error-free utterance) as

frequently as, or even more frequently than, recasts (i.e., teachers repeat students' erroneous utterance in a corrected form), and both types of feedback are used for similar discourse functions—to keep students' attention on their current interaction. Waring (2008) argued that NCF as praise often prevented additional learning opportunities from arising, because this type of NCF may signal the end of the interaction, which makes learners feel that no further discussion is needed on the subject.

Other studies on NCF categorized types of NCF and observed how teachers chose among those types. For example, Ferreira et al. (2007) observed teachers' responses after learners' error-free utterances and found that about 30% were followed by teachers' non-corrective repetition, 10% by non-corrective rephrasing, another approximately 30% simply by a sign of acknowledgement or acceptance, while the remaining 30% did not receive any feedback (i.e., followed by another question or change of topic). Södergård (2008, p.167) observed a teacher's feedback strategy in Swedish kindergarten immersion and noted that non-corrective repetition was more frequent than CF, but the teacher refrained from using praise as NCF probably because she believed that "language and linguistic achievements should not have too much focus in immersion contexts". After observing both NCF and CF in an English as a foreign language class, Wasding (2013) noted that NCF showed less variety in types than CF; the NCF that occurred in his dataset fell in three categories—repetition, praise, or affirmation—while CF occurred in various types, such as prompt, recast, clarification, explicit correction, and so forth. Fagan (2014) analyzed the classroom discourse of a teacher and found that she used positive feedback systematically: when the focus of the current interaction was on the linguistic correctness of learners' output, she either invited peer assessment (e.g., asking the class: "Is that good?") or provided explicit positive feedback (e.g., "Exactly!" or "You are right."); when the focus was not on the correctness itself

but on other larger purposes, she utilized implicit positive feedback, including non-corrective repetition.

This review of previous literature implies that NCF has been mainly viewed as something related to classroom discourse, not as a tool that can directly contribute to L2 linguistic attainment. However, in other research domains, NCF—or positive feedback—has been shown to have a direct and beneficial effect on learning. One such domain is motor skill learning, where limb movements are more extensively investigated but speech has also received a good deal of attention. In fact, researchers have found that some of the findings in limb motor learning can be applied to speech learning (Maas et al, 2008; Bislick et al., 2012). Moreover, feedback is one of the most well-investigated topics in motor learning, such that some studies have already investigated and suggested the benefit of positive feedback in motor learning. Given this context, it is of interest to consider some of the studies in motor learning to seek insights on the effectiveness of NCF on L2 speech learning.

1.2. Positive and negative feedback in motor learning

Positive feedback has recently drawn attention in motor learning research because several recent studies found its advantage over negative feedback, contrary to the traditional view of feedback. Motor learning research has viewed learners' ability to detect and correct errors in their own movement as the crucial factor in learning, and considered feedback as 'guidance' to facilitate the ability (the guidance hypothesis: Salmoni et al., 1984). Feedback used in motor learning research is often neutral and continuous (e.g., how much your arm movement deviates from the desired trajectory) and not dichotomous as positive or negative, but the guidance view inherently implies that negative feedback is more effective than positive feedback, because negative feedback provides abundant information on errors while positive feedback does not.

Thus it was somewhat surprising when a series of studies found that, given control over when to receive feedback, learners preferred to receive feedback after relatively successful (or so they felt) trials rather than less successful trials (Chiviacowsky & Wulf, 2002; 2005) and furthermore, that learners who received feedback only regarding successful trials indeed learned better than those who received feedback on less successful trials (Chiviacowsky & Wulf, 2007; Badami et al., 2011; Saemi et al., 2012). These results did not conform with the above-mentioned view of feedback as guidance, and hence a new perspective, which puts more focus on the motivational effect of feedback, emerged (Wulf & Lewthwaite, 2016). Subsequent studies indeed found that positive feedback enhanced learners' motivation compared to negative feedback (Badami et al., 2011; Saemi et al., 2012). Motivation has long been recognized as an indirect factor in motor learning, such that learners with higher motivation are likely to practice harder or for a longer duration (Salmoni et al., 1984), but such effects were considered to be temporary. More recent studies, however, have suggested that motivation and related affective factors may have more direct effects on learning (Wulf & Lewthwaite, 2016; Sugawara et al., 2012). For example, when given social comparative feedback (i.e., learners are informed of whether their performance is above or below average among peers), participants who were informed that their performance was better than average (even if it was in fact false) performed better at retention (Lewthwaite & Wulf, 2010). Similarly, setting an easier goal for the task, which led learners to perceive themselves as more successful, had a beneficial effect on retention (Trempe et al., 2012). These studies suggest that higher motivation, including higher perceived competence, can facilitate learning. In this context, Wulf and Lewthwaite (2016) proposed a theory that emphasizes the role of motivation and attention in motor learning (i.e., the OPTIMAL theory), in which they argue that motivation can enhance learning through various aspects, especially increased neurological

activity related to retention; a rewarding experience, including positive feedback, would trigger dopamine release (e.g., Schultz, 2013), which enhances motor memory consolidation (Sugawara et al., 2012), which is crucial for long-term memory formation.

The above-mentioned studies have underlined the advantage of positive feedback, especially at retention, but several other studies failed to replicate it; these studies observed no difference in the effectiveness of positive and negative feedback (Patterson & Azizieh, 2012; Carter et al., 2016). More recently, some researchers have begun to explore a possibility that positive and negative feedback may have specific, differential effects on learning, rather than that one is more effective than the other. The advocates of this view argue that while positive feedback is associated with better long-term retention, negative feedback may increase the amount of attention directed to errors at the time of training (Zobe et al., 2019; Krause et al., 2018). In Zobe et al.'s (2019) experiment, participants who received negative comparative feedback improved their accuracy of the target movement but did not change in terms of their automaticity (measured by a dual task), while the opposite pattern—improved automaticity but little change in accuracy—was observed for participants in the positive comparative feedback group. Zobe et al. (2019) speculated that, believing that their performance was sub-par, participants with negative feedback may have invested a greater mount of effort to achieve higher accuracy and direct more attentional resources to errors during training, while participants who were given positive comparative feedback may not have had an incentive to do so since they believed that they were performing well enough. Positive feedback, in return, facilitated higher automaticity which was associated with dopamine-induced memory consolidation benefits because of the "rewarding" learning experience.

Taken together, these recent studies in motor learning has established that positive

feedback can be as effective as, if not more than, negative feedback, contrary to the traditional view where the primary role of feedback was considered to be error correction. Further studies are warranted to understand possible differential effects of positive and negative feedback, which may impact differently on learners' motivation, error awareness, and retention.

1.3. The present study

Given the above discussion, one of the primary questions is whether NCF (positive feedback) in the acquisition of L2 speech (more specifically, pronunciation of a phoneme) benefits learning, as suggested in motor learning research. If NCF indeed can be effective on accurate linguistic performance, it is most likely shown at a retention test where learners who received NCF retain accurate performance better than their CF counterparts, as it is suggested that the benefit of positive feedback lies in better memory consolidation (Wulf & Lewthwaite, 2016; Chiviacowsky & Wulf, 2007; Zobe et al., 2019). On the other hand, we would see the advantage of CF during training, as negative feedback may increase learners' effort and attention invested to the task (Zobe et al., 2019).

To further examine why NCF and CF follow (or do not follow) the above prediction, we also look at two additional measures: error detection ability and motivation. As discussed above, error detection ability is the ability that allows learners to notice their own mistakes or deviations from the target. This ability is traditionally considered as a key to enhance retention (Salmoni et al., 1984), since learners must rely on it when feedback is no longer present at a retention test. The guidance theory implies that negative or corrective feedback, which informs learners of errors, can facilitate this ability, while there is no theoretical reason to assume that positive feedback would do so. A recent study where negative feedback may have directed learners' attention to accuracy (Zobe et al., 2019) also predicts the facilitative effect of negative feedback in error detection, but

for a different reason; the guidance theory assumed that it was because of increased information about errors, while Zobe et al. (2019) hypothesized that it was due to increased attention. In this context, the present study is interested in whether CF in L2 speech learning enhances learners' ability to notice their own errors whereas NCF does not, and whether better error detection ability leads to better learning. On the other hand, recent studies (e.g., Wulf & Lewthwaite, 2016) argue that positive feedback increases learners' motivation, which plays a significant role in motor learning; thus it is also of interest to see whether NCF in L2 speech learning indeed increases learners' motivation and whether it is, in fact, associated with better learning outcomes.

Which types of NCF and CF to use in the current study is an important question, as there are various forms of feedback used in L2 learning, some of which may have differential effects (Lyster & Saito, 2010 for a review). As discussed above, CF is more variable and more well-investigated, thus we decided to choose the types of CF first and then use what we considered were the closest NCF counterparts to them. Sheen (2011) classified CF into two major types: CF that is intended to elicit the correct form from learners and CF that provides the correct form to learners. The elicitation-type of feedback includes prompting self-correction (by a request such as "Pardon?" or "Try again?"), whereas the provision-type of feedback includes recasts, where a teacher repeats the erroneous utterance produced by a learner but with the proper linguistic form, providing a correct model that the learner can imitate. Previous studies suggest that the two types of CF may have differential effects on both factors in which we are interested in the present study—learners' sensitivity to errors and motivation. First, considering the definition of the elicitation and provision type of feedback, it is not unreasonable to assume that the elicitation type of feedback may encourage learners' ability to detect errors in their productions, while the provision type may not. With regard to this assumption, Gooch et al. (2016) found potentially relevant evidence: learners

who received prompts during pronunciation training showed higher generalization ability (i.e., improved in terms of untrained items and a spontaneous task); recasts, on the other hand, seemed to facilitate more precise (target-like) productions for trained items. These results suggest that, in the context of the present study, elicitations may be beneficial to establishing learners' internal criteria about errors that were generalizable to untrained items or spontaneous situations. In addition, the two types of feedback have also been suspected to have differential impacts on affective aspects of learners' outcomes. Some teachers believe that the provision-type feedback may be less damaging to learners' motivation (Yoshida, 2008), partly because the elicitation type of feedback might be more difficult for learners to respond. A recent study indeed found that, specifically with regard to pronunciation training, instructions with CF generally reduced learners' anxiety for speaking in L2, but some of the elicitation-type feedback (clarification requests, specifically) seemed to increase learners' anxiety (Lee, 2016). Given these findings, we had sufficient reason to suspect that the NCF counterparts of prompts and recasts may also have differential effects on learners' error detection ability and/or motivation, and consequently, on overall performance, which led us to include both the elicitation-type and provision-type in the present investigation.

We then determined NCF counterparts to prompts and recasts, which are hereafter referred to as NC elicitations and NC repetitions, respectively (see Table 1). NC repetitions, the NC counterpart of recasts, simply refers to repeating a learner's successful production to acknowledge its correctness, and as reviewed above, NC repetitions are one of the most frequently observed types of NCF. On the other hand, a NC counterpart of prompts has rarely been mentioned in the previous L2 literature. In fact, a similar type of feedback has been sometimes referred to as "reinforcement" (e.g., Lyster & Ranta, 1997). For example, after providing CF and learners'

		Corrective/Non-corrective	
		NCF (positive)	CF (negative)
Type of feedback	Elicitation	**NC-elicitation**	**Prompt**
	Provision	**NC-repetition**	**Recast**

Table 1. The four feedback conditions included in the study.

successful self-repair, teachers may provide more opportunities for learners to practice the target by saying "Correct! Try again." or by repeating the same question (e.g., Lee & Lyster, 2016). This type of feedback can also be applied when not preceded by CF (i.e., question → correct answer → "Correct! Try again."), which is what we considered to be the closest counterpart to prompts. This type of NCF has seldom been the subject of research in the literature despite its apparent presence in practice. In this study, we refer this counterpart of prompts as NC elicitation.

Besides the above-mentioned four feedback conditions, the present study also included a control condition where participants practiced without any feedback. Thus, the primary analysis was performed in terms of comparisons between the control condition and each of the four feedback conditions to examine whether the NCFs and CFs were "effective", which was defined as inducing better learning outcomes compared to a no-feedback condition.

In sum, the objective of the present study was not to provide evidence for a general advantage of NCF over CF, but we believed that, despite the relative indifference toward NCF in the context of L2 research, NCF may have a beneficial effect on learners' linguistic attainment, not just on classroom discourse. Furthermore, we suspect that the effect of NCF may vary, depending on whether it is the elicitation-type or provision-type of NCF. From a practical standpoint, the present study may offer useful, if tentative, pedagogical suggestions with regard to NCF, which would encourage more effective use of NCF. From a theoretical standpoint, the present study challenges a traditional view of feedback in the L2 learning context. That is, as discussed above, feedback in L2 learning has been viewed as error correction, thus the finding that feedback with no error correction could be effective would warrant a different view as to the primary role of feedback.

2. Methods

2.1. Participants

Eighty-one participants were recruited in Montreal, Quebec, Canada from the community, mostly consisting of students and alumni of McGill University and Concordia University via online advertisements. Participants filled out a demographic questionnaire upon arrival (Table 2). The requirements for participants were identical to Study 1: aged between 18 and 40 (M = 21.2), having no prior knowledge of the target language (i.e., Japanese), and speaking English as their most dominant language. Many participants spoke other languages with varying proficiency (M = 1.75 languages besides English; the most common second language was French), but speakers or learners of Japanese or other languages that feature the geminate-singleton contrast were excluded from the study.

Participants were randomly divided into five groups, namely the prompt, NC elicitation, recast, NC repetition, and control group, resulting in fifteen to seventeen participants per group[13]. Note that the prompt and recast groups were the participants from Study 1 (labelled as the 100% prompt and 100% recast group, respectively) and that the control group was also the same as in Study 1 as well. Only the non-corrective elicitation and non-corrective repetition groups were newly recruited for the current study.

2.2. Target words and materials

The target non-words and materials used in the experiment were identical to our previous study (Study 1). A Japanese phoneme, the geminate plosive /tt/, was selected as a target and the

[13] One participant in the control group was excluded due to procedural failure.

	Prompt	NC-elicitation	Recast	NC-repetition	Control
N	16	17	16	17	15
Age	21.1	21.8	20.1	21.6	21.9
Gender (female : male)	9 : 7	12 : 4	15 : 1	15 : 2	12 : 3
Average N of languages	1.4	1.7	1.8	2.1	1.7

Table 2. Demographic information of participants. N = number of participants; Age = average age at the time of experiment; Average N of languages = the average number of languages that participants spoke besides English.

	singleton-singleton (SS)	geminate-singleton (SG)	singleton-geminate (GS)	geminate-geminate (GG)
trained item	atata ototo	attata ottoto	atatta ototto	attatta ottotto
untrained item	akaka okoko	akkaka okkoko	akakka okokko	akkakka okkokko
	akota otako	akotta okatto	attoka ottako	akkotta ottakko

Table 3. Target Japanese non-words. Eight trained items featured the /tt/-/t/ contrast. Untrained items featured either only the /kk-k/ contrast or both contrasts. Untrained items appeared in the pre- and post-test but not in the training phase.

eight non-words were constructed according to the following rules: the word consisted of three syllables, VCVCV; Vs were either all /a/s or all /o/s; the two Cs were either the target /tt/ or its singleton counterpart /t/, resulting in four word types, namely C1=geminate and C2=singleton (GS), C1=singleton and C2=geminate (SG), C1=geminate and C2=geminate (GG), and C1=singleton and C2=singleton (SS). Consequently, we constructed eight non-words: /attata, atatta, attatta, atata, ototo, ototto, ottoto, ottotto/. Another two set of eight non-words, the one which used /kk/-/k/ contrast instead of /tt/-/t/ and the one which used both /tt/-/t/ and /kk/-/k/, were also used in the pre- and post-test to see if the effect of training can be generalized to untrained words (See Table 3). Model stimuli were recorded using a digital field recorder (22050 Hz, 16-bits) by four native Japanese speakers (two female and two male speakers, all Tokyo dialect speakers), who were asked to produce the words clearly at natural speed. Each speaker recorded each word five times, but only one recording (usually the third one) per speaker per word was used in the experiment.

2.3. Procedures

After an exposure phase on Day 1, participants performed the pre-test and training on Day 2 (the next day), then the post-test on Day 3 (one week after Day 2). The exposure phase, pre-test, training, and post-test were performed using OpenSesame software (Mathôt et al., 2012). All procedures were identical to our previous study (Study 1), except for the feedback methods for the non-corrective elicitation and repetition group in the training phase.

Day 1: The exposure phase

A forced choice listening task was performed on Day 1. Participants listened to one word per trial from the eight target words, then chose one of four options (i.e. the options were either /attata, atatta, attatta, atata/ or /ototo, ototto, ottoto, ottotto/. Vowels were kept consistent in order

to have participants focus on distinguishing /tt/ from /t/). Participants were informed of whether their answer was correct on every trial. The task consisted of 160 trials (8 words * 4 speakers * 5 repetitions), which were completed in approximately 10 to 15 minutes. Please refer to Study 1 for further details.

Day 1: Motivational questionnaire (baseline)

The purpose of the baseline motivational questionnaire was to take into consideration the individual differences in motivation toward pronunciation prior to training. Participants' general motivation toward pronunciation exercises were measured using the Intrinsic Motivation Inventory (IMI: McAuley et al. 1989; Self-Determination Theory Research Group, n.d.). We used three subscales of IMI, namely Interest/Enjoyment, Perceived Competence, Tension/Pressure. In other words, in the present study participants were considered "motivated" when they enjoyed a task, felt competent about their own ability, and felt less pressured during the task. Participants answered twelve items on a one-to-seven Likert scale (one = not at all true and seven = very true) as to how they feel about pronunciation exercises in general (see Table 4), not as to the specific training in the present study that they were going to undergo (which was measured on Day 2, see blow). Note that, in the original subscale of "Tension/Pressure", lower scores on the Likert scale represented higher motivation (i.e., participants felt more relaxed), while higher scores were considered to represent higher motivation in the other two subscales. Thus, the scores for the "Tension/Pressure" was reversed (e.g., one was translated into seven) when the statistical analysis was conducted.

Day 2: Pre-test

The pre-test was a read aloud task where the 24 target words (including all three sets: /tt/-/t/ contrast, /kk-k/ contrast, and /tt/-/kk/ mixed) were randomly presented to participants. In addition, participants judged their own performance after every trial: they clicked either "Perfect!"

Interest/Enjoyment
I enjoy pronunciation exercises very much.
I think pronunciation exercises are boring activities. ®
I would describe pronunciation exercises as very interesting.
Pronunciation exercises do not hold my attention at all. ®
Perceived Competence
I think I am pretty good at pronunciation in foreign languages.
After practicing pronunciation for awhile, I feel pretty competent.
I am satisfied with my performance at pronunciation in foreign languages.
Pronunciation exercises are an activity that I can't do very well. ®
Tension/Pressure
I feel very tense while practicing pronunciation.
I am very relaxed in practicing pronunciation. ®
I do not feel nervous at all while practicing pronunciation in foreign languages. ®
I feel pressured while practicing pronunciation in foreign languages.

Table 4. Intrinsic Motivation Inventory (IMI), modified for the current study. The R after an item indicates a reverse item.

or “I can do better” button, depending on whether they thought their pronunciation was successful. Participants’ performance for each trial was also judged by the experimenter (the author), given either “pass” or “fail” with regard to their pronunciation of geminate/singleton contrast (for further details on the judgement criteria, please refer to Study 1). Participants were not informed of the experimenter’s judgement. The test consisted of 96 trials (24 words x 4 repetitions).

Day 2: Training phase

The training phase was a read-aloud task similar to the pre-test, but this time participants were provided a type of feedback that they were assigned to, except for the control group where participants continued to receive no feedback and simply read aloud the target words. Note that all participants were explicitly informed of how and when they were going to receive feedback before the training phase started. That is, participants the NCF groups (NC elicitation and NC repetition) knew that they would only receive feedback when they were successful, and those in the CF group (prompt and recast) were aware that feedback would occur only when there are some errors in their pronunciation.

For the prompt group and the recast group, feedback was given only when a production was judged as a “fail” by the experimenter, while participants simply proceeded to the next word when they were given a “pass”. For a prompt, a text (“Pardon? Try again!”) and the target word appeared on the screen along with a female character illustration with a frowning face (Figure 1, the upper left panel), while for a recast, a recording of a native speaker of the target word was presented via headphones along with the same illustration and the target word (Figure 1, the upper right panel). Both groups were explicitly instructed to produce the target again when they received feedback. In other words, they were expected to “correct” their mistakes at the second trial. There was no further feedback, whether or not they succeeded in the correction.

Figure 1. Examples of the screen presented to participants when they were provided with feedback. The upper panels were for the prompt and recast group respectively, while the lower panels for the non-corrective groups. The upper-right and lower-right panels (the recast and non-corrective repetition group) were presented along with a recording of a native speaker producing the target word.

For the non-corrective elicitation and the non-corrective repetition group, on the other hand, the condition was reversed: feedback was given only when participants were given a "pass" while nothing happened when they obtained a "fail". Participants in the non-corrective elicitation group, when they got a "pass", saw a prompt text saying "Great! Let me hear it again!" and the target word along with a female character illustration with a smiling face (Figure 1, the lower left panel). Those in the non-corrective repetition group were presented with the same illustration and the target word, while listening to a recording of the target word produced by a native speaker (Figure 1, the lower right panel). Identical to the recast and prompt group, these two groups were also explicitly instructed to produce the target word again following feedback, which means that they were expected to "repeat" their successful production. No further feedback was provided, regardless of their success at the repetition.

For each participant to have the same amount of practice, the number of total productions were kept consistent across participants by controlling the number of word presentation: a participant who received more feedback (consequently produced a word twice at a time more often) saw the word less often. Similarly, although participants were allowed to listen to model recordings (twice per word) at the beginning of each block to refresh their memory, the number of audio presentation was also kept consistent across participants, to avoid the situation where the recast and non-corrective repetition group would have more auditory input than other groups. The training phase consisted of 256 trials[14] (8 words x 4 repetition x 8 blocks), completed in approximately 20-25 minutes. Please refer to Study 1 for further details.

[14] In reality, participants may have performed a few extra trials depending on their situations in training. If a participant failed at the last trial of the target word, he/she had to pronounce it one extra time in the post-feedback trial. Since this "extra" trial did not occur when he/she succeeded at the last trial of the target word, whether he/she had extra trials was largely by chance. As a result, each participant on average pronounced one or two extra word. Thus the average number of productions in training in fact was 257 or 258 (depending on groups), instead of 256 (see Table 10).

Day 2: Motivational questionnaire (task-specific)

Participants filled out the Intrinsic Motivation Inventory for the second time immediately after the training phase to measure their perceived motivation toward this particular training that they had just finished. The twelve questions were changed from the baseline questionnaire accordingly to measure task-specific motivation; for example, the item "I enjoy pronunciation exercises very much" was changed into "I enjoyed this session very much." The scores were calculated in the same way as the baseline.

Day 3: Post-test

The post-test was the same read-aloud task as the pre-test. Participants' productions were judged both by the experimenter and the participant themselves. In addition, to measure participants' possible improvements in perceptual skills, participants also performed the same forced-choice task as the exposure phase, but with no feedback this time.

2.4. Analysis

2.4.1. Reliability of the native listener judgement

Since in the present study, learner's accuracy in pronunciation was measured by a Japanese listener's (i.e., the author) perceptual judgement, we first verified the reliability of the judgement before proceeding to the analysis of learners' accuracy. Specifically, we examined to what extent the acoustic correlates of the geminate-singleton contrast could predict the listener's judgement. The acoustic measures that have been shown to be the primary correlates of the contrast include closure duration of the plosives (from the offset of the preceding vowel to the burst of the subsequent plosive), the duration of the preceding and following vowels, fundamental frequency (F0) and intensity changes from the preceding to the following vowel (Idemaru & Guion, 2008; Kawahara, 2015; Hirata & Whiton, 2005). See Table 5 for details on the acoustic measures. All

Measurements	How to take measurements	Name of variables
Closure duration	From the offset of the preceding vowel to the burst of the subsequent plosive (Kawahara, 2015). To account for speech rate, raw values were then divided by the duration of the entire word (Amano & Hirata, 2010; Hirata & Whiton, 2005; Idemaru & Guion, 2010).	C1 C2
Closure duration ratio	The relative closure duration of C1 and C2: dividing C1 by C2.	C1/C2
Vowel duration	The duration of V1 and V2 (for V2, including voice onset time of the previous plosive). As with closure duration, raw values were then divided by the duration of the entire word to account for speech rates.	V1 V2
Intensity ratio	Extracted intensity for all the voiced portions of V1, V2 and V3 using Praat's autocorrelation method (Boersma, 1993), calculated the mean for each vowel, and then divided the mean of S1 by S2, and S2 by S3.	Int.V1/V2 Int.V2/V3
F0 ratio	Extracted F0 for all the voiced portions of V1, V2 and V3, and calculated the same way as the intensity ratio	F0 V1/V2 F0 V2/V3

Table 5. Acoustic measurements used for the analysis of reliability of Japanese listener's judgement.

measurements were taken using Praat (Version 5.4.19, Boersma & Weenink, 2020) and all statistical analyses (as well as in the following section) were performed in R (R Core Team, 2020) using the lme4 package (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017) for GLMMs.

We then built a Generalized Linear Mixed model (GLMM) using the listener's judgement (pass or fail) as the dependant variable and all the above-mentioned acoustic measurements as fixed effects, with a random intercept for participants. After building the full model, fixed effects that did not reach significance ($p>.05$) were removed using the backward elimination procedure. The current paper only presents the final model, where all the remaining fixed effects were significant. All productions that featured the eight targets with the /tt/-/t/ contrast were used in the analysis (i.e., the generalization targets, which featured the /kk/-/k/ contrast, were not included). After removing 162 productions due to difficulty in taking acoustic measurements (e.g., omission or insertion of syllables, devoiced vowels, or noises), 26,064 productions[15] in total were included in this analysis.

To calculate the agreement between the prediction from the models and the listener's judgement, we then obtained the log odds from the models, and labelled the productions with the log odds larger than zero as "predicted pass", otherwise "predicted fail". The agreement percentages (i.e., the percentages of productions whose predicted judgement and actual judgement are matched) and Cohen's kappa coefficients were calculated for each word type.

[15] The total number of 26,064 was calculated as follows: one participant was supposed to have 320 productions in total throughout the experiment (32 in the pre-test + 256 in the training phase + 32 in the post-test), which led us to the total number of 25,920 productions. As explained above, participants may produce a few extra words during training, which accounted for the remaining 144 productions (M = 1.8 extra words per participant).

2.4.2. Accuracy of participant production

Pre-post changes

The accuracy analyses were conducted in terms of two aspects: the pre-post changes and changes during the training phase. To examine participants' pre-post changes in accuracy, a GLMM was first built using 5184 productions (8 target words x 4 repetitions x 2 tests [pre and post] x 81 participants) that only featured the /t-tt/ contrast (i.e., trained items), with the listener's judgement as a binary dependant variable. The fixed effects of primary interest included Group (treatment-coded, with the control group as the reference), Test (the pre- and post-test), and the interaction between Group and Test (i.e., how well each group learned between the pre-and post-test). To control for possible covariates, we added some factors that have been suggested to have impacts on learning of a new foreign sound, which included: learners' age (Flege, 1999); the number of languages spoken (Onishi, 2016; Wrembel, 2010); listening score at the exposure phase (Hao & de Jong, 2016 for complex relationships between perception and production in L2 phoneme learning), baseline motivational scores and error detection scores at the pre-test. In addition, since the present study investigated the possibility that learners' error detection ability and motivation may affect learning, we also included the baseline motivation score and the error detection ability (d-prime, see below) at the pre-test to control for potential individual differences prior to training in the two measurements. In this model, as well as all the linear models described below, continuous variables were centered by subtracting the mean and then rescaled by dividing by the standard deviation unless described otherwise

After building the full model that included all the fixed effects above, a backward elimination process was used to obtain the final model. For the random structure, we chose to include as many random effects as theoretically possible (Barr et al, 2013): a random intercept by

participant and by word; a random slope for Test by participant and by word; a random slope for Word Type by participant.

Finally, to examine whether feedback methods affected learner's ability to generalize the learned geminate-singleton contrast to untrained items, the same analysis as described above was also performed regarding participants' productions that contained /k-kk/ contrast. A total of 10,368 productions (16 target words x 4 repetitions x 2 tests [pre and post] x 81 participants) were used in the analysis.

Changes during training

To examine how participants changed in accuracy over the course of the training phase, we only used normal trials (i.e., where learners produced the targets voluntarily without elicitation or model-presentation by feedback), since the accuracy of post-feedback trials may be largely subject to the types of feedback that the participant received (post-feedback trials after NCF would have much higher in accuracy than those after CF). 16129 productions in total were used in this analysis. A GLMM was built using the same fixed and random effects as described above for the pre-post model, with two modification: the factor Test was replaced by Block (there was eight blocks during training, treated as a continuous variable); to adjust for learners' individual differences in production ability prior to training, the pre-test score (calculated as percentages of "pass") and the interaction between Block and the pre-test score were added as fixed effects. After building the full model, the backward elimination procedure was again used to achieve the final model.

2.4.3. Error detection ability

Participants' ability to find errors in their own productions was measured by the agreement between participants' self-judgement and the Japanese listener's judgement. The

sensitivity index (d′) of signal detection theory (Green & Swets, 1966) was calculated for each participant for both the pre- and post-test separately. First, we obtained the number of trials in which the participant successfully detected errors (i.e., self-judged a trial as fail when the listener judged it as fail) and of trials in which the participants falsely detected a non-existent error (i.e., self-judged a trial as fail when the listener judged it as pass). A "hit rate" was defined as the percentage of correct detection in relation to all the trials that contained errors, and a "false alarm rate" as the percentage of detection of a non-existent error in relation to all the error-free trials. D-prime was then calculated according to the formula: $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$.

Once d′s were obtained, a Linear Mixed Model was fitted to investigate whether error detection ability improved between the pre- and post-tests. The fixed effects of primary interest included Group, Test (pre or post), and the interaction of the two (Group x Test). For the random structure, a random intercept by Participant was included.

2.4.4. Motivational score

To examine whether certain feedback methods used during the training phase motivated participants more than the no-feedback condition, an ordinal logistic regression (using the *ordinal* package in R) was performed with the seven-point Likert scale score as the ordinal dependant variable, using Group (with the control group as the reference), Questionnaire (baseline and task-specific, with the former as the reference level), and the interaction of the two (Group x Questionnaire) as fixed effects and an intercept by Participant and by Item (twelve items in each questionnaire) as random effects. After finishing analysis for all twelve items, we performed the same model building as described above but this time for each subscale (Enjoyment, Perceived Competence, and Pressure/Tension) to further understand which aspect of motivation was the most affected.

3. Results

3.1. Reliability of the native listener's judgement

The fixed effect estimates from the GLMM for each word type are shown in Table 6. The agreement rates between the predictions from these models and the actual listener judgements were 92.1% (Cohen's kappa coefficient (κ) = .837) for the GS targets, 95.3% (κ = .884) for the SG targets, 90.5% (κ = .804) for the GG targets, and 94.5 % (κ = .710) for the SS targets, resulting in 93.2% on average. The high agreement rates indicate that the listener's perceptual judgements were justifiable by objective (acoustic) data. In the following analyses, the listener's perceptual judgements were used as the index of pronunciation accuracy.

3.2. Accuracy of participant production

To provide an overview of each group's overall performance, Figure 2 shows the average raw percentages of "pass" that participants in each group received by the judge throughout the pre-test, each of the eight training blocks and the post-test. For the eight training blocks, only the data of normal trials were used for Figure 2 (post-feedback trials were excluded)[16]. We analyzed the data in two parts: the pre-post change and training performance. Subsequently, we look at error detection ability and motivation separately.

Pre-post changes

Figure 2 suggests that all groups appeared to do better in the post-test than in the pre-test, while the amount of change differed across groups. Estimates for the fixed effects from the final

[16] Since the number of normal trials varied across participants, the percentages were calculated by first calculating percentages of pass for each participant, and then for each group.

	GS					SG				
	β	Std. Err.	z value	p		β	Std. Err.	z value	p	
(Intercept)	-2.78	0.21	-12.94	0.000	*	2.88	0.25	-11.66	0.000	*
C1 duration	2.03	0.12	17.10	0.000	*	-3.86	0.20	-19.63	0.000	*
C2 duration	-3.34	0.17	-19.36	0.000	*	3.29	0.15	22.66	0.000	*
C1/C2	-0.30	0.05	-6.25	0.000	*					
V1 duration	-0.34	0.08	-4.45	0.000	*	-1.06	0.15	-7.09	0.000	*
V2 duration	-1.14	0.11	-10.43	0.000	*					
Int. V1/V2	0.45	0.08	5.83	0.000	*	-0.42	0.09	-4.67	0.000	*
Int. V2/V3										
F0 V1/V2										
F0 V2/V3										
	GG					SS				
	β	Std. Err.	z value	p		β	Std. Err.	z value	p	
(Intercept)	-3.70	0.27	-13.53	0.000	*	1.55	0.32	4.85	0.000	*
C1 duration	7.97	0.30	26.68	0.000	*	-0.83	0.24	-3.42	0.001	*
C2 duration	-4.26	0.23	-18.13	0.000	*	-3.80	0.26	-14.70	0.000	*
C1/C2	-19.27	0.82	-23.43	0.000	*	-4.24	0.47	-9.11	0.000	*
V1 duration						-0.91	0.12	-7.61	0.000	*
V2 duration	-0.59	0.09	-6.93	0.000	*	0.64	0.13	4.77	0.000	*
Int. V1/V2	0.62	0.08	7.77	0.000	*					
Int. V2/V3										
F0 V1/V2	-0.17	0.06	-2.87	0.004	*					
F0 V2/V3										

Table 6. The estimates for fixed effects from the final GLMMs with Japanese listener's perceptual judgement as the dependent variable and acoustic measurements as fixed effects. Blank rows indicate that those variables were eliminated during the backward elimination process.

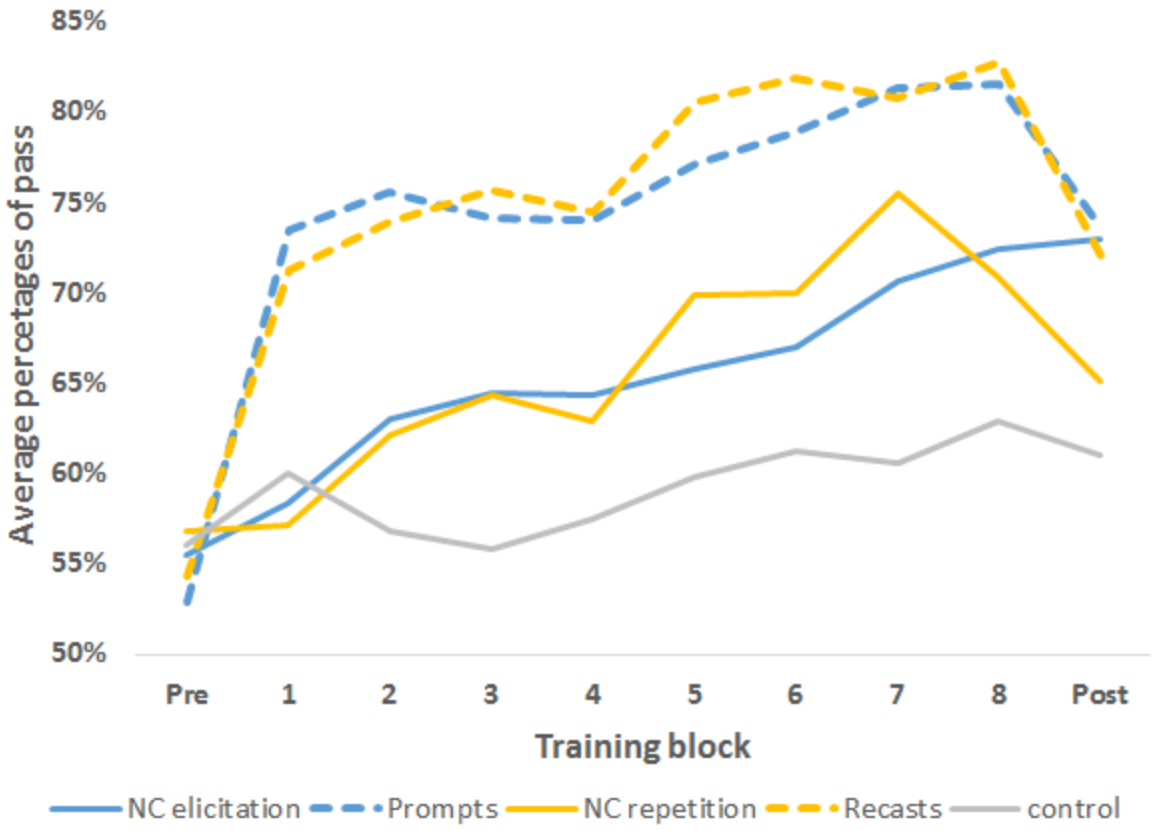

Figure 2. The average percentages of "pass" that participants in each group received in the pre-test, each of the eight blocks in the training phase, and the post-test. For the training phase, this figure only presents the data for normal trials. Due to the unbalanced number of normal trials in the training phase, percentages were first calculated for each word within participant, and then averaged for each participant, then for each group.

model are shown in Table 7. The effect of Test represents the change between the pre- and post-test for the reference level (the control group), which was significant with a positive estimate (β = 0.72, z = 1.99, p = .046), suggesting that participants could learn the target sound to some extent even without any feedback. This led us to the next question, that is, whether the improvements for each of the four feedback groups were over and above the improvements found for the control group. The four effects of Group x Test represent the differences in changes between the pre- and post-test for each group compared to the control group: three of the four, namely the NC elicitation (β = 1.12, z = 2.54, p = .011), prompt (β = 1.69, z = 3.63, p < .001), and recast groups (β = 1.01, z = 2.25, p = .024), significantly improved more between the pre- and post-test compared to the control group, while the NC repetition group did not (β = 0.03, z = 0.08, p = .938).

Other covariates were not the primary interest of the study, but the effects of Pre-Listening and Pre-Listening x Test were both significant with positive estimates, which suggests that higher initial listening scores predicted higher pre-test scores, as well as greater improvement between the pre- and post-test.

Finally, the estimates from the final model for production that featured the /kk/-/k/ contrast at the pre- and post-test are shown in Table 8. There were no significant group differences in the changes between the pre- and post-test (Group x Test effects), although all groups seemed to improve in the /kk/-/k/ contrast (because the Test effect was significant for the control group, and the β estimates for the Group x Test effects implied other groups improved more or less to the similar extent). This suggests that although participants were able to generalize their knowledge to untrained items, the effect of feedback was not observed in generalization.

Training performance

See Figure 2 again for the average raw percentages of ‘pass’ that participants in each group

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.71	0.75	0.95	0.342	
Test (for the control group)	0.72	0.36	1.99	0.046	*
Group (NC-elicitation vs. control)	-0.16	0.49	-0.32	0.751	
Group (Prompt vs. control)	-0.43	0.50	-0.86	0.390	
Group (NC-repetition vs. control)	-0.09	0.49	-0.18	0.856	
Group (Recast vs. control)	-0.29	0.51	-0.57	0.566	
Pre-Listening	1.16	0.18	6.44	0.000	*
Pre-Listening x Test	0.52	0.15	3.47	0.001	*
Pre-Error Detection	0.39	0.18	2.17	0.030	*
Group x Test (NC-elicitation vs. control)	1.12	0.44	2.54	0.011	*
Group x Test (Prompt vs. control)	1.69	0.47	3.63	0.000	*
Group x Test (NC-repetition vs. control)	0.03	0.44	0.08	0.938	
Group x Test (Recast vs. control)	1.01	0.45	2.25	0.024	*

Table 7. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the pre-post change in productions of trained items (/tt-t/ contrast).

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.33	0.52	0.63	0.532	
Test (for the control group)	1.00	0.27	3.66	0.000	*
Group (NC-elicitation vs. control)	-0.01	0.43	-0.03	0.979	
Group (Prompt vs. control)	0.02	0.43	0.05	0.962	
Group (NC-repetition vs. control)	-0.19	0.42	-0.46	0.646	
Group (Recast vs. control)	-0.12	0.45	-0.26	0.794	
Pre-Listening	1.18	0.14	8.55	0.000	*
Pre-Listening x Test	0.33	0.12	2.82	0.005	*
Group x Test (NC-elicitation vs. control)	0.56	0.36	1.53	0.126	
Group x Test (Prompt vs. control)	0.28	0.37	0.76	0.448	
Group x Test (NC-repetition vs. control)	-0.08	0.36	-0.23	0.821	
Group x Test (Recast vs. control)	0.54	0.37	1.46	0.145	

Table 8. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the pre-post change in productions of untrained items (/kk-k/ contrast).

received on normal trials throughout the eight training blocks. The results suggest that the percentage of 'pass' increased over the course of the eight training blocks for all groups, although it seems that some groups showed advantages as early as the first block of training while other groups displayed slow improvement over the course of training. The estimates for the fixed effects from the final model are presented in Table 9. Here the effects for the Group factor indicate the differences between each group as compared to the control group at the first block of training: both the prompt and recast groups already outperformed the control group (β = 0.59, z = 2.13, p = .033 for the prompt group and β = 0.82, z = 2.83, p = .005 for the recast group), while the NC elicitation and NC repetition group did not (p = .609 and p = .769 respectively).

In terms of the rate of change over the rest of the blocks, the Group x Block interaction indicates that the NC repetition group (β = 0.92, z = 2.57, p = .010) and recast group (β = 0.81, z = 2.26, p = .024) showed faster positive changes over blocks than the control group. The Group x Block interaction for the NC elicitation group also indicated that they improved faster than the controls over the blocks at a marginally significant level (β = 0.67, z = 1.89, p = .059). The prompt group, however, did not show a significant difference from the control group (β = 0.50, z = 1.42, p = .157). This suggests that the three groups (NC-elicitation, recast, NC-repetition) improved faster than the control group after the first block, while the rate for the prompt group was somewhat slower.

The other covariates were not the primary interest of analysis: the positive estimates for Pre-Test and Pre-Listening indicated that higher pre-test and listening scores predicted higher performance in the first block of training, and the positive estimate for Pre-Test x Block suggested better learning during training for participants who obtained higher scores in the pre-test. These results suggest that initial individual differences in both perception and production skills prior to

Fixed effects	**Estimate (β)**	**Std. Error**	**z value**	**p**	
Intercept	0.74	0.44	1.70	0.089	
Block (for the control group)	0.46	0.25	1.88	0.061	
Group (NC-elicitation vs. control)	0.14	0.28	0.51	0.609	
Group (Prompt vs. control)	0.59	0.28	2.13	0.033	*
Group (NC-repetition vs. control)	-0.08	0.28	-0.29	0.769	
Group (Recast vs. control)	0.82	0.29	2.83	0.005	*
Pre-Test Score	1.03	0.12	8.24	0.000	*
Pre-Test Score x Block	0.25	0.13	1.98	0.048	*
Pre-Listening	0.49	0.12	4.12	0.000	*
Group x Block (NC-elicitation vs. control)	0.67	0.35	1.89	0.059	.
Group x Block (Prompt vs. control)	0.50	0.35	1.42	0.157	
Group x Block (NC-repetition vs. control)	0.92	0.36	2.57	0.010	*
Group x Block (Recast vs. control)	0.81	0.36	2.26	0.024	*

Table 9. The fixed effect estimates from the final GLMM fitted to the listener's perceptual judgement regarding the change during the training phase (normal trials only).

	Normal trial		post-feedback trial	
	n	Percent of pass	n	Percent of pass
Control	256	59.3%	-	-
NC elic.	160	65.0%	99	93.1%
Prompt	213	77.4%	44	56.1%
NC rep.	160	65.9%	99	96.3%
Recast	215	78.1%	42	86.6%

Table 10. The number and average percentage of "pass" of normal and post-feedback trials during the training phase.

training had a significant impact on training performance.

Post-feedback trials

Refer to Table 10 again to see that the average percentages of 'pass' for post-feedback trials seemed higher for the NCF groups than the CF groups: 93.1% vs. 56.1% for NC elicitations vs. prompts; 96.3% vs. 86.6% for recasts vs. NC repetitions. This is not surprising because of the nature of NCF and CF: responding to NCF should be easier since participants were required to repeat their previous successful production in response to NCF, as opposed to CF which requires repairing errors. One thing to note was that post-feedback trials after prompts seemed particularly difficult (56.1%) compared to the other three feedback conditions.

3.3. Error detection ability

Figure 3 presents raw average d-prime scores for each group. Table 11 shows the estimates from the final model. The estimates for the Group x Block factor showed that the NC elicitation (β = 0.56, z = 2.41, p = .018) and prompt (β = 0.57, z = 2.40, p = .018) groups showed improvement between the pre- and post-tests compared to the pre-post change for the control group, which itself showed virtually no (if anything, negative) change (β = -0.15, z = -0.87, p = .388). On the other hand, the estimates for the NC repetition and recast groups indicated that there were minimal differences between each of the two groups and the control group, which suggests that participants in these two groups showed little improvement between the pre- and post-tests (β = 0.18, z = 0.76, p = .448 for NC-repetition and β = 0.01, z = 0.05, p = .961, for recast, respectively).

3.4. Motivation

Average ratings on a seven-point Likert scale from the baseline and task-specific motivation questionnaires (Figure 4) indicate that all groups tended to rate motivation as higher

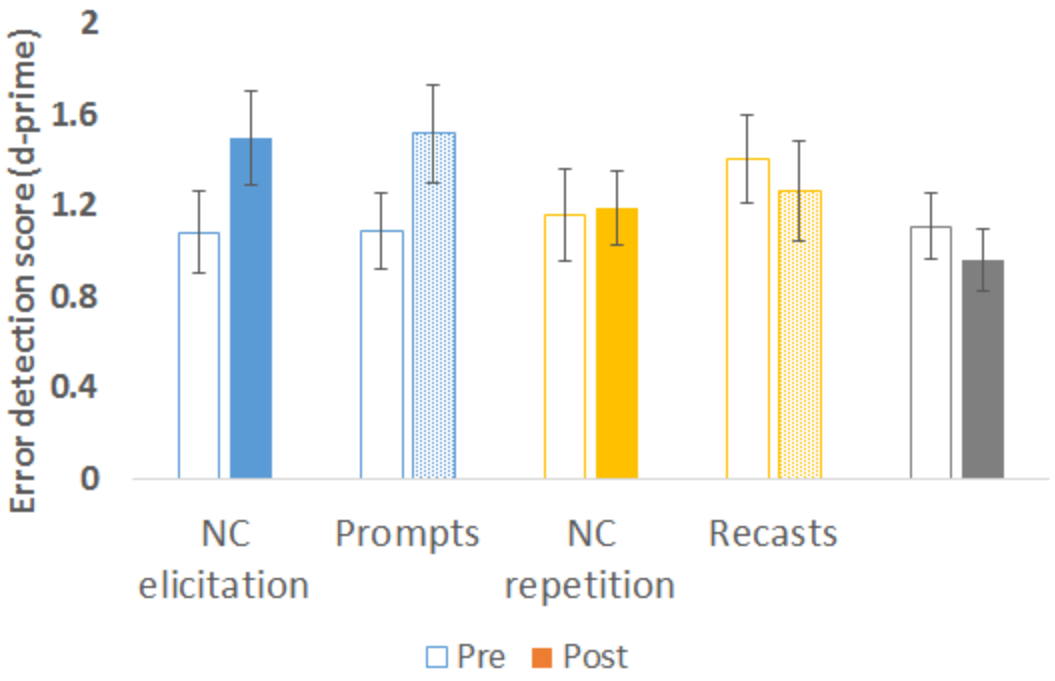

Figure 3. Average error detection ability (d-prime) for each group at the pre- and post-test. The error bars indicate the standard error of the mean.

Fixed effects	**Estimate (β)**	**Std. Error**	**t value**	**p**	
Intercept	1.11	0.19	5.73	0.000	*
Test	-0.15	0.17	-0.87	0.388	
Group (NC-elicitation vs. control)	-0.03	0.27	-0.10	0.920	
Group (Prompt vs. control)	-0.02	0.27	-0.07	0.943	
Group (NC-repetition vs. control)	0.05	0.27	0.19	0.850	
Group (Recast vs. control)	0.29	0.27	1.09	0.278	
Group x Test (NC-elicitation vs. control)	0.56	0.23	2.41	0.018	*
Group x Test (Prompt vs. control)	0.57	0.24	2.43	0.018	*
Group x Test (NC-repetition vs. control)	0.18	0.23	0.76	0.448	
Group x Test (Recast vs. control)	0.01	0.24	0.05	0.961	

Table 11. The fixed effect estimates from the LMM fitted to d-prime (error detection ability).

for the task-specific questionnaire than the baseline, but the extent differed across groups. The estimates for fixed effects from the model (Table 12) indicate that, first, the control group showed significantly higher motivation for this specific task than the baseline (β = 0.46, z = 2.50, p = .012). Second, the NC elicitation and recast group showed positive β estimates (β = 0.38 for NC elicitation and β = 0.38 for recast) for the Group x Questionnaire effect, which implies significantly higher task-specific motivation than the baseline for the two groups (because the differences were larger than the control group and the control group itself showed significantly higher task-specific motivation). Third, the negative β estimates for the Group x Questionnaire effect for the prompt and NC repetition group (β = -0.31 for prompt and β = -0.17 for NC repetition) only indicate that the differences between the baseline and task-specific motivation for these two groups were smaller than the control group, which suggested that we still needed to verify whether the differences for these two groups reached significance.

Given the above results, we performed post-hoc pairwise comparisons between the baseline and task-specific motivation scores for each of the five groups using the package *emmeans* (Russell Lenth, 2020) with Bonferroni correction for the alpha level (now α=.01). The results confirmed that the NC elicitation (β = 0.69, z = 4.72, p < .001) and recast groups (β = 0.70, z = 4.60, p < .001) were more motivated by the current training than they would normally be in pronunciation training, but the prompt (β = 0.12, z = 0.78, p = .439) and NC repetition groups (β = 0.23, z = 1.59, p = .112) were not. The control group now only showed a marginally significant increase (β = 0.38, z = 2.51, p = .012), which suggests that evidence for higher task-specific motivation in the control group was weaker than the NC-elicitation or the recast group.

To further illustrate in which aspect of motivation the difference was the most obvious, Figure 5 presents the difference in average scores between the baseline and task-specific

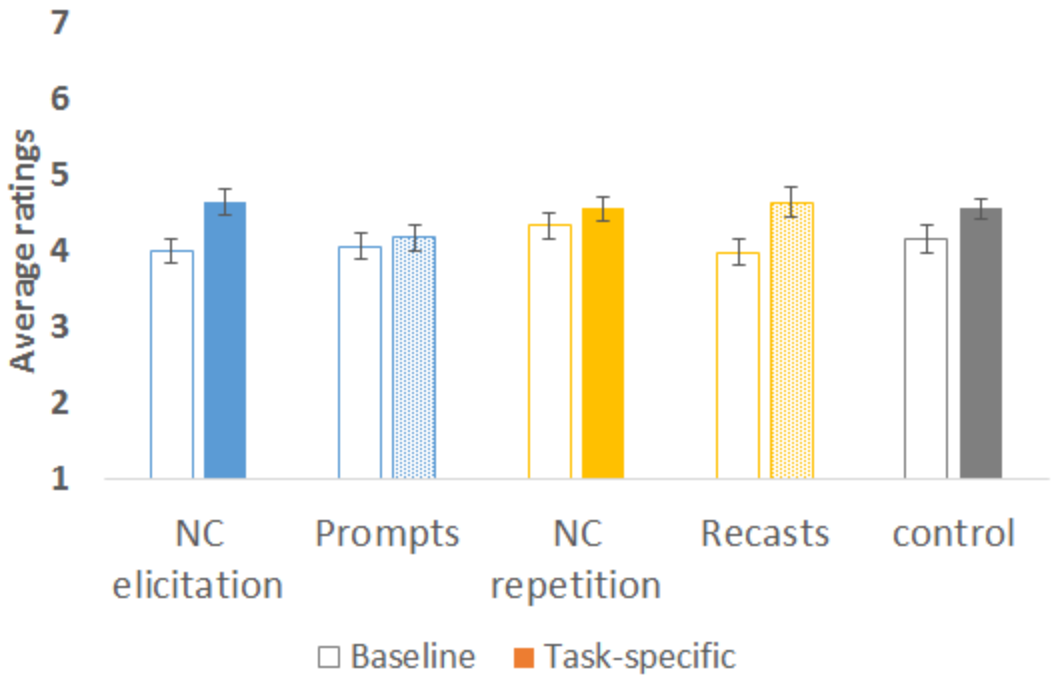

Figure 4. Average motivational scores (7-poing Likert scale) for each group at the pre- and post-test. The error bars indicate the standard error of the mean.

Fixed effects	**Estimate (β)**	**Std. Error**	**t value**	**p**	
Questionnaire	0.46	0.18	2.50	0.012	*
Group (NC-elicitation vs. control)	-0.17	0.28	-0.61	0.542	
Group (Prompt vs. control)	-0.10	0.28	-0.35	0.726	
Group (NC-repetition vs. control)	0.27	0.28	0.97	0.334	
Group (Recast vs. control)	-0.21	0.28	-0.74	0.458	
Group x Questionnaire (NC-eli. vs. control)	0.38	0.26	1.47	0.141	
Group x Questionnaire (Prompt vs. control)	-0.32	0.26	-1.23	0.218	
Group x Questionnaire (NC-rep. vs. control)	-0.18	0.26	-0.70	0.481	
Group x Questionnaire (Recast vs. control)	0.38	0.26	1.45	0.146	

Table 12. The fixed effect estimates from the ordinal logistic regression model fitted to Likert-scale cores in the motivational questionnaires (base and task-specific).

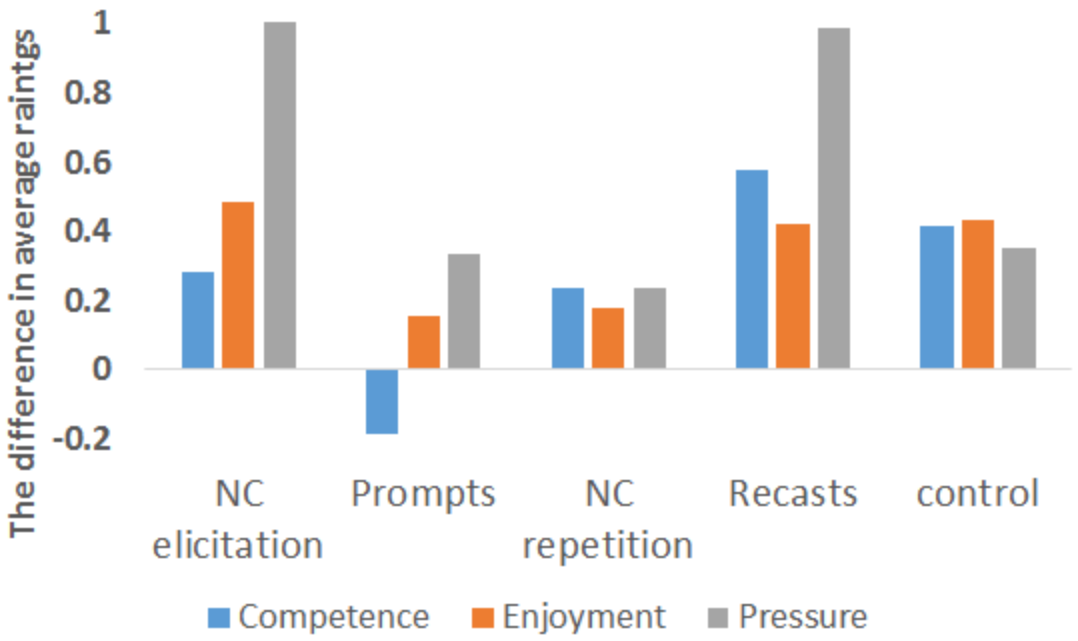

Figure 5. Differences in average scores between the baseline and task-specific motivation scores. Positive values indicate higher task-specific scores than the baseline. Note: for Pressure, higher scores mean that the participant felt less pressured.

motivation questionnaires, which are broken down into the three subscales, namely Enjoyment, Perceived Competence, and Pressure/Tension (positive values indicate higher scores for the task-specific motivation). When we conducted the same modeling and post-hoc comparisons as we did above but within each of the three subscales, the only significant differences found were in the Pressure/Tension subscale: the NC elicitation ($\beta = 1.18$, $z = 4.32$, $p < .001$) and recast groups ($\beta = 0.99$, $z = 3.48$, $p < .001$) showed significantly higher scores (i.e., participants felt less pressured) in the task-specific questionnaire than in the baseline.

4. Discussion

In the present study, we investigated the effects of non-corrective feedback (NCF) in comparison with corrective feedback (CF) on L2 phoneme pronunciation learning with regard to two different feedback types, the elicitation type (NC elicitations vs. prompts) and the provision type (NC repetitions vs. recasts). The existing literature suggested that NCF would facilitate retention and motivation, while CF would enhance training performance and error detection ability, but we expected that these effects may be different for the elicitation and provision types of feedback. In the following section, we will first discuss the results for each type respectively.

4.1. The effect of NCF and CF in the elicitation-type feedback

For the elicitation feedback type, the results followed the pattern stated above: during the training, the NC elicitation group started out slowly, with scores as low as the control group after the first training block, but then improved faster than the controls. In contrast, the prompt group already outperformed the control group at the first block. These results conform to previous findings showing that negative feedback may be associated with higher training performance (Zobe et al, 2019). On the other hand, the pre- and post-tests revealed that both the NC elicitation

and prompt groups improved to a greater extent than the controls. The NC elicitation group learned slower during training, but retained what they learned quite well, resulting in similar performance as the prompt group at the post-test (see Figure 2 to observe the smaller decrease for the NC elicitation group between the last block of training and the post-test), which is also in line with previous studies which suggested better retention for positive feedback (Wulf & Lewthwaite, 2016; Chiviacowsky & Wulf, 2007). The results for motivation were also consistent with the previous assumption that NCF may facilitate motivation: the NC elicitation group showed higher task-specific motivation, but the prompt group did not. What was surprising was the finding on error detection ability, where both the NC elicitation and prompt groups showed improvement between the pre- and post-tests, because NCF has not been considered to facilitate error detection ability, given the nature of NCF is not likely to raise learners' awareness to errors.

In the previous L2 learning literature, NCF has often been discussed with respect to its function in classroom discourse (Lyster, 1998; Waring, 2008; Fagan, 2014), rather than its effect on the acquisition of specific linguistic features, whereas the role of CF in specific aspects of language learning has been investigated extensively (Russell & Spada, 2006; Li, 2010; Lyster & Saito, 2010). The present study revealed that NCF, at least the elicitation-type, can be effective for L2 pronunciation learning with a different learning trajectory than CF; NCF is slower to exert an effect on training, but then yields a greater likelihood that learners will retain what was learned, while CF tends to induce quick improvement in training followed by substantial forgetting. The mechanisms behind the effects of positive feedback have been associated with increased dopaminergic activity in the brain that leads to increased motivation and benefits memory consolidation (Widmer et al., 2016), and that of negative feedback with a temporary increase in attentional control and possibly increased sensitivity to errors (Zobe et al., 2019). The present

study, however, found both higher motivation and increased error detection ability in the NC elicitation group. Therefore, whether the observed beneficial effect of NCF can be attributed to increased motivation and to better memory consolidation still requires further investigation.

The function of NCF does not involve error correction, but the current results showed that participants who received NC elicitations were still able to correct errors and attain higher accuracy in pronunciation through training. One alternative explanation, aside from increased motivation and memory consolidation, is that the absence of NCF on a given trial may convey similar information to the presence of CF. That is, if one's performance is not praised or acknowledged, it suggests that it was ill-formed, or at least not as good as other productions that had been praised. In other words, NCF in the current study may have been a subtle, implicit form of CF, indicating the presence of error by not praising instead of directly correcting learners, which explains why the NC elicitation group showed improvement in error detection ability. If NCF can be viewed as an implicit version of CF (to be clear, both NCF and CF in the present experiment were highly explicit due to the training setting and the participants were definitely aware of the presence of feedback, but NCF was "implicit" in the way it points to errors), the underlying reason for the effectiveness of NCF could be its similarity to CF, rather than its association with "rewarding experience" and increased dopaminergic activity; some researchers in the L2 field have argued that the effect of implicit CF is retained better while explicit CF appears more effective in the short term (Li, 2010; Lyster et al., 2013, for a review) because implicit CF may develop implicit knowledge (Li, 2010). Either way, the current results warrant more research on the effectiveness and the underlying mechanisms of NCF in the context of L2 learning.

4.2. The effect of NCF and CF in the provision-type feedback

On the other hand, the results for the provision type feedback did not follow the anticipated

pattern in most respects. First, the NC repetition group did not show better retention at the post-test. During training, they started out slowly and then improved better than the controls, as predicted, but at the time of the post-test, they failed to outperform the controls, which implies a substantial amount of forgetting. To highlight this point, recall that the NC *elicitation* group showed similar training performance but still outperformed the controls at the post-test. In addition, the results of the motivation questionnaires revealed a pattern inconsistent with the notion that NCF would facilitate motivation more than CF: the NC repetition group did not show higher task-specific motivation than at baseline, whereas the recast group did. Finally, neither the NC repetition nor recast group showed any improvement in error detection ability, as opposed to the notion that CF (recasts) may have a beneficial effect on error detection ability.

NC repetitions were the only feedback condition in the current study that failed to yield improvement relative to the no-feedback control group at the post-test. According to the previous literature, NC repetitions are one of the most commonly used types of NCF in practice (Ferreira et al., 2007; Södergård, 2008), yet our results did not support its beneficial effect on learning of the target phoneme. One of the reasons why NC repetitions did not facilitate retention may be related to motivation. NCF is intended to acknowledge or praise learners' successful performance, which is supposed to increase learners' motivation (i.e., enhance perceived competence and/or make the task more enjoyable and relaxing), but NC repetitions may have distracted learners from the fact that their performance was praised and instead made them focus on auditory models. Moreover, although the participants in the NC repetition group were explicitly briefed (and constantly reminded throughout the training phase) that they would receive feedback only when their performance was successful, it is still possible that participants may have confused NC repetitions with recasts. That is, the provision of auditory models by native speakers itself may have implied

that their productions could be further improved. If participants perceived as such, they in fact received a mild version of CF on the trials where they performed well. This may partially explain why the NC repetition group seemed less motivated than the recast group; the former may have been frustrated by getting 'corrected' after their most successful performance, while the latter group felt more relaxed because they gain access to correct auditory models after erroneous performance.

An alternative account besides motivation is that providing auditory models may have directly hindered learners' long-term memory. In theory, the provision of models was meant to be a further opportunity to establish learners' memory as to what the target phoneme sounded like. However, right after learners' output, learners' short-term memory should still store the sound that they successfully produced, and the subsequent exposure to an auditory model may have masked it and made it difficult for learners to transfer their own successful production into the long-term memory.

In sum, the provision of the correct model as NCF seemed to cancel out supposedly beneficial effects of NCF (increased motivation and better retention), which suggests that this form of NCF may not facilitate linguistic attainment at least in L2 pronunciation learning. Note that, as discussed in the Introduction, in classrooms, NC repetitions are often purposefully used in a context where the linguistic correctness is not the priority and its intended function is mostly related to discourse, as opposed to other types of NCF that can be used to focus on accuracy (Fagan, 2014; Södergård, 2008). This may imply that some teachers intuitively or anecdotally understand NC repetition's functions and prefer it in a certain context, and the current results provide empirical evidence for that preference.

4.3. Limitations of the present study

One of the limitations of the present study is that it was conducted in a highly form-focused context, and thus the results cannot directly be extended to other learning contexts. For example, the absence of NCF could be much less salient in meaning-oriented interactions, because topic continuation without NCF can potentially be perceived by learners as acknowledgement of linguistic correctness. In the present study, the sole focus of the training was the target phoneme, therefore the presence or absence of feedback was extremely salient by design (as it tends to be in laboratory or quasi-experimental contexts). As Lyster and Mori's (2006) Counterbalance Hypothesis posits, learners in a form-focused context tend to be more attentive to implicit feedback, thus it is possible that NCF—if it can be viewed as more subtle, implicit way to provide information about errors (as discussed as an alternative interpretation) worked in the current study because of the highly form-focused context. If that is the case, for NCF to be effective, one might need to set up a learning context properly so that learners can notice both the presence *and* absence of NCF. This is in fact not a specific issue for NCF, however. Pragmatic ambiguity that makes feedback less salient is also a challenge for CF (Lyster, 1998). The present study suggests that a certain type of NCF can be effective in a context where the saliency of feedback is ensured, specifically in a form-focused training setting, and future work is required to determine how to implement it in a different, more meaning-oriented learning setting.

5. Conclusion

In the present study, we investigated the effectiveness of NCF (and, as a comparison, CF) on acquisition of an L2 phoneme, to explore the possibility that NCF can be effective in this type of context. Findings revealed that the elicitation-type NCF indeed enhanced learning, especially yielding an advantage in retention. Not all NCF showed advantages, however; the provision type

of NCF, NC repetition, did not yield improvement compared to a no-feedback condition. These findings offer both practical and theoretical implications for future research on feedback in L2 speech learning. Practically, NC-elicitations, which have not been well recognized as a type of feedback currently, may be utilized more often because findings suggest that this type of feedback can be beneficial to pronunciation accuracy while also providing motivation. Future studies are needed to replicate this finding and yield additional insights about how to implement it in a more realistic learning environment. Theoretically, the current findings did not offer a conclusive view as to whether the primary role of feedback in L2 speech learning is error correction because the results could be attributed to either motivation or to other possible factors such as "subtle" error correction (NC-elicitation) or hindered long-term memory (NC-repetition). Nevertheless, the current study suggests that, at least, feedback that does not include "explicit" error correction can be effective, which challenges the traditional view of feedback solely as a tool for error correction. As the alternative view of feedback gains ground in the related discipline of motor learning (e.g., Wulf & Lewthwaite, 2016), the current study suggests a need for further studies on the role of feedback in L2 speech learning.

Preface to Paper 3

The prior two papers, in which we explored the effects of feedback in L2 pronunciation training in a new light, only utilized acoustic data as supporting evidence that verified the reliability of the listener's judgement. To keep our discussion focused on the effect of feedback, the papers also did not discuss the details of the target linguistic item: Japanese geminates. In other words, we have so far discussed *how* learners learn in pronunciation training but have not discussed *what* they were learning in the training. For example, the Japanese geminate-singleton contrast is primarily durational, but were learners in our study aware of that? What acoustic features of geminates did they learn? This may be of interest in our interpretation of the findings of the prior two papers, because the effect of feedback might be affected by the characteristics of the target sound (see general discussion).

To answer these questions, the third paper of the present thesis focuses on acoustic data gathered from the same dataset collected for the prior two papers. Because the primary objective of this paper was to investigate acoustic characteristics of native and non-native production of the Japanese geminate-singleton contrast, the present paper employed a cross-sectional approach (i.e., comparing English and Japanese speakers at one specific point in time) unlike the pre-post design of the prior two papers. The present paper not only offers complementary information to understand the pronunciation training that we conducted in the present thesis, but also contributes to further understanding of non-native production of the Japanese geminate-singleton contrast.

Paper 3: Non-native production of Japanese geminate and singleton in three syllable words by English speakers

1. Introduction

Native-like production of geminate consonants is one of the well-known difficulties that non-native speakers of Japanese (JP) face in the acquisition of oral skills. Geminates, the obstruent consonants that have longer duration than their singleton counterparts, are one of the few non-syllabic moraic elements in JP: they rhythmically count as one unit (mora) just as a syllable does, but they are not considered a syllable by themselves. Language teachers of Japanese often find non-native speakers' geminates "not long enough", while non-native speakers are also prone to produce singletons that are too long due to confusion about the contrast (Sukegawa, 1993).

Previous acoustic analyses of Japanese geminate-singleton contrasts demonstrated that constriction duration (i.e., duration of closure for stops and of friction for fricatives) was the primary acoustic cue that governed the contrast, stable and consistent even in the context of different speech rates, and particularly strong as a cue when measured relative to duration of the word (Hirata & Whiton, 2005; Amano & Hirata, 2010). (Hereafter we only refer to closure duration since the present and most previous studies focus on stop consonants.) JP native speakers' geminate-to-singleton ratio of closure duration ranges from approximately two to three (Kawahara, 2015, for a review), although the exact ratio varies across studies, depending on the type of consonant, adjacent vowel, speech rate, and method of measurement. JP native speakers

also utilize other secondary acoustic cues to differentiate the contrast, such as duration and intensity of the preceding and following vowel; compared to singletons, geminates have increased duration and intensity of the preceding vowel but decreased values of the following vowel (Idemaru & Guion, 2008). The geminate-singleton contrast also affects pitch structure, which is, in JP, primarily specified by the location of lexical accent. JP lexical accent is implemented by a sudden drop in fundamental frequency (F0) from the accented vowel to the following vowel, and geminates yield a larger such drop if they occur between the two relevant vowels (Kawahara, 2005; Idemaru & Guion, 2008).

Previous studies on non-native production of JP geminates report that non-native speakers can produce adequate closure duration distinctions for geminates and singletons, but that their closure durations are not as clearly distinguished compared to JP native speakers (Han, 1992; Harada, 2006; Grenon & White, 2008; Lee & Mok, 2016; Hirata, 2017). For non-native speakers, the geminate-to-singleton ratio of closure duration tends to be lower than two, due to both insufficient closure duration for geminates (Han, 1992) and excessive closure duration for singletons (Toda, 1994; Muraki & Nakaoka, 1990). On the other hand, some individuals may also show a tendency towards overshoot, which results in an exaggerated ratio of geminate-singleton closure duration (Toda, 2007).

Only a few studies have examined the use of secondary acoustic cues in non-native production of JP geminates. Among the secondary cues, vowel duration has been considered, with results revealing that non-native speakers tend not to show a distinctive, consistent pattern (Toda, 1994; Lee & Mok, 2016; Guillemot, 2018). Non-native speakers sometimes followed a similar pattern as L1 Japanese speakers (i.e., longer duration for the vowel preceding the geminate and shorter duration for the vowel following the geminate), but this pattern was

inconsistent across speakers (Lee & Mok, 2016). Toda (1994) found some individuals who lengthened the preceding vowel of geminates excessively, presumably in an attempt to control syllable duration by extending vowels instead of consonants. Non-durational secondary cues associated with JP geminates, such as intensity and F0 of the adjacent vowels, are rarely reported with respect to non-native speakers. One of the few studies that investigated non-native speakers' F0 and intensity with regard to geminates is that of Hirata and Kato (2018), who investigated whether those secondary cues affected non-native speakers' degrees of foreign accent. When the F0 structures of non-native speakers' productions were digitally replaced by those of native speakers, JP listeners' ratings for foreign accent were improved, whereas the replacement of the intensity structure did not affect the rating. These studies seem to suggest that non-native speakers' secondary acoustic cues contain some deviations from native norms, especially in terms of F0, although it is still not clear what exactly the deviations are and how they affect overall success of geminate production (Toda, 1994; Lee & Mok, 2016; Guillemot, 2018; Hirata & Kato, 2018).

The present study aims to further investigate characteristics of non-native production of JP geminates and singletons with two notable differences from prior investigations. First, the present study uses three-syllable instead of disyllabic words as the target words. The majority of previous studies investigated non-native production of geminates and singletons in disyllabic words (Han,1992; Harada, 2006; Grenon & White, 2008; Lee & Mok, 2016; Hirata, 2017), which was reasonable, given that previous acoustic analyses of JP geminates produced by native JP speakers were mainly conducted on disyllabic words (but see Hirata & Amano, 2012). Nevertheless, using three-syllable words makes it possible to shed light on an interesting aspect of non-native production: the effect of position. It is known that non-native speakers are more

susceptible to errors at one position in a word relative to another, even when they produce (or perceive) the same target linguistic elements (Bent et al., 2007). For example, studies have shown that the Japanese long-short vowel contrast is more difficult for learners to both produce and perceive at the word-final position than the word-initial position (Muroi, 1995: Minagawa et al., 2002). In some studies, the word-initial position was also easier for non-native speakers than the word-medial position (Oguma, 2001), while there was no difference between the two positions in other studies (Minagawa et al., 2002). Unlike the long-short vowel contrast, there are very few studies on the effect of position on the geminate-singleton contrast, which is important because the majority of real words that involve the contrast have more than two syllables. According to Amano and Kondo's (2003) JP vocabulary database, the number of three-syllable words that contain one geminate is three times greater than that of disyllabic words with one geminate. Furthermore, among those three-syllable words, the geminate occurs between the first and second vowel (or syllabic nasal) six times more often than between the second and third (Amano & Kondo, 2003). This unbalanced distribution of geminates makes the effect of position, should there be any, more significant for non-native speakers.

Second, we investigate non-native speakers' use of secondary cues for JP geminates such as duration, intensity and F0 of the preceding and following vowels, in addition to the primary cue of closure duration. Non-native speakers are known to use secondary acoustic cues differently than native speakers (Iverson et al., 2003; Schertz et al., 2015; Zhang, 2008); non-native speakers sometimes focus on secondary cues more than the primary cue, especially when the primary cue is not utilized phonemically in their L1 and is therefore hard to perceive and produce (e.g., JP speakers tend to focus on the frequency of the second formant, a secondary cue, instead of the third formant, the primary cue, to distinguish English (EN) /r/ and /l/; Iverson et

al., 2003). In other cases, non-native speakers overuse secondary cues because they represent primary cues to a similar contrast in their L1 (e.g., Korean speakers consistently distinguished EN voicing contrasts using F0, which is a significant cue in Korean voicing contrasts, (Schertz et al., 2015); Mandarin speakers produced EN stressed syllables with higher F0 than EN speakers (Zhang, 2008)). The present study recruited English (EN) speakers who were trained to produce JP geminates. Although segmental duration in EN is not phonemic, it does serve as one of the cues for several linguistic contrasts, including tense and lax vowels, voiced and voiceless fricatives, and lexical stress accent (Klatt, 1976). If duration is not a salient cue for EN speakers, it is possible that they may use other secondary cues to systematically distinguish the JP geminate-singleton contrast. The present study, therefore, aims to examine whether non-native speakers' use of secondary cues differ from native speakers, and if so, whether the difference is due to non-native speakers' systematic attempt to distinguish the contrast.

Finally, having examined the use of the primary and secondary acoustic cues, we were also interested in how these cues may affect perception of non-native production of geminates; we included a Japanese listener's perceptual judgements and examined how EN speakers' primary and secondary acoustic cues could predict the listener's perception. This additional analysis was to evaluate how significantly non-native speakers' diversions (if any) from native norms in primary and secondary cues affect the overall success (i.e., being perceived correctly) of geminate production.

2. Methods

2.1. Participants

One hundred and twelve[17] young adult EN speakers were recruited in Montreal, Canada, aged from 18 to 36 (M=21.7, SD=3.5). Participants filled out a language background questionnaire (see Table 1) to confirm that they had no experience in learning of or having been extensively exposed to Japanese, or other languages that feature the geminate-singleton contrast. Most participants spoke other languages besides EN, such as French, to varying degrees of proficiency, although they self-reported that their most dominant language was EN and they used it more than 90% of the time during daily life at the time of the study.

Eight JP speakers served as a reference group (aged from 34 to 65, M=44.2, SD=15.1), all of whom spoke Tokyo dialect and resided in Tokyo, Japan. All of them learned English in secondary school as a formal subject but spoke only in Japanese during daily life at the time of the study and had little experience abroad (except for one participant who had lived abroad for two years after reaching adulthood).

2.2. Materials

The target words were three-syllable nonwords that consisted of two consonants and three vowels (V1C1V2C2V3). The target consonants were the Japanese coronal stop geminate /tt/ or its singleton counterpart /t/, while the three vowels were either all /a/ or all /o/. The primary target nonwords involved both geminates and singletons, which resulted in one of two syllable types: geminate-singleton (GS: /attata/, /ottoto/) or singleton-geminate (SG: /atatta/,

[17] There were two more participants in the original project, who were excluded from this study due to the failure in sound recording for acoustic analysis.

	English speakers	Japanese speakers
N	112	8
Age (M)	21.7 (*SD* = 3.5)	44.2 (*SD* = 18.9)
Gender (female : male)	88 : 24	4 : 4
The number of languages spoken (besides L1)	1.8	1.0

Table 1. Summary of the demographic questionnaire for English and Japanese speakers.

/ototto/). We also used target nonwords that only involved either geminates or singletons, which included geminate-geminate (GG: /attatta/, /ottotto/) and singleton-singleton (SS: /atata/, /ototo/) nonwords. Note that although all four types were three-syllable nonwords, they had inherently different word lengths: both GS and SG words counted as four moras (i.e., three vowels and one geminate), GG as five moras (i.e., three vowels and two geminates) and SS as three moras (i.e., three vowels). All four types of target nonwords were included because we assumed that there may be differences when speakers must distinguish the geminate-singleton contrast in a target production or in separate targets, but the difference in mora count made it necessary to treat each type separately in some analyses (see Analysis below).

All nonwords had lexical accent on the second vowel (V2), which specifies the general pitch pattern for the three vowels; given V1, V2 and V3, the pitch pattern was expected to be Low-High-Low, due to the word-initial rise between V1 and V2 (which is considered as a prosodic phenomenon rather than a lexical accent: Labrune, 2012, pp.180) and the drop (i.e., the pitch accent itself) between V2 and V3. Intensity is not a strong, consistent correlate of Japanese pitch accent (Sugiyama, 2011), although accented vowels have been found to have higher intensity in some studies (e.g., Neustupný, 1966).

As part of a larger research project, there were also other nonwords that participants were presented with in the experiment, but the present study limited its analysis to these eight nonwords that featured the contrast of geminate /tt/ and singleton /t/; (for the complete list of the target nonwords, please refer to Studies 1 and 2).

2.3. Procedures

First, JP speakers recorded the eight target nonwords using a digital field recorder (sampling rate 44100 Hz with a 16-bit resolution) placed on a table approximately 10 cm from them. Speakers were instructed to produce the nonwords presented on the screen in front of them clearly at a moderate speech rate, with lexical accent on the second vowel. Each speaker recorded the nonwords five times each. Three hundred and twenty productions (8 speakers x 8 nonwords x 5 production) were used in the analysis.

To record EN speakers' productions, the following procedures were undertaken. Because EN participants had no prior experience in JP, they first performed a perception task to have exposure to the geminate-singleton contrast. EN speakers were informed that "tt" and "t" were different sounds in JP (no further explanation was provided). The model recordings (which were a subset of the JP speakers' recordings as described above) were presented at the beginning of the task, two times each for all eight nonwords, along with the written target presented on the screen. Participants then proceeded to the task, where they listened to the recordings one by one in randomized order, clicking one of the four options on the screen after each recording: the four options were the SS, SG, GS and GG nonwords (vowels were pre-set: if the recording was /atata/, all options were presented with /a/). Participants were provided with the correct answer after each trial. Participants performed a total of 160 trials in approximately 10 to 15 minutes.

As part of a larger research project that focused on the effect of different types of training (more precisely, feedback), on the next day participants proceeded to a training session, where they practiced the production of the target nonwords based on their assigned training conditions for approximately 30 minutes. The present study, however, is restricted to the data from the post-test, as described below, because our goal was to analyse the detailed acoustic characteristics of

non-native speakers' ultimate geminate-singleton productions, rather than the learning process. This is comparable to analyses conducted in previous studies (Han,1992; Harada, 2006; Grenon & White, 2008; Lee & Mok, 2016), which were cross-sectional and involved learners of Japanese with heterogeneous learning experiences. To see the complete procedure of the experiment including training conditions, and the effects of those conditions, please refer to Study 1 and Study 2.

Non-native productions analysed in the present study were from the post-test in the original studies, which was conducted one week after the training, with the same conditions for all EN speakers. The production task was a read-aloud task, where EN speakers were asked to read the targets presented on a computer screen one by one, with the order of targets randomized. For each production, a JP listener (the author) listened to and judged the production as either correct or incorrect, solely based on perception of the geminate-singleton contrast; (that is, any deviation in vowels was not judged as incorrect). EN participants were not informed of this judgement. The judgement was made with regard to each nonword, not to each consonant (i.e., to be judged as correct, both C1 and C2 needed to be heard as the intended ones). Note that the JP listener was aware that they were judging non-native productions and there were no native productions to serve as controls at the time of judgement; thus the judgement may have been more lenient than it would be under different conditions. Each participant produced 32 targets (8 nonwords x 4 repetitions), resulting in 3648 productions (32 productions x 114 participants) in total.

2.4. Analysis

Closure durations of C1 and C2 were measured using Praat (Version 5.4.19, Boersma & Weenink, 2020), as shown in Figure 1. The beginning of the closure was marked at the end of

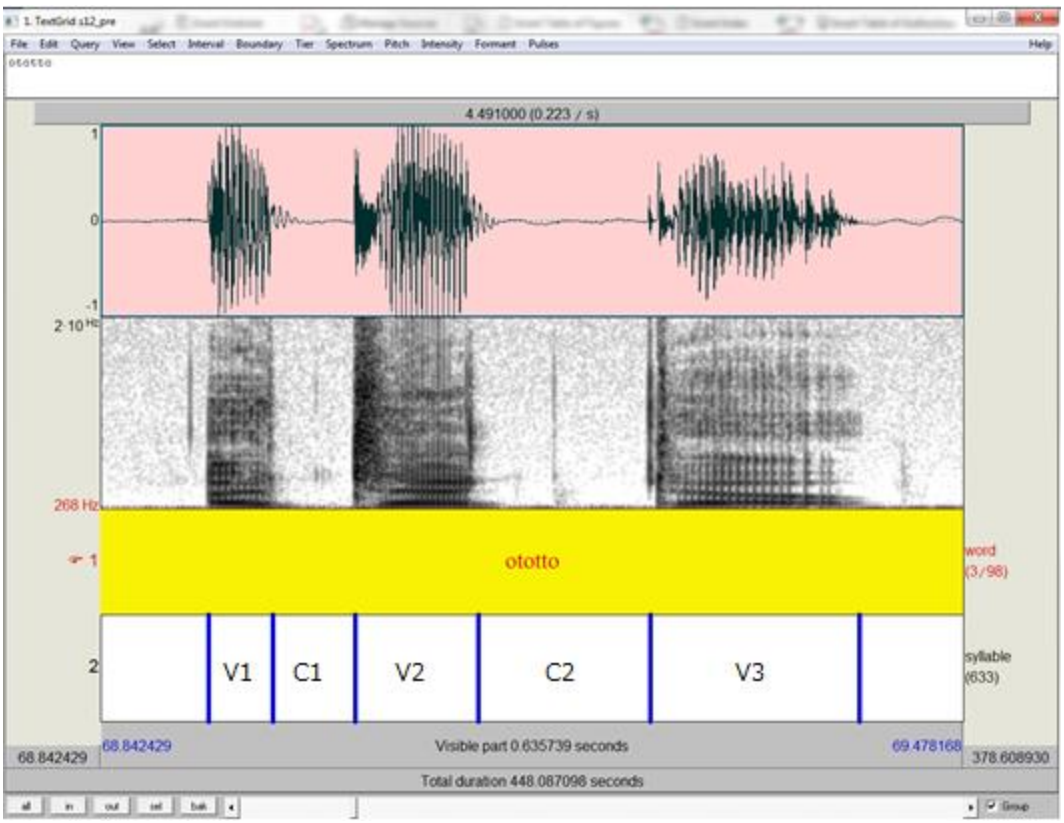

Figure 1. An example of acoustic measurements. C1 and C2 represent closure durations for /t/, while V1, V2 and V3 are the sum of each vowel and, where applicable, VOT.

the preceding vowel, which was determined by spectrograms and waveforms; when the visible offset of all formants was not clear, the periodicity of the waveform was considered as well. The end of the closure was marked at the onset of the release burst of C1 or C2. Closure duration was measured in milliseconds (ms) and then divided by the entire duration of the nonword (from the beginning of V1 to the end of V3) in which that particular consonant was embedded, to calculate the closure-to-word ratio (C/W). Although absolute closure duration is intuitively interpretable, the C/W has been shown to be the most reliable index to classify geminates and singletons in both production and perception by Japanese speakers, especially in the context of varying speech rates (Hirata & Whiton, 2005; Amano & Hirata, 2010; Hirata & Amano, 2012). Considering that non-native speakers tend to speak more slowly than native speakers (e.g.,Trofimovich & Baker, 2006), the present analyses used C/W duration as the primary index for closure duration, although we also present absolute duration where necessary to illustrate different outcomes from the two types of values. Note that when using C/W duration, the targets with different mora counts could not be directly compared to each other, since word length was inherently different (e.g., geminates in GG nonwords and singletons in SS nonwords could not be compared in C/W duration). Thus, in the analysis of closure duration, GS and SG nonwords (four mora words) were analysed together, while GG and SS nonwords were analysed separately.

The present study did not include voice onset time (VOT) as part of closure duration, since VOT is not considered as an acoustic correlate of Japanese geminates (Kawahara, 2015). Therefore, the duration of V2 and V3 indicated in Figure 1 is in fact the sum of the VOT and vowel duration (note that V1 did not have the VOT portion, since there was no word-initial consonant). Previous studies have shown that EN speakers of JP produced longer VOTs than JP speakers (e.g., Grenon & White, 2008), thus the results may need to be interpreted with caution.

Vowel (plus VOT) duration was also divided by word duration to account for different speech rates (vowel-to-word ratio, hereafter V/W). As with C/W, V/W could not be compared across targets with different word lengths, thus in the analysis of V/W we only used the four mora nonwords (i.e., GS and SG nonwords).

F0 of the voiced portion of V1, V2 and V3 were extracted using the autocorrelation method in Praat (Boersma, 1993), and we then calculated the mean for each segment and divided the mean of V1 by V2, and V2 by V3, to obtain the F0 difference between the vowels preceding and following C1 and C2. Intensity was also extracted in dB using Praat, then calculated in the same way as F0: first as the mean for each segment and then the ratio of V1 and V2, as well as V2 and V3. All types of target nonwords were used in these analyses.

EN speakers' productions were excluded from all analyses if they omitted any of the segments in the target or inserted extra segments, or if there were noises such as coughing or laughing or recording failures: 89 productions (2.4%) were excluded according to these criteria. In addition, there was a high occurrence of devoicing or creaky voice at V1 and V3 (especially V3), making the extraction of F0 impossible for some productions (417 productions, 11.4%), due to the low pitch induced by the location of lexical accent. We excluded these productions from the analyses that involved F0, while we kept them for other analyses where corresponding measurements were feasible. The last exclusion criterion was unusually slow speech rate: EN speakers' productions were excluded when word duration was more than 2.5 SD above the mean of word duration for the individual (17 productions, 0.4%). Among productions from JP speakers, 24 productions (7.5%) were excluded from the F0-related analyses due to the low pitch and difficulty in extracting F0.

All statistical analyses were performed in R (R Core Team, 2020). When linear models were built, all fixed effects were deviation coded (the first level coded as 1, the second -1; e.g., if the levels read "geminate or singleton", geminate was coded as 1, singleton as -1) unless described otherwise. Linear mixed effects models were built using the lme4 package (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017). Post-hoc multiple comparisons were performed using the *glht* function of the multcomp package (Hothorn et al., 2008).

3. Results

3.1. Listener's perceptual judgements

To provide an overview of how well EN speakers produced each word type (GS, SG, GG, or SS), Figure 2 presents the percentage of "correct" given by the JP listener for each nonword type across all EN speakers. A generalized linear mixed effects model with nonword type (treatment coded, with SS as the reference level) as a fixed effect and participants as a random intercept was fitted to the listener's binary judgement. Results showed that participants produced SS nonwords more accurately than all three other types (all $p < .001$), and post-hoc multiple pairwise comparisons with Tukey's HSD further confirmed that SG nonwords were more accurate than GS and GG nonwords (both $p < .001$), and that there was no significant difference between GS and GG nonwords ($p = .958$).

3.2. Primary cue: closure duration

Absolute and CW duration

Table 2 presents mean absolute closure duration (in ms) for geminates and singletons, the geminate-singleton ratio, nonword duration (in ms), and C/W duration (in %) for geminates and singletons for each nonword type across all speakers in each L1 group. Note that EN speakers'

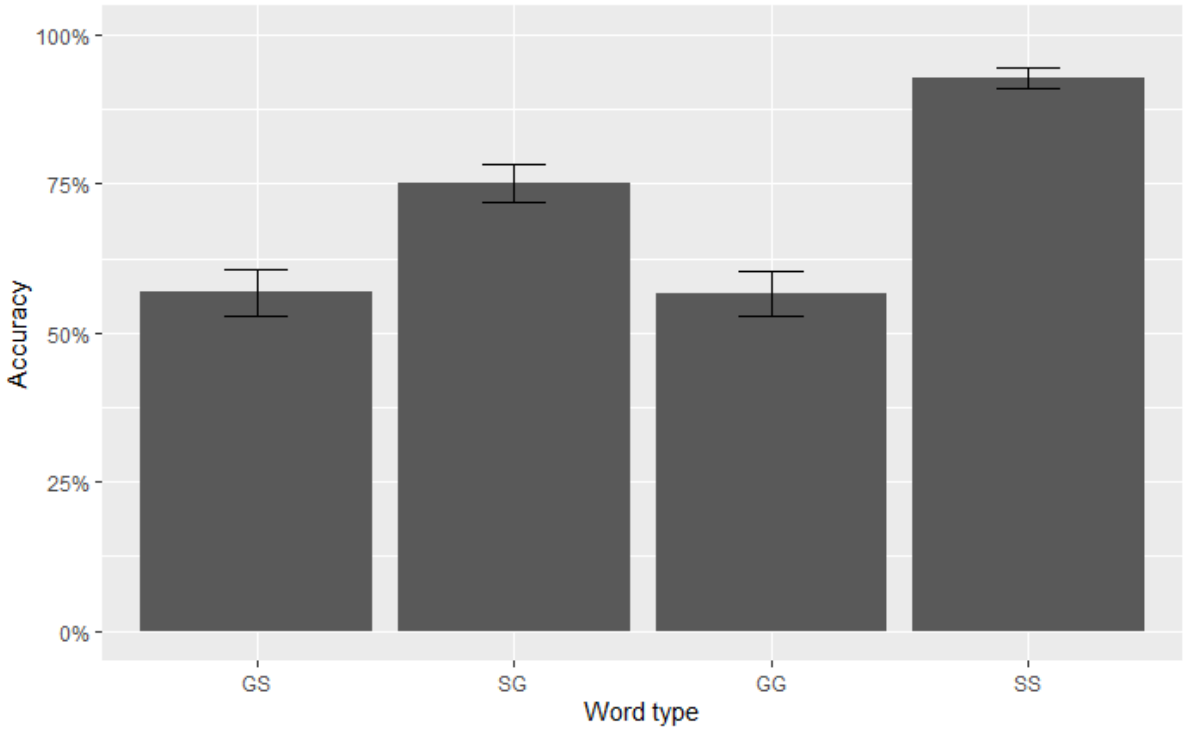

Figure 2. The percentages of "correct" given by the JP listener for each word type across all EN speakers. Error bars indicate standard error of the mean (SEM).

L1 group	wordtype	Geminate (ms)	Singleton (ms)	GS ratio	Word (ms)	CW Geminate (%)	CW singleton (%)
EN	GS	196	122	1.61	758	25.4	16.2
EN	SG	245	111	2.21	786	30.7	14.1
EN	GG	209	-	-	933	22	-
EN	SS	-	88	-	541	-	16.2
L1 group	**wordtype**	**Geminate (ms)**	**Singleton (ms)**	**GS ratio**	**Word (ms)**	**CW Geminate (%)**	**CW singleton (%)**
JP	GS	226	100	2.26	688	32.8	14.5
JP	SG	231	91	2.53	649	35.7	14.1
JP	GG	220	-	-	818	27	-
JP	SS	-	96	-	526	-	18.3

Table 2. Mean absolute closure duration (in ms) for geminates and singletons, the geminate-singleton ratio, word duration (in ms), and C/W duration (in %) for geminates and singletons for each word type across all speakers in each L1 group.

nonword durations were longer than JP speakers' for all nonword types (i.e., slower speech rate) and that the differences in nonword duration were even larger in nonwords that featured one or two geminates (i.e., GS, SG, and GG nonwords) than in nonwords that only contained singletons (i.e., SS nonwords). This made EN speakers' closure duration for geminates appear longer in absolute duration than in C/W duration: for instance, in SG and GG nonwords, although EN speakers' geminates appeared comparable with (or even longer than) JP speakers' in absolute duration, that was in fact not the case in terms of relative C/W duration. In the following sections, we use C/W duration to examine each word type separately.

GS and SG nonwords

Figure 3 presents closure duration for geminates and singletons in GS and SG words produced by each L1 group. A linear mixed model with positions (GS or SG), consonants (geminate or singleton), L1 groups (EN or JP) and all the possible interactions of the three as fixed effects, and participants as a random intercept, was fitted to C/W duration. Results (Table 3) showed that a three-way interaction between positions, consonants, and L1 groups was significant (β = -0.53, *SE* = 0.24, *t* = -2.17, *p* = .029), as well as a two-way interaction between positions and consonants (β = -1.32, *SE* = 0.24, *t* = -5.47, *p* < .001). This suggests that, in general, the distinction between geminate and singleton was less clear when geminates occurred at C1 (GS words) compared to when geminates occurred at C2 (SG words), and this effect was even stronger for EN speakers. Post-hoc multiple comparisons with Tukey HSD tests further revealed that EN speakers' geminates at C1 were significantly shorter than their own geminates at C2 (*z* = 13.66, *p* < .001); EN speakers' geminates at C1 were also shorter than JP speakers' geminates both at C1 (*z* = 3.92, *p* = .002) and C2 (*z* = 5.40, *p* < .001), while EN speakers' geminates at C2 were not significantly different from JP speakers' geminates at either C1 (*z* = 1.11, *p* = 1.00) or C2 (*z* = 2.60, *p* = .261). The difference between JP speakers' geminates at C1

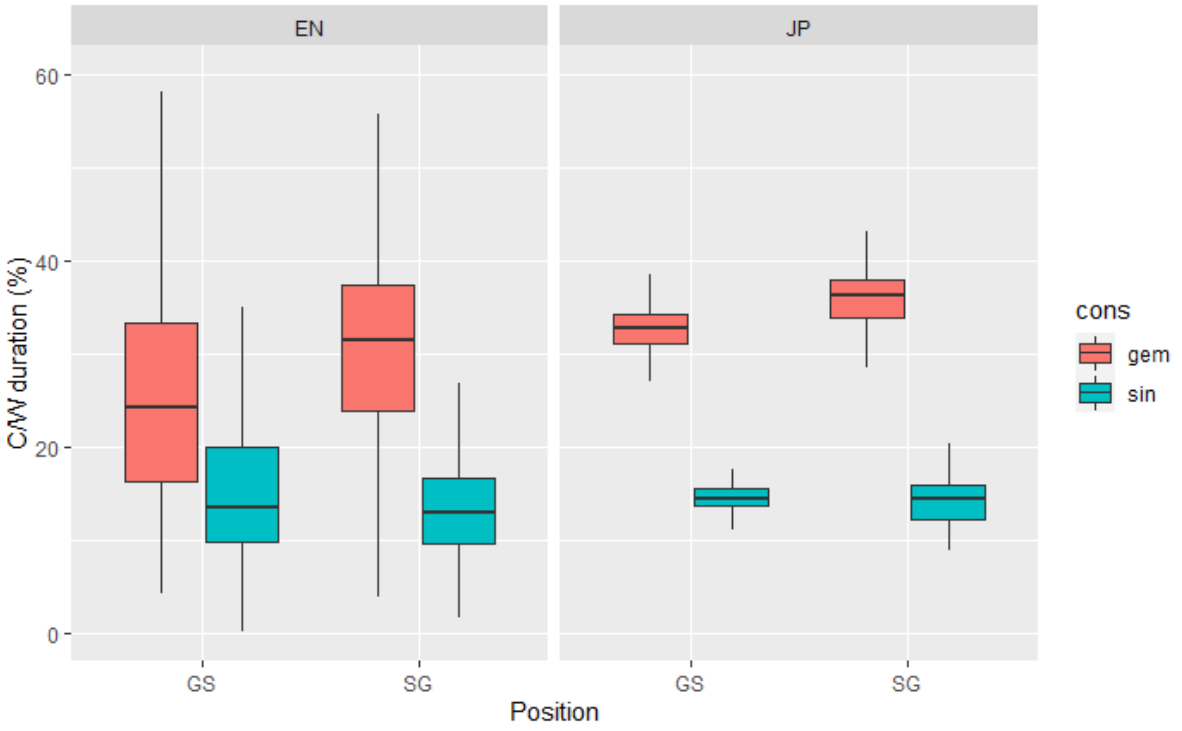

Figure 3. C/W (closure to word ratio) duration for geminates and singletons in GS and SG words by each L1 group. The lines inside the box indicate median. Outliers (outside 1.5±IQR from 25% and 75% percentile) are not shown.

Fixed effects	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	22.95	0.85	114.22	26.98	0.000	***
Position	-0.71	0.24	3728.42	-2.96	0.003	**
Consonant	8.22	0.24	3727.96	34.04	0.000	***
L1 group	-1.33	0.85	114.22	-1.57	0.120	
Position x Consonant	-1.32	0.24	3727.96	-5.47	0.000	***
Position x L1 group	-0.10	0.24	3728.42	-0.41	0.682	
Consonant x L1 group	-1.75	0.24	3727.96	-7.26	0.000	***
Position x Consonant x L1 group	-0.52	0.24	3727.96	-2.17	0.030	*

Table 3. The model estimates for fixed effects from a linear mixed model with closure (C/W) duration as the dependent variable, positions (GS or SG), consonants (geminate or singleton), L1 groups (EN or JP), and all the possible interactions of the three as fixed effects. The asterisks indicate: ***= p<.001, **= p<.01, and *= p< .05.

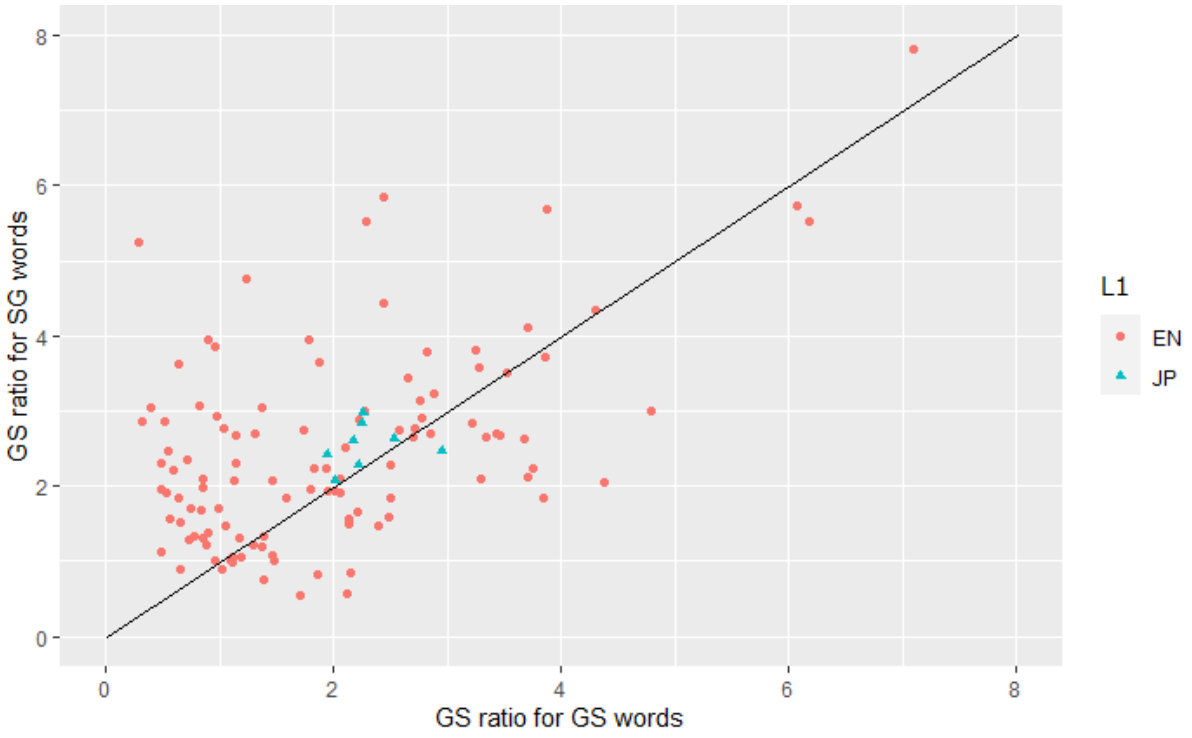

Figure 4. Average geminate-singleton closure ratio for each participant for GS and SG words. The diagonal line indicates the same ratio for both word types: individuals on the upper side of the line showed a larger ratio for SG words, while those on the lower side of the line showed a larger ratio for GS words.

and C2 did not reach significance ($z = 2.15$, $p = .876$). As for singletons, there were no significant differences between positions or between L1 groups (all $p > .05$) except for one pair: EN speakers' singletons at C2 (GS words) were longer than those at C1 (SG words) ($z = -5.30$, $p < .001$). In sum, multiple comparisons confirmed that the EN speakers' geminate-singleton distinction was less clear in GS words, both due to insufficient duration for geminates and excessive duration for singletons, whereas that was not the case for JP speakers.

To further illustrate how individual speakers' productions were affected by position, we calculated the average geminate-singleton closure ratio (using absolute closure duration) for each participant separately for GS and SG nonwords (Figure 4). It was observed that GS nonwords were clearly more difficult than SG nonwords for some individuals: there was a group of individuals who could not produce a sufficient germinate-singleton ratio distinction for GS nonwords (approx. below 1.5) but could produce a larger ratio for SG nonwords (i.e., the distribution was disproportionately concentrated on the upper side of the diagonal line that indicates the same ratio for both word types). On the other hand, when individuals could produce a sufficient ratio for GS nonwords (approx. above 2), the position effect seemed to largely reflect individual differences: that is, the distribution was somewhat balanced at both sides of the diagonal line, which means that some individuals showed clearer distinction for GS nonwords and some for SG nonwords.

GG nonwords

Figure 5 presents closure duration for geminates at C1 and C2 in GG nonwords produced by each L1 group. A linear mixed effects model with position (C1 or C2), L1 group (EN or JP), a two-way interaction of the two as fixed effects, and participants as a random intercept, was fitted to C/W duration ratios. Results showed that there was no significant interaction effect of position

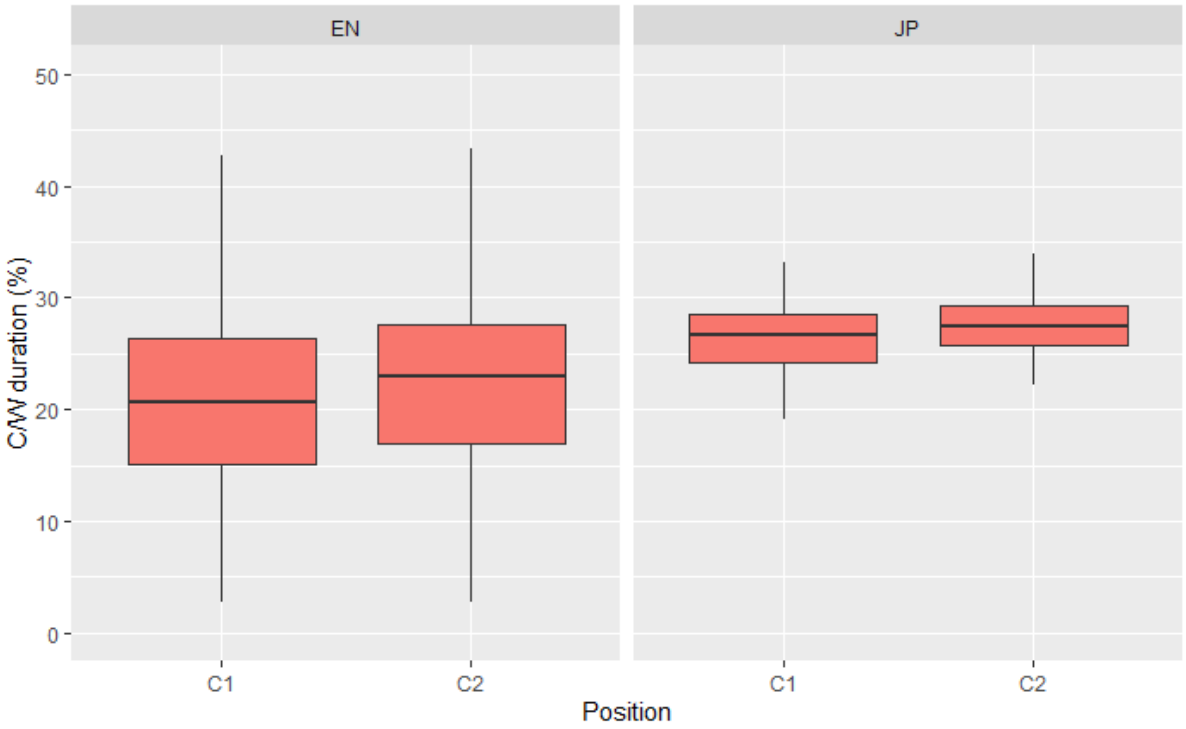

Figure 5. C/W (closure to word ratio) duration for geminates at C1 and C2 in GG words by each L1 group. The lines inside the box indicate median. Outliers (outside 1.5±IQR from 25% and 75% percentile) are not shown.

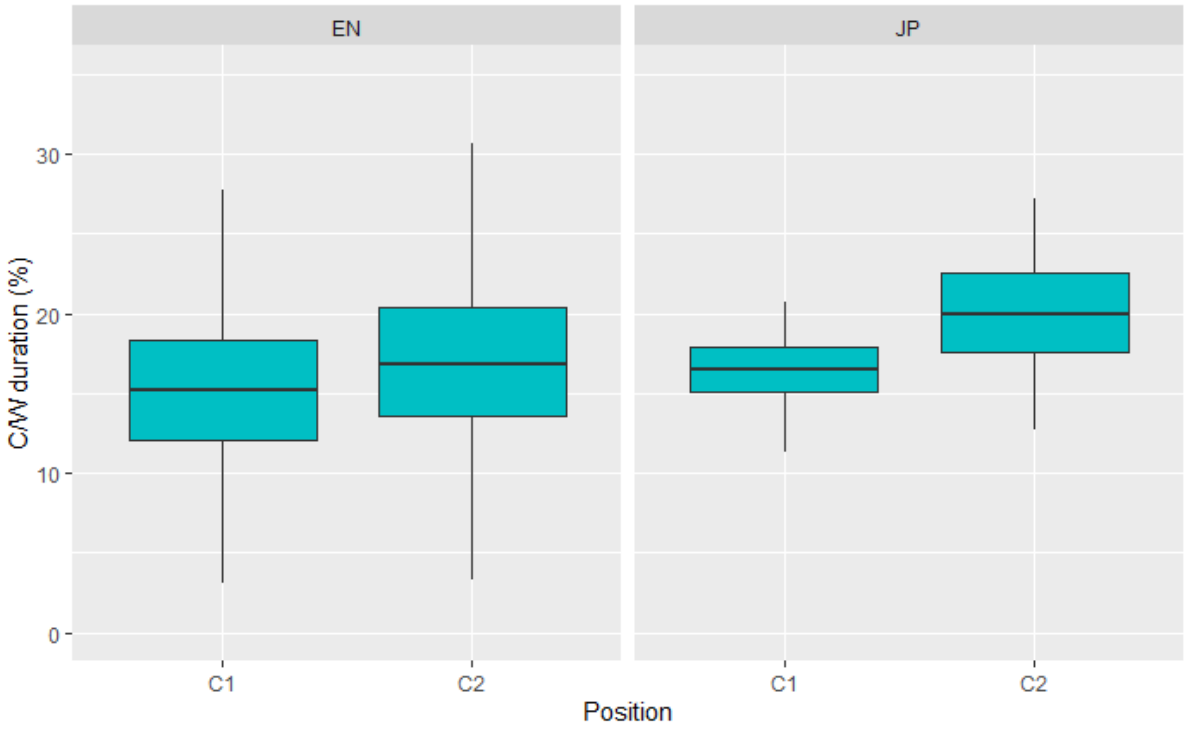

Figure 6. C/W (closure to word ratio) duration for singletons at C1 and C2 in SS words by each L1 group. The lines inside the box indicate median. Outliers (outside 1.5±IQR from 25% and 75% percentile) are not shown.

and L1 group (β = -0.19, *SE* = 0.23, *t* = -0.81, *p* = .415). Geminates at C1 tended to be shorter than those at C2 (the main effect of positions: β = -0.72, *SE* = 0.23, *t* = -3.13, *p* = .002) and EN speakers' geminates were generally shorter than JP speakers (the main effect of L1 groups: β = -2.50, *SE* = 1.03, *t* = -2.42, *p* = .017).

SS nonwords

Figure 6 shows closure duration for singletons at C1 and C2 in SS nonwords produced by each L1 group. The same linear mixed effects model as in the previous section (GG nonwords) was fitted to C/W duration. Results showed that the interaction of position and L1 group (β = 0.58, *SE* = 0.13, *t* = 4.33, *p* < .001) was significant, and that the main effect of position was significant (β = -1.33, *SE* = 0.13, *t* = -9.95, *p* < .001), but that of L1 group was not (β = -1.05, *SE* = 0.73, *t* = -1.44, *p* = .151). Post-hoc multiple comparisons with Tukey HSD tests further showed that the difference between C1 and C2 reached significance for both EN speakers (*z* = 9.92, *p* < .001) and JP speakers (*z* = 7.45, *p* < .001). This suggests that while speakers in both L1 groups showed an effect of position (i.e., C1 was shorter than C2), the effect of position was larger for JP speakers.

3.3. Secondary cues: vowel duration

Preceding vowels

Figure 7 (upper panel) shows the means of V/W duration ratios for vowels that preceded either geminates or singletons in GS and SG nonwords for each L1 group. A linear mixed effects model with consonant, L1 group, an interaction effect of the two as fixed effects and participants as a random intercept was fitted. Results showed that vowels that preceded geminates were longer than those that preceded singletons (the main effect of consonant: β = 2.60, *SE* = 0.17, *t* = 15.19, *p* < .001), while the main effect of L1 group (β = 0.65, *SE* = 0.57, *t* = 1.13, *p* = .261) and

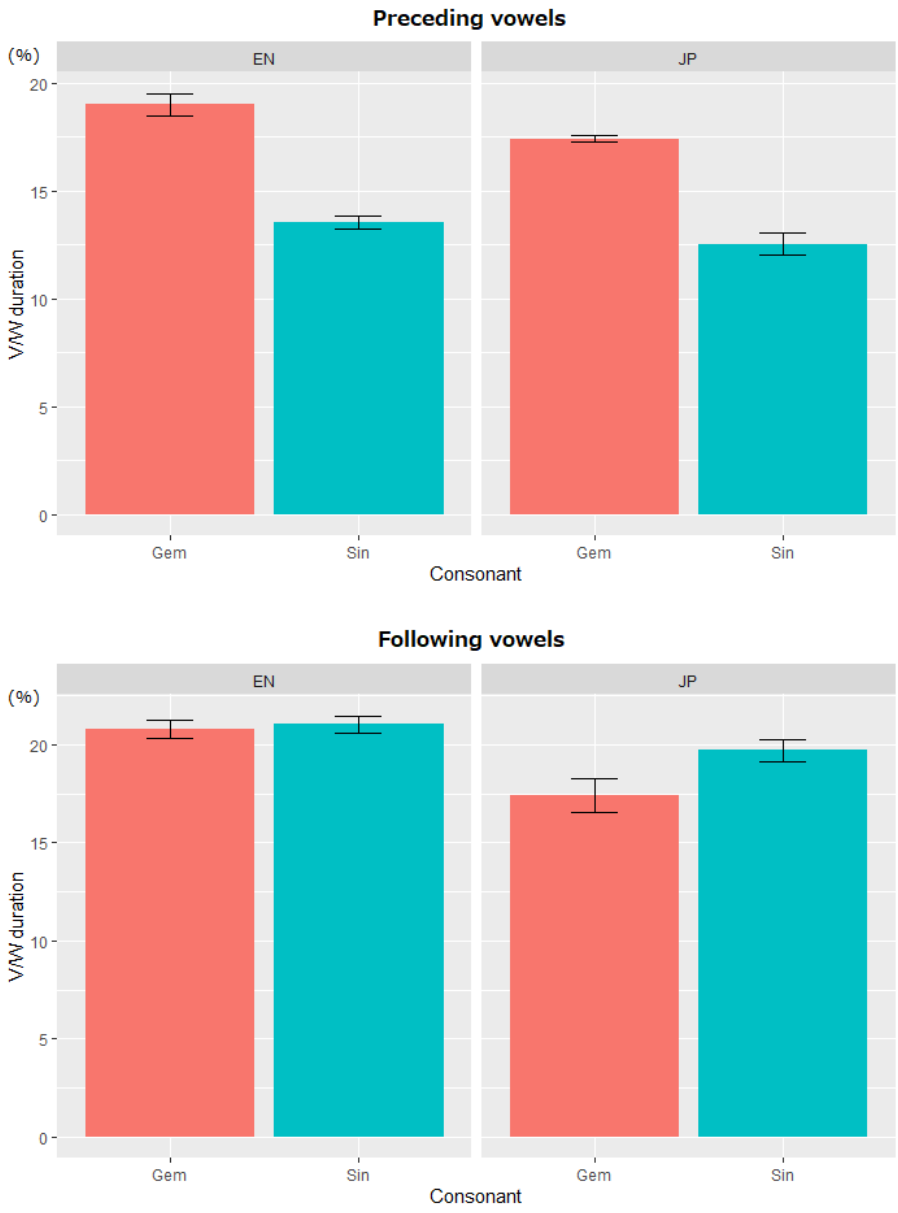

Figure 7. Mean V/W (vowel to word ratio) duration for the preceding (upper panel) and following (lower) vowels of geminates and singletons for each L1 group. Error bars indicate standard error of the mean (SEM).

L1	Word type	V1/word (%)	V2/word (%)	V3/word (%)
EN	GS	19.0	16.2	23.1
EN	SG	10.9	19.0	25.2
EN	GG	15.1	18.1	22.7
EN	SS	15.8	21.0	30.8
JP	GS	17.2	13.6	21.8
JP	SG	11.5	17.6	21.2
JP	GG	13.4	14.3	18.2
JP	SS	15.7	19.7	27.9

Table 4. Mean V/W (vowel to word ratio) duration for V1, V2 and V3 for each word type for each L1 group.

the interaction between consonant and L1 group were not significant (β = 0.13, *SE* = 0.17, *t* = 0.77, *p* = .440).

Following vowels

Figure 7 (lower panel) shows the means of V/W duration ratios for vowels that followed either geminates or singletons in GS and SG nonwords for each L1 group. The same fixed and random effects as used in the previous analysis of preceding vowels were used. Because following vowels either occurred at V2 or V3 in each nonword, and V3 was generally longer due to word-final lengthening (see Table 4), the model also included word-final lengthening as a controlling factor (vowels occurred at V3 coded as 1 while others as -1).

Results showed that, on average, vowels that followed singletons were longer than those that followed geminates (the main effect of consonant, β = -0.65, *SE* = -0.17, *t* = -3.72, *p* < .001) and the effect of consonants was smaller in EN speakers (the interaction effect of consonant and L1 group, β = 0.48, *SE* = 0.17, *t* = 2.79, *p* = .005). The main effect of word-final lengthening was significant (β = 3.25, *SE* = 0.09, *t* = 34.23, *p* < .001) and the main effect of L1 group was not (β = 1.17, *SE* = 0.80, *t* = 1.46, *p* = .148). Post-hoc multiple comparisons with Tukey HSD tests revealed that vowels that followed singletons were significantly longer than those that followed geminates in JP speakers (*z* = 3.39, *p* = .004), but there was no significant difference in EN speakers (*z* = 1.64, *p* = .600). In sum, these results suggest that EN speakers did not differentiate the duration of the following vowels, whereas JP speakers shortened the duration of vowels that followed geminates.

3.4. Secondary cues: intensity

The mean intensity ratio of V1 to V2 and V2 to V3 across all word types in each L1 group, grouped by consonants that occurred between the two vowels, is presented in Figure 8. A

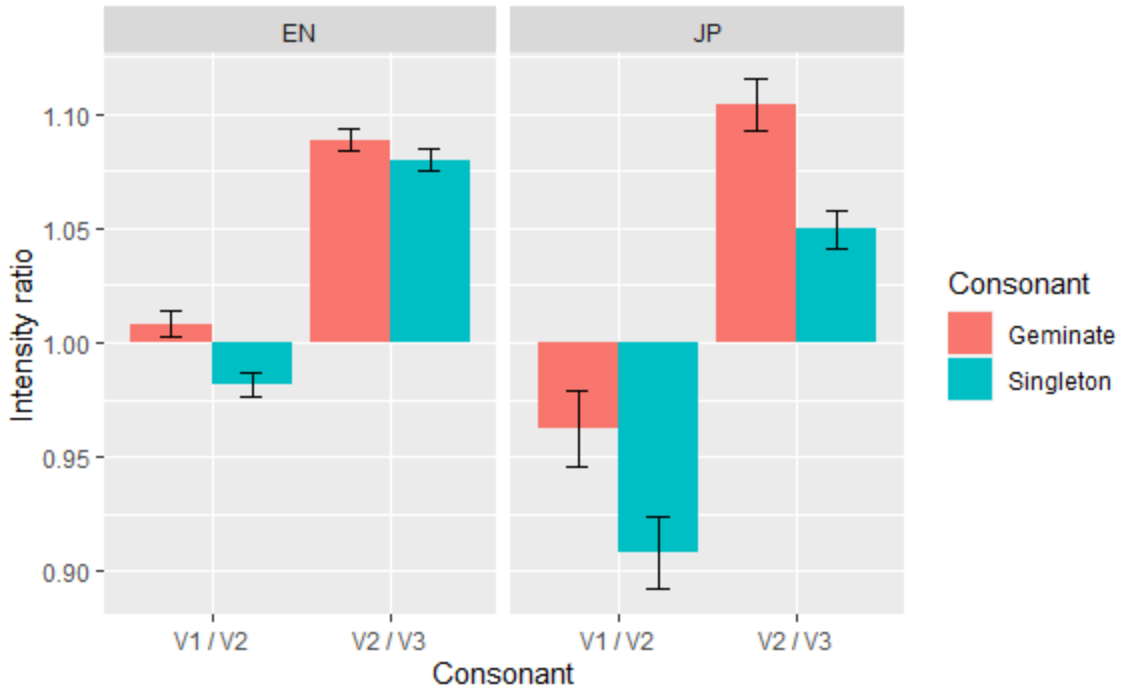

Figure 8. Mean intensity ratio of V1 to V2 and V2 to V3 across all word types for each L1 group. "Consonant" indicates which consonant (geminate or singleton) occurred between the two vowels. Error bars indicate standard error of the mean (SEM).

Fixed effects	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1.02	0.0094	5.27	108.38	0.000	***
Consonant	0.03	0.0016	7384.00	16.13	0.000	***
L1 group	0.02	0.0048	114.30	3.51	0.001	***
Position	-0.06	0.0014	7613.00	-41.60	0.000	***
Consonant x L1 group	-0.01	0.0014	7613.00	-6.61	0.000	***
Consonant x Position	0.00	0.0016	7387.00	-2.66	0.008	**
L1 group x Position	0.01	0.0014	7613.00	9.39	0.000	***
Consonant x L1 group x Position	0.00	0.0014	7613.00	1.62	0.105	

Table 5. The model estimates for fixed effects from a linear mixed model with intensity as the dependent variable, positions (V1/V2 or V2/V3), consonants (geminate or singleton), L1 groups (EN or JP), and all the possible interactions of the three as fixed effects. The asterisks indicate: ***= p<.001, **= p<.01, and *= p< .05.

linear mixed effects model was fitted to the ratios with consonant (geminates or singletons), L1 group (EN or JP), position (V1/V2 or V2/V3) and all possible two-way and three-way interactions as fixed effects and participants and word type as random intercepts. The results are shown in Table 5. What interests us most here is an interaction effect of consonant and L1 group (β = -0.01, *SE* = 0.0014, *t* = -6.61, $p < .001$), which suggests that the effect of consonant (i.e., when geminates occurred between two vowels, the preceding vowel tended to have higher intensity) was different for each L1 group. Specifically, a positive estimate for the effect of consonant and a negative estimate for the interaction effect indicate that the effect of consonant was smaller for EN speakers (although post-hoc multiple comparisons with Tukey HSD tests showed that the differences between geminates and singletons were significant at both positions for both groups, all $p < .001$). There was also a main effect of position (β = -0.06, *SE* = 0.0014, *t* = -41.60, $p < .001$): between V1 and V2, the following vowel had higher intensity than the preceding vowel, while the preceding vowel had greater intensity than the following vowel between V2 and V3, which indicates that V2 (the vowel that had lexical accent on it) had the highest intensity. This effect of position, similar to the effect of consonant, was also smaller for EN speakers (the interaction effect of L1 group and position: β = 0.01, *SE* = 0.0014, *t* = 9.39, $p < .001$).

3.5. Secondary cues: F0

Figure 9 presents the mean F0 ratio of V1 to V2 and V2 to V3, which clearly indicates that there was an effect of position, where the following vowel had higher F0 (i.e., rising pitch) between V1 and V2, and the preceding vowel had higher F0 (i.e., falling pitch) between V2 and V3, which was expected because of the location of lexical accent. A linear mixed effects model with the same fixed and random effects as the previous intensity analysis was fitted to the F0

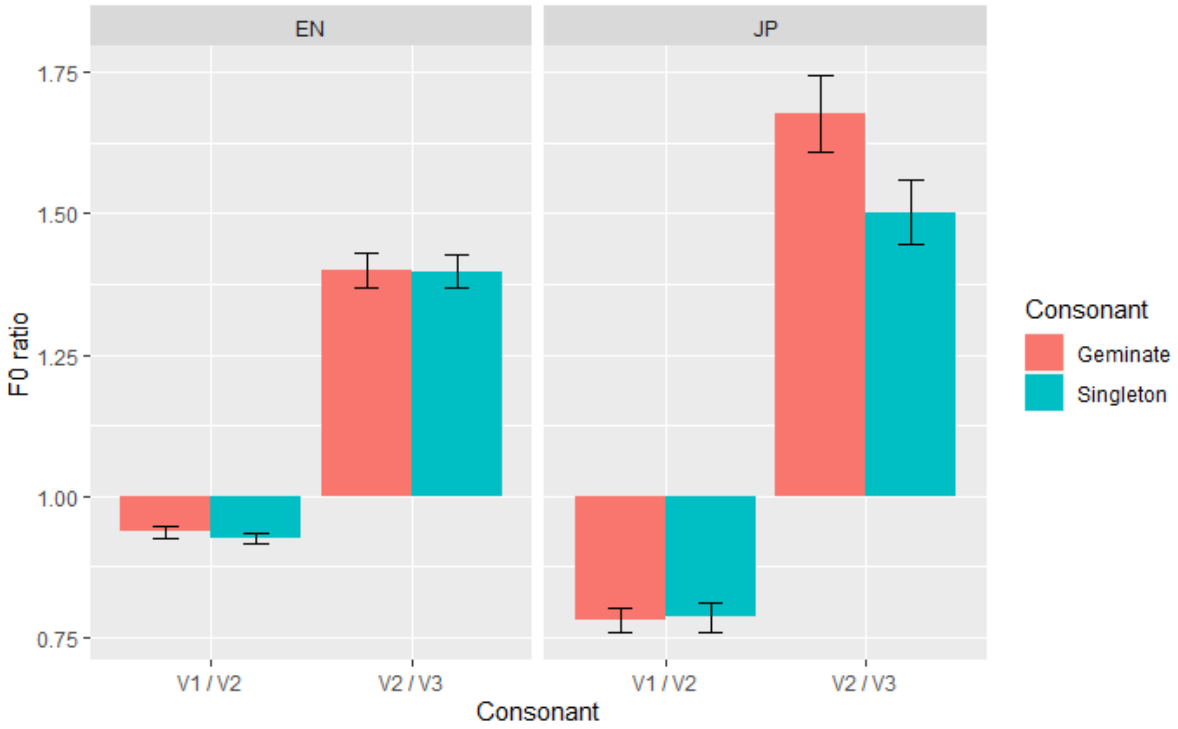

Figure 9. The mean F0 ratio of V1 to V2 and V2 to V3 across all word types for each L1 group. "Consonant" indicates which consonant (geminate or singleton) occurred between the two vowels. Error bars indicate standard error of the mean (SEM).

Fixed effects	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1.17	0.022	73.11	52.78	0.000	***
Consonant	0.03	0.007	347.80	4.23	0.000	***
L1 group	-0.02	0.021	112.60	-0.74	0.460	
Position	-0.31	0.006	7158.00	-49.27	0.000	***
Consonant x L1 group	-0.02	0.006	7153.00	-2.87	0.004	**
Consonant x Position	-0.02	0.007	357.10	-3.16	0.002	**
L1 group x Position	0.09	0.006	7158.00	13.94	0.000	***
Consonant x L1 group x Position	0.02	0.006	7153.00	3.37	0.001	***

Table 6. The model estimates for fixed effects from a linear mixed model with fundamental frequency (F0) as the dependent variable, positions (V1/V2 or V2/V3), consonants (geminate or singleton), L1 groups (EN or JP), and all the possible interactions of the three as fixed effects. The asterisks indicate: ***= p<.001, **= p<.01, and *= p< .05.

ratios. Results (Table 6) showed that the effect of position was weaker for EN speakers (the interaction effect of L1 group and position: $\beta = 0.09$, $SE = 0.006$, $t = 13.94$, $p < .001$). Regarding our primary interest (the effect of consonant), the three-way interaction effect of consonant, L1 group and position was significant ($\beta = 0.02$, $SE = 0.006$, $t = 3.37$, $p = .001$), which suggests that the effect of consonant was different for each L1 group, but only at either V1/V2 or V2/V3. Post-hoc multiple comparisons with Tukey HSD tests revealed that JP speakers showed a larger falling F0 when a geminate, rather than a singleton, was placed between V2 and V3 ($z = 5.99$, $p < .001$), while EN speakers showed no significant difference between geminates and singletons at V2/V3 ($z = -1.77$, $p = .579$). There were no significant differences between geminates and singletons at V1/V2 for either EN speakers ($z = -1.65$, $p = .662$) or JP speakers ($z = -0.26$, $p < .001$).

3.6. Acoustic cues as predictors for perceptual judgement

Finally, to investigate how each acoustic cue explained the JP listener's perceptual judgement of EN speakers' productions, several sets of generalized (logistic) linear models were built for each of the GS, SG, GG, and SS nonwords using the perceptual judgement (correct or incorrect) as a dependent variable, calculating the log odds for each trial (i.e., the log odds for each trial to be judged as "correct"), marking it as predicted "correct" when the log odds were larger than zero, otherwise as predicted "incorrect", then obtaining the prediction accuracy by calculating the percentages of trials whose predicted judgement and actual judgement matched (Table 7). First, if there is no predictor in the models and all predictions are "correct" (i.e., when the prediction accuracy is minimum), the prediction accuracy was identical to the actual percentage of "correct" at 68.3% across all word types. When the primary cues (i.e., closure duration of C1 and C2) were added to each model, the prediction accuracy increased by 19.3% to

	GS	SG	GG	SS	Mean
(1) No predictor	53.9	73.4	53.7	92.3	68.3
(2) C1, C2	90.4	92.4	74.4	93.2	87.6
(3) C1, C2, C1/C2	90.4	92.0	82.8	94.0	89.8
(4) C1, C2, C1/C2 + All secondary cues	92.0	93.4	82.5	93.7	90.4

Table 7. The prediction accuracy for each word type with different predictors. C1 and C2 indicate closure duration for C1 and C2.

	GS				SG				GG				SS			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-15.45	4.99	-3.09	0.002 **	3.80	4.67	0.81	0.416	-9.66	3.14	-3.07	0.002 **	4.52	5.81	0.78	0.436
C1	0.03	0.00	11.04	0.000 ***	-0.04	0.01	-3.91	0.000 ***	0.04	0.00	8.74	0.000 ***	0.01	0.01	1.16	0.246
C2	-0.04	0.00	-9.29	0.000 ***	0.03	0.01	4.43	0.000 ***	-0.02	0.00	-4.83	0.000 ***	-0.04	0.01	-5.14	0.000 ***
C1/C2	-0.03	0.02	-1.47	0.142	-1.63	1.69	-0.96	0.337	-4.96	0.78	-6.39	0.000 ***	-4.16	0.68	-6.08	0.000 ***
V1	-0.01	0.00	-3.18	0.001 **	-0.02	0.01	-2.69	0.007 **	0.00	0.00	-0.27	0.785	-0.02	0.01	-2.95	0.003 **
V2	-0.02	0.00	-4.71	0.000 ***	0.00	0.00	0.76	0.446	-0.01	0.00	-5.83	0.000 ***	0.00	0.01	0.75	0.455
V3	0.00	0.00	-0.06	0.950	0.00	0.00	0.76	0.448	0.00	0.00	-2.66	0.008 **	0.00	0.00	-0.25	0.800
Intensity V1/V2	13.08	2.97	4.41	0.000 ***	-5.42	2.46	-2.20	0.028 *	9.06	1.79	5.07	0.000 ***	5.55	2.84	1.96	0.050 *
Intensity V2/V3	5.59	2.80	2.00	0.046 *	4.01	3.00	1.34	0.181	4.45	1.87	2.38	0.017 *	3.68	3.86	0.95	0.341
F0 V1/V2	-0.89	0.87	-1.03	0.306	-0.30	0.96	-0.31	0.757	0.25	0.70	0.36	0.720	-1.68	1.09	-1.55	0.122
F0 V2/V3	-0.17	0.33	-0.51	0.610	-0.98	0.34	-2.88	0.004 **	-0.22	0.23	-0.96	0.337	-0.72	0.35	-2.05	0.041 *

Table 8. The model estimates for fixed effects from generalized (logistic) linear models with JP listener's perceptual judgement as the dependent variable and all acoustic cues as fixed effects.

87.6%. Considering the prediction accuracy for GG nonwords (74.4 %) was lower than other word types (> 90%), we also added the ratio of C1 to C2 (C1/C2) to the models, because if there were two geminates in a production, the JP listener may feel the duration of C1 and C2 needed to be balanced, even if each closure were independently sufficiently long. The addition of C1/C2 improved the prediction accuracy for GG nonwords by 8.4% (74.4% → 82.8%), which led to an average 89.9 % of JP listener's judgments correctly predicted by closure duration. Finally, to see whether other secondary cues had additional predictive power with respect to the listener's judgment, we added all secondary acoustic cues (i.e., vowel duration of V1, V2 and V3, intensity ratio of V1/V2 and V2/V3, F0 ratio of V1/V2 and V2/V3) to the third sets of models in addition to closure duration. Some of these secondary cues had significant effects (see Table 8), while the effects on prediction accuracy were small: the GS nonwords (90.4% → 92.0%, +1.6%) and the SG nonwords (92.0% → 93.4%, +1.4%) slightly improved by the addition of secondary cues, while the prediction accuracy for SS and GG nonwords did not change much (if anything, decreasing slightly). Across all four word-types, the addition of secondary cues improved the prediction accuracy by only 0.8% (M = 89.8% → M = 90.4%).

4. Discussion

4.1. The effect of position in word

The present study investigated whether the position of JP geminates in a word affected non-native production by novice EN speakers. Results revealed that (a) there was indeed an effect of position, and this effect was not only observed in EN (non-native) speakers, but also in JP (native) speakers; (b) in addition, EN speakers seemed more affected by position than JP speakers were. First, we observed a main effect of position in closure duration for SS and GG

nonwords, where geminates at C2 were longer than those at C1, both for EN and JP speakers (even stronger for JP speakers in SS nonwords, which underlines the finding that this position effect observed in Japanese speakers was not limited to geminates, and probably originated from factors other than consonant types). We also found the same effect of position for GS and SG nonwords, even though the difference between geminates at C1 and C2 for JP speakers did not reach significance after adjustment for multiple comparisons. We speculate that, at least for JP speakers, the effect was not caused by the position of the geminate itself, but by the position of lexical pitch accent. All target nonwords in the present study had lexical accent on V2, which means that there was a pitch drop between V2 and V3 (i.e., over the duration of C2). As discussed above, JP lexical accent is realized by falling pitch, while rising pitch at the beginning of the word (i.e., between V1 and V2, over the duration of C1) is considered as a prosodic phenomenon (Labrune, 2012). Thus, as was seen in F0 in the present data, the pitch difference was more distinct between V2 and V3 than between V1 and V2. We assume that consonants at C2 tended to be slightly longer than C1 because C2 needed to accommodate a larger pitch movement than C1. Previous research, however, has not considered lexical accent as a major factor that affects closure duration in JP geminate-singleton contrasts (e.g., Beckman, 1982; Hirata 2017, footnote 2). We assume that it was because the effect is relatively weak and might be cancelled or overpowered by other more influential factors that affect closure duration, such as place of articulation or voicing of the consonant (e.g., Idemaru & Guion, 2008). Since these factors were constant for both C1 and C2 in the current dataset, we believe that the observed effect of position in JP speakers may be attributed to lexical accent.

What interests us more is that EN speakers seemed to be affected by the geminate's position beyond what can be explained by lexical accent, as observed most clearly in closure

duration for four-mora (GS and SG) targets. There was a three-way interaction that suggested that EN speakers' geminate-singleton distinction was more affected by position than JP speakers' was: EN speakers' geminates at C1 were significantly shorter than those at C2, such that there was a larger overlap between geminates and singletons for GS nonwords (see Figure 3). GS nonwords were also judged more poorly than SG nonwords in the JP listener's perceptual judgement. Although the effect of lexical accent, as discussed above, may partially explain why EN speakers did relatively well on C2, it does not account for the difficulty that they faced when geminates occurred at C1. This implies that there may be other factors underlying the finding.

The fact that geminates at C1, closer to the word-initial position, were more difficult for EN speakers than C2 is interesting because it is inconsistent with the results reported for the JP long-short vowel contrast, which was found to be easier for non-native speakers in word-initial position (Muroi, 1995: Minagawa et al., 2002; Oguma, 2001). The JP geminate-singleton consonant and long-short vowel contrasts have similar durational structures in terms of mora-timing and have often been investigated alongside each other in previous research on the acquisition of JP non-syllabic moraic elements (e.g., Toda, 1998); /aatata/ (/aa/ is a word-initial long vowel) counts as four units of moras, just as /attata/ counts as four units of moras. The two examples are identical except for the second mora, which is filled either with the extended portion of the preceding vowel, or with the initial portion of the following consonant. We speculate that geminates near the word-initial position might be more challenging because production requires a sudden stop immediately after the onset of a word, which might need more articulatory planning before starting to speak than producing geminates later in a word or extending an initial vowel. It is, however, also worth noting that the strong position effect for EN speakers can only be seen when both geminates and singletons occurred in a word (i.e., GS or

SG nonwords), not when there were only geminates; in GG nonwords, the position effect for EN speakers was less evident and did not differ from that of from JP speakers, which could be explained by the effect of lexical accent. One possible explanation may be because non-native speakers needed to pay more attention to distinguish the contrast when both geminates and singletons occur in a word, and the position effect may emerge only under the demanding condition due to limited attentional resources that can be used for articulatory planning. In any case, the present study suggests that the JP geminate-singleton contrast indeed presents different difficulties to EN speakers depending on its position in a word, and the effect may differ from what has been reported for the JP long-short vowel contrast despite similarities in durational structures across the two types of contrasts.

We also find the position effect important in practical terms because, if true, the findings suggest that EN speakers will have substantial difficulty producing geminates in real words, which tend to be multisyllabic. As discussed above, geminates at C1 are six times more common than geminates at C2 in four-mora words (Amano & Kondo, 2003). Further research is needed to understand whether the effect of position can be observed in other phonetic contexts (especially in other lexical accent patterns), and if so, to investigate what underlies the effect, because it directly addresses the practical difficulty that non-native speakers may face in real-life oral communication.

4.2. The usage of secondary acoustic cues

The present study also investigated non-native speakers' use of secondary acoustic cues for the JP geminate-singleton contrast, such as duration, intensity, and F0 of the vowels preceding and following the consonant. The results showed that EN speakers, in general, did not utilize these cues as effectively as JP speakers did. EN speakers did not differentiate the duration

of the following vowels between geminates and singletons, where vowels following geminates should be shorter than those following singletons, although they did correctly differentiate the duration of vowels preceding geminates (i.e., longer duration for vowels that preceded geminates). EN speakers' distinction in intensity between geminates and singletons, where vowels that preceded geminates have higher intensity, was not as clear as that of JP speakers. EN speakers' inability to utilize secondary acoustic cues could be seen most clearly in F0: when pitch fall due to lexical accent (i.e., between V2 and V3 in the current target nonwords) occurred over geminates rather than singletons, a larger fall was observed for JP speakers, whereas EN speakers produced almost identical F0 structures for syllables containing geminates and singletons.

Non-native speakers, in some cases, have been shown to overuse secondary acoustic cues, especially when they are not familiar with the use of the primary cues, resulting in non-native-like productions (Schertz et al., 2015; Zhang, 2008). However, the current results showed that, in the case of the JP germinate-singleton contrast, EN speakers underused secondary acoustic cues even though their native language does not feature durational phonemic contrasts. A previous study suggested that, although non-native speakers appeared to be aware that duration was the primary cue for JP geminates, some incorrectly used duration of the preceding vowel instead of consonant closure duration to control syllable duration (Toda, 2007), possibly because vowels may be more salient than is consonant closure, which is mere silence. The current results, however, did not find the extended vowel duration common among EN speakers either. While a closer look at individual data did find a few participants who extended the preceding vowel of geminates instead of increasing closure duration, EN speakers as a group did not show longer duration for the preceding vowel of geminates compared to JP speakers. While

we do not exclude the possibility that EN speakers may systematically use other secondary acoustic cues that were not investigated in the present study, such as VOT, the present study suggests that, if there were deviations from native norms in the use of secondary acoustic cues by non-native speakers, it is more likely due to underuse rather than overuse in the case of EN speakers' production of the JP geminate-singleton contrast. The finding that the JP listener's perceptual judgements were predominantly explained by closure duration and that secondary acoustic cues only had a small impact further highlights the underuse of secondary cues. This, however, does not necessarily mean that secondary acoustic cues do not significantly contribute to perception. For instance, secondary cues, especially F0, may also increase the degree of perceived foreign accent, as shown in a previous study (Hirata & Kato, 2018). The present study indeed found that EN speakers showed a lack of distinction between geminates and singletons the most clearly in F0 structures of syllables that contained pitch accent, which suggests that learning to differentiate pitch may also be important in mastering the JP geminate-singleton contrast.

5. Conclusion

The current study examined both durational and non-durational acoustic cues of non-native speakers' JP geminate-singleton contrast in three-syllable nonwords, exploring the effect of position in the nonword and the use of secondary acoustic cues. We found the position effect, where geminates that occur near the word initial position may be more challenging than those that occur later in a word, which showed that multi-syllabic words that contain the geminate-singleton contrast may pose a specific difficulty to non-native speakers that disyllabic words may not. Considering that the learners of JP are often offered minimal pair practice using disyllabic

words when learning the geminate-singleton contrast (e.g., Kokusai kōryū kikin nihongo kyōiku kokusai sentā [The Japan Foundation Japanese-Language Education Center], 2008), the present findings suggest the importance of including multi-syllabic words in practice and of being aware that non-native speakers' mastery of the contrast should be assessed using varied phonetic contexts. The finding that the position effect for geminates seemed different from that for JP long vowels, which are similar to geminates in terms of durational properties, warrants further studies on underlying factors that make JP non-syllabic moraic elements challenging for non-native speakers. The secondary acoustic cues for geminates were underused by non-native speakers, which suggests that these secondary cues are less salient to them. Thus it may be beneficial to direct their attention to these cues (by, for example, providing explicit explanation). Although non-native speakers' use of secondary cues may only have a relatively small impact on JP listener's perceptual distinction between geminates and singletons, they may still have influence on other aspects in terms of the mastery of the contrast, such as perceived degrees of foreign accent.

General discussion

The present thesis set out two types of objectives: from a practical standpoint, we aimed to seek effective feedback conditions that optimize L2 pronunciation training for adult learners; from a research standpoint, we aimed to challenge the existing context where the focus of feedback research in instructed L2 acquisition is predominantly placed on corrective feedback (e.g., Russell & Spada, 2006; Li, 2010; Lyster & Saito, 2010; Ellis, 2017) and to discuss whether error correction is the primary role of feedback in L2 pronunciation learning. To these ends, inspired by research in motor skill learning (e.g., Maas et al, 2008; Bislick et al., 2012; Wulf & Lewthwaite, 2016), we focused on two factors: frequency (frequent vs. reduced) and type (non-corrective vs. corrective) of feedback. The former investigated whether the frequent provision of feedback yields better learning outcomes compared to less frequent feedback, whereas the latter examined the effectiveness of non-corrective feedback. Both studies investigated two types of feedback, the elicitation-type feedback (e.g., prompts) and provision-type feedback (e.g., recasts), because different results were expected for the two types of feedback.

The results from the first study of this thesis showed that when the frequency of corrective feedback was reduced to 50%, that is, only half of errors were corrected, it was still effective (i.e., yielding better learning outcomes than a no-feedback condition) when the type of feedback given was recast, but not when it was prompt. Prompts increased participants' ability to find errors in their own productions but might negatively affect learners' motivation when given frequently (100% of the time). Recasts did not increase error detection ability irrespective of frequency, and increased motivation when given frequently. These results led us to conclude that the two types of feedback may encourage learners to employ different learning strategies.

The second study revealed that non-corrective feedback, whose discourse function has been investigated (e,g., Lyster, 1998; Södergård, 2008) but not its direct effect on the learning of linguistic targets, was in fact effective much like corrective feedback was, but only when it was of the elicitation type. Non-corrective elicitation also increased learners' error detection ability *and* motivation. The results for the provision-type of non-corrective feedback, on the other hand, presented a striking contrast: it did not positively affect pronunciation accuracy, error detection, or motivation.

The third study examined the acoustics of participants' productions of the target sound, the Japanese geminate-singleton contrast. We found that participants in the present studies seemed correctly aware that consonant closure duration is the primary cue in the geminate-singleton contrast and almost solely focused on it; that is, they did not utilize other, secondary acoustic cues. The study also revealed that participants seemed to face difficulty in producing the target sound when it occurred in a specific position in the word, which may be partially due to articulatory difficulty.

Some of these findings did not replicate results from previous studies. A number of studies have demonstrated that participants who received reduced feedback outperformed those who received frequent feedback (e.g., Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012; Austermann Hula et al., 2008; Maas et al., 2012; Adams et al. 2002; Katz et al., 2010; Van Stan et al., 2017), which was not the case in the first study of this thesis; the reduced recast group performed comparably to the frequent recast group, but they did not outperform the latter. Frequent and corrective feedback (regardless of whether it was the elicitation- or repetition-type) yielded fast improvement especially at the beginning of the training, and the improvement was so significant that participants still retained some of it at the post-test despite

seemingly poorer retention rates compared to reduced feedback. We assume that this discrepancy from prior studies may stem from the very fact that we investigated corrective feedback. In the previous studies of feedback frequency on speech learning, feedback was neutral (i.e., given to both successful and unsuccessful performance) (e.g., Adams & Page 2000; Steinhauer & Greyhack, 2000; Kim et al., 2012) and Study 1 was the first study to empirically examine the frequency of corrective feedback. As recent studies suggest that corrective (negative) feedback is characterized by quick and significant improvement during training (recall our discussion in Study 2), it seems reasonable that we observed a stronger training effect for frequent corrective feedback than in previous studies on feedback frequency. This underscores the importance of replication of previous findings with respect to corrective feedback, which was, in fact, part of the motivation for the present studies.

When looking at the combined findings from the present studies, what was interesting was that we observed opposite results for the first and second studies with respect to the type of feedback. In the first study, it was the provision-type feedback that remained effective in the context of reduced frequency, whereas in the second study it was the elicitation-type feedback that remained effective when used in a non-corrective way. Across the first and second studies, all the groups that received the elicitation-type feedback (P100, P50, and NC-elicitation) improved in their ability to detect errors in their own productions, whereas all the groups that received the provision-type feedback (R100, R50, and NC-repetition) did not show changes in error detection ability. These results further highlight our discussion of how the two types of feedback seemed to promote different learning strategies. Our results suggest that eliciting learners' post-feedback actions itself encourages learners to acquire the target sound by actively figuring out the rules about errors by themselves, which is why it needed to be frequent and it

still had the similar effectiveness even when it was non-corrective. On the other hand, providing learners with correct auditory models seemed to have them employ an entirely different strategy, which was probably learning via imitation. Learning via imitation may not require as much awareness of learners' own errors because the provision of the correct model could distract learners' attention from their own productions.

One of the reasons why we observed such differences between the elicitation-type and provision-type of feedback might also be related to the nature of the target sound. As seen in Study 3, participants in our study were (correctly) focused on the durational feature of the geminate-singleton contrast. While there is little empirical evidence on which sounds can or cannot be imitated easily by L2 learners, we speculate that a phonetic feature such as duration might be easier to produce by imitation without any explicit instruction than other sounds such as vowels or laterals, which may require new, unfamiliar shapes of vocal tract. This was also suggested in our data, in which the post-feedback trials after recasts showed a high percentage of success (86.6 %). Thus, we cannot exclude the possibility that, if the target sound were more difficult to imitate, participants who received the provision-type feedback would have opted for employing other learning strategies, such as consciously seeking errors in their own productions. To further understand the differences among the types of feedback and their differential effects on L2 pronunciation learning, it would be necessary to accumulate evidence using different targets that pose different challenges to learners.

The present thesis had set out the question whether the role of feedback in L2 pronunciation training is primarily error correction, or alternatively, whether other roles such as motivation come into play more significantly than previously thought in L2 acquisition research. The answer to the question remains inconclusive, in a sense that we observed different results for

different types of feedback. For the elicitation-type feedback including prompts, we might conclude that figuring out errors seemed to play a primary function in learning a target sound, and even non-corrective elicitation may be considered as a subtle version of error correction. On the other hand, the provision-type feedback, such as recasts, showed little relationship to error awareness, and it indeed functions as a motivator for learners. The findings of the present studies suggest that feedback is not merely a tool for error correction or motivation, but something that affects how learners approach the learning task; the nature of the feedback thus helps to shape the learning experience. The effect of feedback on learning strategies and learners' perception on the whole experience, not just on linguistic attainment, would be an interesting avenue to be explored in future studies on feedback in L2 speech learning.

Finally, the original findings of the present thesis discussed in this section and throughout the thesis are the fruit of our interdisciplinary approach that connected two different domains of learning: L2 speech learning and motor skill learning. Looking across different disciplines with similar interests may offer immense benefits because it enables us to apply findings in one area to another (in our case, the effect of feedback frequency and positive feedback). In doing so, a simple replication may not be sufficient due to various significant differences between the two domains, as the nature of different types of CF led us to quite different results from what was suggested in the previous studies in motor learning. The originality and contribution of the present thesis lie in our effort to be conscious of the uniqueness and research context of CF when applying knowledge from motor learning, and by doing so, to generate new insights and implications for L2 pronunciation learning.

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of acquisition and nativelikeness in a second language: Listener perception vs. linguistic scrutiny. Language Learning, 59, 249–306.

Adams, S. G., & Page, A. D. (2000). Effects of selected practice and feedback variables on speech motor learning. Journal of Medical Speech-Language Pathology, 8, 215–220.

Adams, S. G., Page, A. D., & Jog, M. S. (2002). Summary feedback schedules and speech motor learning in Parkinson's disease. Journal of Medical Speech-Language Pathology, 10, 215–220.

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of Experimental Psychology: Human Perception and Performance, 35(2), 520-529.

Ahangari, S., & Amirzadeh, S. (2011). Exploring the teachers' use of spoken corrective feedback in teaching Iranian EFL learners at different levels of proficiency. Procedia—Social and Behavioral Sciences, 29, 1859–1868.

Amano, S. & Hirata, Y. (2010). Perception and production boundaries between single and geminate stops in Japanese. Journal of the Acoustical Society of America, 128(4), 2049–2058.

Amano, S. & Kondo, K. (2003). NTT Database series: Nihongo-no goitokusei: Lexical properties of Japanese]. Sanseido: Tokyo.

Anderson, D. I., Magill, R. A., Sekiya, H., & Ryan, G. (2005). Support for an explanation of the guidance effect in motor skill learning. Journal of Motor Behavior, 37(3), 231–238.

Austermann Hula, S. N., Robin, D. A., Maas, E., Ballard, K. J., & Schmidt, R. A. (2008). Effects of feedback frequency and timing on acquisition, retention, and transfer of speech skills in acquired apraxia of speech. Journal of Speech, Language, and Hearing Research, 51, 1088–1113.

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

Badami, R., Vaezmousavi, M., Wulf, G., & Namazizadeh, M. (2011). Feedback after good versus poor trials affects intrinsic motivation. Research Quarterly for Exercise and Sport, 82, 360-364.

Badets, A., & Blandin, Y. (2004). The role of knowledge of results frequency in learning through observation. Journal of Motor Behavior, 36, 62–70.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), Self-efficacy beliefs of adolescents (pp. 307–337). Greenwich, CT: Information Age.

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(3), 255–278.

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. CATESOL Journal, 30, 177–194.

Bates, D, Mächler, M, Bolker, B, & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

Beckman, M. E. (1982). Segmental duration and the "mora" in Japanese. Phonetica, 39, 113-35.

Bent, T., Bladlow, A., & Smith, B. (2007). Segmental errors in different word position and their effects on intelligibility of non-native speech. In Language Experience in Second Language Speech Learning: In Honor of James Flege.

Bilodeau, E. A., & Bilodeau, I. M. (1958a). Variable frequency knowledge of results and the learning of a simple skill. Journal of Experimental Psychology, 55, 379-383.

Bislick, L. P., Weir, P. C., Spencer, K., Kendall, D., & Yorkston, K.M. (2012). Do principles of motor learning enhance retention and transfer of speech skills? A systematic review. Aphasiology, 26, 709–728.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences, 17, 97–110.

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.16, retrieved from http://www.praat.org/

Brooks, N. (1960). Language learning: Theory and practice. Harcourt, Brace and Company: New York.

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. Language Teaching Research, 20(4), 436-458.

Carter, M.J., Smith, V., & Ste-Marie, D.M. (2016). Judgments of learning are significantly higher following feedback on relatively good versus relatively poor trials despite no actual learning differences. Human Movement Science, 45, 63-70.

Chiviacowsky, S., & Wulf, G. (2002). Self-controlled feedback: does it enhance learning because performers get feedback when they need it? Research Quarterly for Exercise and Sport, 73, 408-415.

Chiviacowsky, S., & Wulf, G. (2005). Self-controlled feedback is effective if it is based on the learner's performance. Research Quarterly for Exercise and Sport, 76, 42–48.

Chiviacowsky, S., & Wulf, G. (2007). Feedback after good trials enhances learning. Res Q Exerc Sport, 78, 40–47.

Chiviacowsky, S., Wulf, G., de Medeiros, F. L., Kaefer, A., & Wally, R. (2008). Self-controlled feedback in 10-year-old children: higher feedback frequencies enhance learning. Research Quarterly for Exercise and Sport, 79, 122–127.

Derwing, T. M. (2003). What do ESL students say about their accents? Canadian Modern Language Review, 59, 547–566.

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. System, 30, 155-166.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. Language Learning, 64, 526–548.

Dhawale, A. K., Smith, M. A., & Olveczky, B. P. (2017). The role of variability in motor learning. Annual Review of Neuroscience, 40, 479–498.

El Tatawy, M. (2002). Corrective feedback in second language acquisition. Working papers in TESOL & Applied Linguistics, 2(2), 1e19.

Ellis, R. (2009). Corrective feedback and teacher development. L2 Journal, 1, 3-18.

Ellis, R. (2016). Focus on form: A critical review. Language teaching research, 20(3), 405-428.

Ellis, R. (2017). Oral Oral corrective feedback in L2 classrooms: What we know so far. In Nassaji, H. & Kartchava, E. (Ed.), Corrective Feedback in Second Language Teaching and Learning: Research, Theory, Applications, Implications (pp. 3-18). Abingdon-on-Thames, U.K.: Routledge.

Fagan, D. S. (2014). Beyond "excellent": uncovering the systematicity behind positive feedback turn construction in ESL classrooms. Novitas-ROYAL (Research on Youth and Language), 8(1), 45–63.

Ferreira, A., Moore, J. D., & Mellish, C. (2007). A Study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computer-assisted language learning systems. International Journal of Artificial Intelligence in Education, vol. 17(4), 389-422.

Flege, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), Second language acquisition and the Critical Period Hypothesis (pp. 101–131). Mahwah, NJ: Erlbaum.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. Journal of the Acoustical Society of America, 97, 3125–3134.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? A lexical decision study in French listeners. Journal of Experimental Psychology: Human Perception and Performance, 32, 1276–1293.

Fu, T., & Nassaji, H. (2016). Corrective feedback, learner uptake, and feedback perception in a Chinese as a foreign language classroom. Studies in Second Language Learning and Teaching, 6, 161–183.

Gass, S., Mackey, A., & Ross-Feldman, L. (2005). Task-based interactions in classroom and laboratory settings. Language Learning, 55, 575–611.

Gluszek, A., & Dovidio J. F. (2010a). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. Personality and Social Psychology Review, 14, 214–237.

Gluszek, A., & Dovidio, J. F. (2010b). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging to the United States. Journal of Language and Social Psychology, 29, 224–234.

Gooch, R., Saito, K., & Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English /ɹ/ to Korean adult EFL learners. System, 60, 117-127.

Green, D.M., & Swets, J.A. (1966). Signal detection theory and psychophysics, Wiley: New York.

Grenon, I., & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In Chan, H., Jacob, H. & Kapia, E., Proceedings of the 32nd Annual Boston University Conference on Language Development, 155–166. Somerville: Cascadilla.

Guadagnoli, M. A., Dornier, L. A., & Tandy, R. D. (1996). Optimal length for summary knowledge of results: The influence of taskrelated experience and complexity. Research Quarterly for Exercise & Sport, 67, 239–248.

Guadagnoli, M., & Kohl, R. (2001). Knowledge of results for motor learning: Relationship between error estimation and knowledge of results frequency. Journal of Motor Behavior, 33(2), 217–224.

Guillemot, C. (2018). The role of L1 durational correlates in L2 acquisition: A production study of Japanese geminates by Italian, French and English L2 learners. In Proceedings of ISAPh 2018 International Symposium on Applied Phonetics, 57–61.

Han, J., & Jung, J. (2007). Patterns and preferences of corrective feedback and learner repair. Korean Journal of Applied Linguistics, 23(1), 243-260.

Han, M. S. (1992). The timing control of geminate and single stop consonants in Japanese: A challenge for nonnative speakers. Phonetica, 49, 102-127.

Hao, Y., & de Jong, K. (2016). Imitation ofsecondlanguagesoundsinrelationtoL2perception and production. Journal of Phonetics, 54, 151–168.

Harada, T. (2006). The acquisition of single and geminate stops by English-speaking children in a Japanese immersion program. Studies in Second Language Acquisition, 28(4), 601-632.

Harvranek, G. (2002). When is corrective feedback most likely to succeed?. International Journal of Educational Research, 37(3-4), 255-270.

Hirata, Y. (1990). Tango/bun reberu-ni okeru sokuon-no kikitori [Perception of geminate consonants at word and sentence level]. Onsei Gakkai Kaihou [Phonetic Society Reports] 194, 23–28.

Hirata, Y. (2017). Second language learners' production of geminate consonants in Japanese. In The Phonetics and Phonology of Geminate Consonants.

Hirata, Y., & Amano, S. (2012). Production of single and geminate stops in Japanese three- and four-mora words. The Journal of the Acoustical Society of America 132, 1614-1625.

Hirata, Y., & Kato, H. (2018). Acoustic and perceptual evaluation of Japanese geminates produced by L2 learners. In Proceedings of NINJAL ICPP 2018, retrieved on January 2021, from http://crosslinguistic-studies.ninjal.ac.jp/prosody/wp-content/uploads/sites/9/2018/04/ICPP2018_Hirata.pdf

Hirata, Y., & Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. The Journal of the Acoustical Society of America, 118, 1647–1660.

Hirata, Y., & Whiton, J. (2005). Effects of speech rate on the singleton/geminate distinction in Japanese. Journal of the Acoustical Society of America, 118, 1647–1660.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal 50(3), 346-363.

Idemaru, K., & Guion-Anderson, S. (2010) Relational timing in the production and perception of Japanese singleton and geminate stops. Phonetica 67: 25–46.

Idemaru, K., & Guion, S. (2008). Acoustic covariants of length contrast in Japanese stops. Journal of the International Phonetic Association, 38 (2), 167-186.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition, 87, B47-B57.

Jean, G., & Simard, D. (2011). Grammar teaching and learning in L2: Necessary, but boring? Foreign Language Annals, 44, 467–494.

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. Frontiers in Psycholosy, 5(1246), 1–17.

Katz, W., McNeil, M., & Garst, D. (2010). Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback. Aphasiology, 24, 826–837.

Kawahara, S. (2005). Voicing and geminacy in Japanese: an acoustic and perceptual study. University of Massachusetts Occasional Paters in Linguistics 31, 87-120.

Kawahara, S. (2015). The phonetics of obstruent geminates, sokuon. In H. Kubozono (Ed.), The Mouton Handbook of Japanese Language and Linguistics. Berlin: Mouton de Gruyter.

Kennedy, S. (2010). Corrective feedback for learners of varied proficiency levels: A teacher's choices. TESL Canada Journal, 27(2), 31–50.

Kim, I. S., LaPointe, L. L., & Stierwalt, J. A. (2012). The effect of feedback and practice on the acquisition of novel speech behaviors. American Journal of Speech-Language Pathology, 21, 89–100.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. Journal of the Acoustical Society of America 59, 1208-21.

Kokusai kōryū kikin nihongo kyōiku kokusai sentā [The Japan Foundation Japanese-Language Education Center] (2008). Kyōshiyō nihongo Kyōiku hando bukk 6 Hatsuon kaitei ban [Japanese-language education hand book 6: pronunciation (revised edition)] , Bonjinsha: Tokyo.

Krashen, S. (1981). Second language acquisition and second language learning. Oxford : Pergamon.

Krause, D., Agethen, M., & Zobe, C.(2018). Error feedback frequency affects automaticity but not accuracy and consistency after extensive motor skill practice. Journal of Motor Behavior, 50, 144-154.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." _Journal of Statistical Software_, *82*(13), 1-26. Doi: 10.18637/jss.v082.i13 (URL: https://doi.org/10.18637/jss.v082.i13).

Labrune, L. (2012). The phonology of Japanese. Oxford: Oxford University Press

Lai, Q., & Shea, C. H. (1998). Generalized motor program (GMP) learning: Effects of reduced frequency of knowledge of results and practice variability. Journal of Motor Behavior, 30, 51–59.

Landis, R. J., Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Lasagabaster, D., & Sierra, J. M. (2005). Error correction: Students' versus teachers' perceptions. Language Awareness, 14, 112-127.

Lee, A. & Mok, P. (2016). Durational correlates of Japanese phonemic quantity contrasts by Cantonese-speaking L2 learners. In Proceedings of Speech Prosody 2016, 597-601.

Lee, A. H., & Lyster, R. (2015). The effects of corrective feedback on instructed L2 speech perception. Studies in Second Language Acquisition, 38, 1–30.

Lee, E. J. (2013). Corrective feedback preferences and learner repair among advanced ESL students. System, 41, 217–230.

Lee, E. J. (2016). Reducing international graduate students' language anxiety through oral pronunciation corrections. System, 56, 78–95.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. Language Learning, 60, 309–365.

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), Understanding second language process (pp. 27-44). Clevedon, UK: Multilingual Matters.

Lochtman, K. (2002). Oral corrective feedback in the foreign language classroom: how it affects interaction in analytic foreign language teaching. International Journal of Educational Research, 37, 271–283.

Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). L2 learners' beliefs about grammar instruction and error correction. The Modern Language Journal, 93(1), 91-104.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. Journal of the Acoustical Society of America, 89, 874-886.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, D. Coste, R. Ginsberg, & C. Kramsch (Eds.), Foreign language research in cross-cultural perspective (pp. 39–52). Amsterdam: Benjamins.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. R. Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468), San Diego: Academic Press.

Long, M. (2006). Problems in SLA. Mahwah, NJ: Erlbaum.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. Studies in Second Language Acquisition, 20, 51-81.

Lyster, R., & Izquierdo, J. (1998). Prompts versus recasts in dyadic Interaction. Language Learning, 59, 453-498.

Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. Studies in Second Language Acquisition, 28, 269–300.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. Studies in Second Language Acquisition, 19, 37–66.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. Studies in Second Language Acquisition, 32, 265–302.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. Language Teaching, 46, 1–40.

Maas, E., Butalla, C. E., & Farinella, K. A. (2012). Feedback frequency in treatment for childhood apraxia of speech. American Journal of Speech-Language Pathology, 21, 239–257.

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders.

Magill, R. A. (2007). Motor learning and control: Concepts and application. McGraw-Hill.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods, 44(2), 314–324.

McAuley, E., Duncan, T., & Tammen, V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. Research Quarterly for Exercise and Sport, 60, 48–58.

Méndez, E. H., & Cruz, M. R. R. (2012). Teachers' perceptions about oral corrective feedback and their practice in EFL classrooms. Profile, 14(2), 63-75.

Minagawa, Y., Maekawa, K., & Kiritani, S. (2002). Nihongo washa no cho tan boin no doutei ni okeru picchi gata to onsetsu ichi no kouka [Effects of Pitch Accent and Syllable Position in Identifying Japanese Long and Short Vowels: Comparison of English and Korean Speakers]. Journal of the Phonetic Society of Japan, 6(2), 88-97.

Mori, R. (2011). Teacher cognition in corrective feedback in Japan. System, 39, 451-467.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 45, 73–97.

Muraki, M., & Nakaoka, N. (1990). Hatsuon to sokuon - eigo, chugokugo washa no hatsuon [Syllabic nasals and geminates - pronunciation of English and Chinese speakers]. In M. Sugimoto (Ed), Kouza nihongo to nihongo kyoiku 3 nihongo no onsei, onin (ge), 139-177.

Muroi, K. (1995). Eigo washa no nihongo no tokushuhaku no chikaku to sanshutsu ni okeru sho mondai [Problems of the perception and production of Japanese morae - the Case of native English speakers]. Sophia Lunguistica, 38, 41-60.

Neustupný, J. V. (1966). Is the Japanese accent a pitch accent? Onsei-Gakkai Kaihoo, 121. Reprinted in M. Tokugawa (ed.), Akusento (pp. 230-239). Tokyo: Yuuseidoo, 1980.

Nicholson, D. E., & Schmidt, R. A. (1991). Scheduling information feedback to enhance training effectiveness. Proceedings of the Human Factors Society Annual Meeting, 35, 1400-1402.

Nieuwenhuis, S., Slagter, H. A., von Geusau, N. J., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: Differential activation of human cortical areas by positive and negative outcomes. European Journal of Neuroscience, 21, 3161–3168.

Oguma, R. (2001). Nihongo gakushusha no cho-on no sanshutsu ni kansuru shutoku kenkyu - cho-on ichi ni yoru nanido to shutoku junjo [The study on the acquisition of the production of long-vowls by learners of Japanese - the difficulty and order of acquisition based on the position of long-vowels in word]. Nihongo kyoiku [Japanese Language Education], 109, 110-117.

Onishi, H. (2013). Cross-linguistic influence in third language perception: L3 and L3 perception of Japanese contrasts (Unpublished doctoral dissertation). University of Arizona.

Panova, I., & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. TESOL Quarterly, 36, 573–595.

Park, J., Shea, C. H., & Wright, D. L. (2000). Reduced-frequency concurrent and terminal feedback: A test of the guidance hypothesis. Journal of Motor Behavior, 32(3), 287-296.

Patterson, J.T., & Azizieh, J. (2012). Knowing the good from the bad: Does being aware of KR content matter?. Human Movement Science, 31, 1449-1458.

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

Rezaei, S., Mozaffari, F., & Hatef, A. (2011). Corrective feedback in SLA: classroom practice and future directions. International Journal of English Linguistics, 1(1). 21-29.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A metaanalysis of the research. In J. M. Norris & L. Ortega (Eds.), Synthesizing research on language learning and teaching (pp. 133–164). Amsterdam: Benjamins

Russell, V. (2009). Corrective feedback, over a decade of research since Lyster and Ranta (1997): Where do we stand today? Electronic Journal of Foreign Language Teaching, 6(1), 21-31.

Saemi, E., Porter, J. M., Ghotbi-Varzaneh, A., Zarghami, M., & Maleki, F. (2012). Knowledge of results after relatively good trials enhances self-efficacy and motor learning. Psychology of Sport and Exercise, 13, 378-382.

Safari, P. (2013). A descriptive study on corrective feedback and learners' uptake during interactions in a communicative EFL class. Theory and Practice in Language Studies, 3(7), 1165-1175.

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: teaching the perception and production of English /ɹ/ to adult Japanese learners. Language Learning, 63, 499–529.

Saito, K. & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English. Language Learning, 62(2), 595–633.

Salmoni A. W., Schmidt, R. A., & Walter, C. A. (1984). Knowledge of results and motor learning: A review and critical reappraisal, Psychological Bulletin, 95(3), 355-386.

Sato, M. (2013). Beliefs about peer interaction and peer corrective feedback: Efficacy of classroom intervention. The Modern Language Journal, 97, 611–633.

Schertz, J. L., Cho, T., Lotto, A. J., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. Journal of Phonetics, 52, 183–204.

Schmidt, R. A. (1991). Frequent augmented feedback can degrade learning: Evidence and interpretations. In J. Requin & G. E. Stelmach (Eds.), Tutorials in motor neuroscience, pp. 59–75, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. Psychological Science, 3(4), 207–217.

Schmidt, R. A., & Lee, T. M. (2011). Motor control and learning: a behavioural emphasis, 5th ed. IL: Human Kinetics.

Schultz, W. (2013). Updating dopamine reward signals. Current Opinion in Neurobiology, 23, 229–238.

Schulz, R. (2001). Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA–Colombia. The Modern Language Journal 85.2, 244–258.

Self-Determination Theory Research Group. (n.d.). Intrinsic Motivation Inventory. Retrieved from http://www.selfdeterminationtheory.org/intrinsic-motivation-inventory/

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. Language Teaching Research, 8, 263–300.

Sheen, Y. (2007). The effects of corrective feedback, language aptitude and learner attitudes on the acquisition of English articles. In A.Mackey (Ed.), Conversational interaction in second language acquisition: A collection of empirical studies (pp. 301–22). Oxford, England: Oxford University Press.

Sheen, Y. (2011). Corrective feedback, individual differences and second language learning. Dordrecht: Springer.

Sidaway, B., Ahn, S., Boldeau, P., Griffin, S., Noyes, B., & Pelletier, K. (2008). A comparison of manual guidance and knowledge of results in the learning of a weight-bearing skill. Journal of Neurologic Physical Therapy, 32, 32–38.

Silva, L. C. D., Pereira-Monfredini, C. F., & Teixeira, L. A. (2016). Improved children's motor learning of the basketball free shooting pattern by associating subjective error estimation and extrinsic feedback. Journal of Sports Sciences, 35, 1825–1830.

Sippel, L., & Jackson, C. N. (2015). Teacher vs. peer oral corrective feedback in the German language classroom. Foreign Language Annals, 48(4), 688–705.

Södergård, M. (2008). Teacher strategies for second language production in immersion kindergarten in Finland. In T. Fortune & D. Tedick (Eds.), Pathways to bilingualism and multilingualism: Evolving perspectives on immersion education, 152–173. Multilingual Matters: Clevedon

Sparrow, W. A., & Summers, J. J. (1992). Performance on trials without knowledge of results (KR) in reduced relative frequency presentations of KR. Journal of Motor Behavior, 24, 197-209.

Steinhauer, K., & Grayhack, J. P. (2000). The role of knowledge of results in performance and learning of a voice motor task. Journal of Voice, 14, 137–145.

Sugawara, S. K., Tanaka, S., Okazaki, S., Watanabe, K., Sadato, N. (2012). Social rewards enhance offline improvements in motor skill. PLoS ONE, 7, doi: 10.1371/journal.pone.0048174

Sukegawa, Y. (1993). Bogo betsu ni mita hatsuon no keikou: ankeeto chosa no kekka kara [The tendency of pronunciation based on speakers' first language: results of a survey]. In Nihongo onsei to nihongo kyouiku monbu shou juuten ryouiki kenkyuu "Nihongo onsei ni okeru inritsu teki tokuchou no jittai to sono kyouiku ni kansuru sougou teki kenkyuu", 187-222.

Sullivan, K. J., Kantak, S. S., & Burtner, P. A. (2008). Motor learning in children: feedback effects on skill acquisition. Physical Therapy, 88, 720–732.

Suzuki, M. (2004). Corrective feedback and learner uptake in adult ESL Classrooms. Working Papers in TESOL and Applied Linguistics, 4(2). Retrieved from http://journals.tc-library.org/index.php/tesol/article/view/58

Toda, T. (1994). Gaikokujin gakushusha no tokushuhaku no shutoku [Acquisition of special morae in Japanese as a second language]. Onsei-kenkyu [Journal of Phonetics Society of Japan], 7(2), 70-83.

Toda, T. (2007). Nihongo kyoiku ni okeru sokuon no mondai [Issues regarding geminate consonants in Japanese language education]. Onsei-kenkyu [Journal of Phonetics Society of Japan], 11(1), 35-46.

Trempe, M., Sabourin, M., & Proteau, L. (2012). Success modulates consolidation of a visuomotor adaptation task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 52–60.

Trofimovich, P. & Baker, W., "Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech", 28:1-30, 2006.

Van Stan, J. H., Mehta, D. D., Petit, R., Sternad, D., Muise, J., Burns, J. A., & Hillman, R. E. (2017). Integration of motor learning principles into real-time ambulatory voice

biofeedback and example implementation via a clinical case study with vocal fold nodules. American Journal of Speech-Language Pathology, 26, 1-10.

Vander Linden, D. W., Cauraugh, J. H., & Greene, T. A. (1993). The Effect of Frequency of Kinetic Feedback on Learning an Isometric Force Production Task in Nondisabled Subjects, Physical Therapy, 73 (2), 79-87.

Vásquez, C., & Harvey, J. (2010). Raising teachers' awareness about corrective feedback through research replication. Language Teaching Research, 14(4), 421-443.

Waring, H. Z. (2008). Using explicit positive assessment in the language classroom: IRF, feedback, and learning opportunities. The Modern Language Journal, 92(4), 577-594.

Wasding, R. (2013). Feedback expressions used by an English teacher of Tour and Travel Department. Indonesian Journal of Applied Linguistics, 3(1), 53-67.

Weeks, D. L., & Kordus, R. N. (1998). Relative Frequency of Knowledge of Performance and Motor Skill Learning. Research Quarterly for Exercise and Sport, 69(3), 224-230.

Winstein, C. J. & Schmitt, R. A. (1990). Reduced Frequency of Knowledge of Results Enhances Motor Skill Learning. Journal of Experimental Psychology Learning Memory and Cognition, 16(4), 677-691.

Wrembel, M. (2010). L2-accented speech in L3 production, International Journal of Multilingualism, 7(1), 75-90.

Wulf, G , & Lewthwaite, R. (2016). Optimizing performance through intrinsic motivation and attention for learning: the OPTIMAL theory of motor learning. Psychonomic Bulletin & Review, 23, 1382–1414.

Wulf, G., & Schmidt, R. A. (1989). The learning of generalized motor programs: Reducing the relative frequency of knowledge of results enhances memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4), 748–757.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. Psychonomics Bulletin, 9, 185–211.

Wulf, G., Lee, T. D., & Schmidt, R. A. (1994). Reducing Knowledge of Results about Relative versus Absolute Timing: Differential Effects on Learning. Journal of Motor Behavior, 26(4), 362-369.

Wulf, G., Schmidt, R. A., & Deubel, H. (1993). Reduced feedback frequency enhances generalized motor program learning but not parameterization learning. Journal of Experimental Psychology: Learning, Memory, and Cognition,19, 1134–1150.

Wulf, G., Shea, C. H., &Matschiner, S. (1998). Frequent feedback enhances complex motor skill learning.Journal of Motor Behavior, 30, 180–192.

Yang, Y. & R. Lyster (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. Studies in Second Language Acquisition 32.2, 235-263.

Yoshida, R. (2008). Teachers' choice and learners' preference of corrective feedback types. Language Awareness, 17(1), 78-93.

Zhang, L. J., & Rahimi, M. (2014). EFL learner's anxiety level and their beliefs about corrective feedback in oral communication classes. System, 42, 429–4.

Zhang, Y., Nissen, S. L. & Francisa, A.L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. The Journal of the Acoustical Society of America, 123, 4498; https://doi.org/10.1121/1.2902165

Zobe, C., Krause, D., & Blischke, K. (2019). Dissociative effects of normative feedback on motor automaticity and motor accuracy in learning an arm movement sequence. Human Movement Science, 66, 529-540.