Kobe University Repository : Kernel

PDF issue: 2026-11-26

Perception and Production of Secondary Acoustic Cues to Syllable-final Stop Voicing by Mandarin Native Speakers

Zhang, Yaming

(Degree)
博士（学術）

(Date of Degree)
2019-03-25

(Date of Publication)
2020-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第7369号

(URL)
https://hdl.handle.net/20.500.14094/D1007369

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。

博士論文

Perception and Production of Secondary Acoustic Cues to Syllable-final Stop Voicing by Mandarin Native Speakers
（中国語母語話者による音節末破裂音の有声性の知覚と生成―副次的音響手がかりの知覚訓練を中心に―）

平成 31 年 1 月

神戸大学大学院国際文化学研究科

張　亜明

ZHANG YAMING

CONTENTS

1. Introduction 1

1.1 Nature of English Syllable-final Stop Voicing 1

1.2 Underlying Causes to the Acquisition Difficulty for Mandarin Speakers 5

1.3 Possibility to Improve Perceptual Ability of Foreign Language 12

1.4 Purpose of the Study 14

2. Previous Studies 16

2.1. Syllable-final stop voicing perception and production by English speakers 16

2.2. Syllable-final stop voicing perception and production by Mandarin speakers ... 20

2.3. Phonemic vowel duration contrast for Mandarin speakers 24

2.4. Vowel duration as a secondary acoustic cue for Mandarin speakers 28

2.5. Previous Studies on Distribution Learning 34

3. Experiment 1 – Production of Syllable-final Voicing 40

3.1. Reading Material 40

3.2. Participants 40

3.3. Procedures and Measurements 41

3.4. Results 41

3.5. Discussion 44

4. Experiment 2: One - dimensional Distributional Learning 47

4.1. Participants 47

4.2. Stimuli 47

4.3. Procedures 49

4.4. Results 53

4.5. Discussion 55

4.6. Follow-up Experiment 57

4.7. Results and Discussions 58

5. Experiment 3: Two-dimensional Distributional Learning 60

5.1. Participants 60

5.2. Stimuli 60

5.3. Procedures 63

5.4. Results 66

5.5. Discussion 68

6. General Discussion 70

7. Conclusion 75

8. Further Studies 79

Acknowledgements 80

References 81

Appendix 93

1. Introduction

This study is undertaken in order to investigate the perception and production of English syllable-final stop voicing by Mandarin native speakers, and also to examine the effect of distributional learning on the acquisition of secondary acoustic cues.

Foreign language skill has always been seen as an important part in education in China. According to a national census conducted in 2000, a total of 425.95 million people in China had studied one or more foreign languages (National Bureau of Statistics, 2000). Among them, 93.8% had studied English, which makes it by far the most popular foreign language in China.

It is known that in second language acquisition, one's first language (L1) has a potential impact on the acquisition process, and the influence of L1 could appear in both perception and production. Therefore when it comes to English learning, learners with different language backgrounds generally encounter a variety of problems. For Mandarin speaking English learners, one of the difficulties is the perception of unreleased syllable-final stop voicing. For instance, Mandarin speakers can only distinguish the English minimal pair /bat/ and /bad/ when the final consonant is fully released. However, in daily English conversation, these words are often pronounced in an unreleased way as [bæt̚] and [bæd̚].

The fact that Mandarin speakers have difficulty distinguishing this type of pronunciation variant will potentially cause misunderstanding in communication.

1.1 Nature of English Syllable-final Stop Voicing

In phonetics a sound is usually described by the place of articulation and the

manner of articulation. Based on the manner of articulation, a stop consonant is also called a plosive consonant. The articulation of a plosive consists of three processes: catch (the airway closes so that no air can escape), occlusion (the airway stays closed causing a slight pressure difference), and release (the closure is opened and the released airflow produces a sudden impulse causing an audible sound). Based on the place of articulation, we can further divide English stop consonants into three types, which are bilabial sounds /p/ and /b/, alveolar sounds /t/ and /d/, and velar sounds /k/ and /g/. And also, each pair of these sounds is a contrast between voiceless sound and voiced sound. The voicing feature of a sound refers to the articulatory process in which the vocal folds vibrate.

A minimal pair is formed by two words that differ by only one phoneme, usually by one feature. In English, we can find many minimal pairs that only differ in voicing feature but have completely different meanings, i.e. /bet/ - /pet/, /pɪk/ - /pɪg/.

After Lisker & Abramson (1964), VOT (Voice onset time) has been regarded as the most important cue to identify stop voicing especially in syllable-initial position. Voice Onset Time is defined as the duration of the period of time between the release of a stop consonant and the onset of the vocal fold vibration. Because the vibration of the vocal fold starts before the release, VOT value is usually minus in the case of voiced sound. On the contrary, voiceless stop sound usually has a positive value.

In English however, most syllable-final consonants are pronounced unreleased, which means there is no audible burst at the end for listeners to capture. Because every process of articulation could provide certain acoustic cues, the loss of a release burst will make its perception more difficult. For instance, no release of the consonant also means that VOT no longer serves as an effective cue in perception. In this case, listeners will have to rely on other secondary acoustic cues to identify the sound.

Figure 1 shows the spectrogram of an English minimal pair *lad* and *lat*. From this figure we can see that there is a difference in both vowel duration and F1 offset frequency between the two words.

It is reported that English native speakers mainly rely on preceding vowel duration cues in perceiving syllable-final stop voicing, and they rely on preceding vowel duration cue so much that even the removal of the final release burst does not change their perception significantly (Raphael, 1972; Rodgers, 2008; Hillenbrand et al., 1984). They tend to perceive the following consonant as a voiced sound when the vowel duration is relatively long, and as a voiceless sound when the vowel duration is shorter.

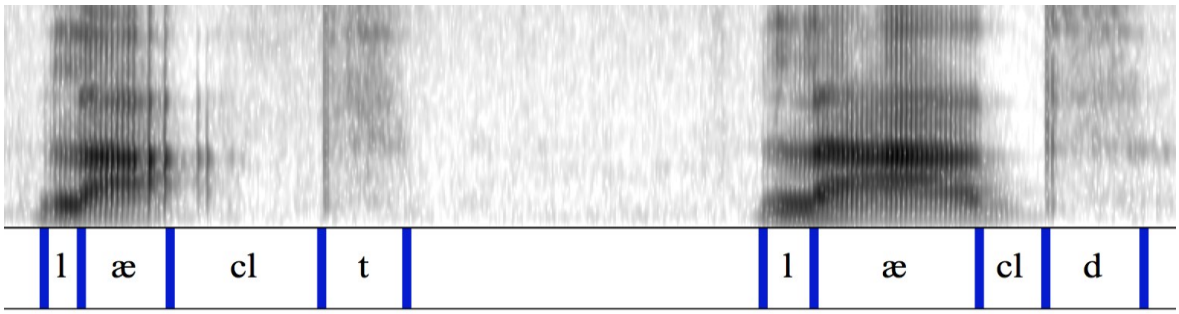


Figure 1 | Spectrogram of English minimal pair *lad* and *lat*.

In addition，it is reported that the first formant (F1) of preceding vowels also varies with the voicing of the following consonant, i.e. F1 offset frequency of the preceding vowel usually has a lower value when it is followed by a voiced stop consonant. English native speakers can also rely on this vowel to consonant transition to perceive syllable-final voicing (Hillenbrand et al., 1984).

Therefore, based on the perception strategy of English native speakers, a long vowel and (or) a lower F1 offset indicate a voiced stop consonant, while a short vowel and (or) a higher F1 offset indicate a voiceless stop consonant.

Some studies show that English native speakers have obtained the sensitivity to vowel duration differences conditioned by voicing at a very early age. Eilers (1977) reported that English learning infants at around two months of age already use vowel duration as a supplementary cue for discriminating final consonantal voicing. Eilers et al. (1984) also reported that 5 to 11 months old infants have the ability to discriminate vowel duration differences, though their performance was much poorer than that of adults. Ko et al. (2009) further examined English learning infants' sensitivity to vowel duration conditioned by syllable-final consonant voicing. The results suggested that the infants' sensitivity to extrinsic vowel duration begins to develop between 8 and 14 months.

In addition to perception, English native speakers also produce different vowel durations as a cue for the listeners. It has been reported that English learning infants start to control vowel durations conditioned by voicing in their speech production before the age of two (Ko, 2007).

Some researchers have interpreted vowel length effect as a universal process that is due to articulatory contingencies. They proposed production explanation, which states that a later closure onset is selected in the context of voiced stops because it allows for more root advancements within closure (Lehiste, 1970). However, this account could not explain some cases in which the durational patterns are independent of voicing. For instance, Flege (1979) reported that in Arabic, long vowel /a:/ was not significantly longer when it was preceding word-final /d/ than /d/. And also, Keating (1979) measured Polish vowel durations before voiced and voiceless consonants. It was found that Polish vowel duration does not vary systematically according to the voicing of the following consonant.

These exceptions have suggested that it may be a language-specific pattern that has

to be learned. Therefore, some proposed the auditory explanations, which claim that speakers in a language community produce the vowel length effect as a strategy to perceptually enhance the voicing feature of the consonants by indicating contrast effects between the vowel duration and the closure-duration of consonants (Kluender et al., 1988).

As a second language learner, is it essential to receive the enhanced features in perception when talking to a native speaker, and from aforementioned studies we can see that the ability to master the vowel length effect would benefit English learners in both production and perception of syllable-final consonant voicing in English.

However, this is not the case for Mandarin native speakers. It was reported by previous studies that native speakers of Mandarin Chinese are able to perceive released syllable-final voicing very well, but they have difficulties in perceiving unreleased syllable-final stop voicing (Flege, 1989). And also, they do not make use of preceding vowel duration cues in perception (Zhang & Hayashi, 2015).

1.2 Underlying Causes to the Acquisition Difficulty for Mandarin Speakers

What causes the difference in perception of syllable-final voicing between Mandarin English learners and English native speakers?

It is known that in second language acquisition, one's first language (L1) has a potential impact on the acquisition process. In order to find the underlying cause to the acquisition difficulty, it is necessary to look for the differences between L1 and L2.

First, in Mandarin there is no contrast between a voiced stop and a voiceless stop, but instead a contrast between a voiceless unaspirated stop and a voiceless aspirated stop. As a consequence, unlike English native speakers who are used to identify stop

categories by voicing feature, Mandarin native speakers are more used to identify stop categories by aspiration.

The second difference comes from syllable structure. Syllable is a unit of organization for a sequence of speech sounds. In most theories of phonology, the general structure of a syllable consists of three parts: onset, nucleus and coda.

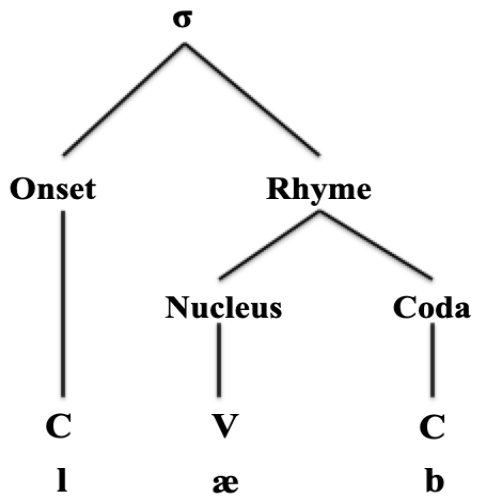


Figure 2 | Syllable structure of English word *lab* /læb/.

Onset is present in most syllables and represents the sounds at the beginning of a syllable. Some languages restrict onset to be only one single consonant (like Japanese and Mandarin Chinese), while others allow multi-consonant onset. For instance, in English the maximum number of the consonant onsets is three.

Nucleus is at the middle of a syllable and consists of at least one vowel. Generally, every syllable requires a nucleus, and the minimal syllable consists only of a nucleus like English words eye /aɪ/ and owe /o/. A combination of a nucleus and an optional coda is also called rhyme.

Coda represents the final sounds following the nucleus. In most Japanese words,

there are no other sounds following the vowel so Japanese is also called an open-syllable language. As an open-syllable language, in Japanese, most syllables have a simple CV structure. Only in several circumstances can consonants be permitted to appear at the end of a syllable, i.e. in the case of geminate consonants, the first half /p/, /t/, /k/ and /s/ are permitted to be in coda position while in the case of ん, /m/, /n/, /ŋ/ and /N/ are possible.

Mandarin Chinese also restrains the consonants in a syllable-final (coda) position. There are in total 22 consonants and 14 vowels in Mandarin. All consonants except for /ŋ/ can appear in syllable-initial position, while only the nasal sounds /n/ and /ŋ/ are permitted to be syllable-final consonants.

Meanwhile, in English words with a syllable structure of CVC are very common. Almost every consonant (except for /h/) can serve as a syllable-final consonant, and it can become even more complex if it is consonant cluster. Kessler & Treiman (1997) calculated the frequencies of different consonants in the 2,001 mono-morphemic CVC words found in the unabridged Random House Dictionary (Flexner, 1987). It is found that 63% of the words have an obstruent coda, among which 41% are voiceless obstruent and 22% are voiced obstruent.

It has long been recognized that the syllable is the basic unit in the perception of spoken language, and the difference in syllable structure between L1 and L2 can potentially influence the acquisition of L2 speech sounds. For instance, Kubozono (1995) showed that Japanese native speakers have a different segmentation strategy compared to English native speakers. It is indicated that the L1 segmentation strategy used by learners makes it difficult for them to process some within-syllable information in L2, including certain secondary acoustic cues.

In English, a voicing contrast of stop consonants occurs both syllable initially and

syllable finally. However, stop consonants never occur in syllable-final position in Mandarin. This makes voicing contrast of stop consonants in syllable-final position unfamiliar to Mandarin native speakers.

L2 learners often have difficulties in acquiring a contrast that occurs in an unfamiliar syllable position. In the case of unfamiliar syllable-final contrast, learners usually apply to syllable-final decisions the same procedures that they use for syllable-initial decisions.

Nevertheless, the same perceptual strategy cannot deal with all pronunciation variants. In daily conversation, the reduction of sounds is very common. It has been stated in some studies that the reduction of consonants is closely related to the position of the sound in a syllable. Consonant reduction happens relatively easy in a syllable-final position than in a syllable-initial position (Ohala & Kawasaki, 1984; Recasens, D., 2004).

When a consonant is reduced, the first that changes is a shortened duration, which forces the listener to get more phonemic information during a certain time. For instance, the cue of a stop release is often absent in American English. Native speakers are usually more sensitive to secondary acoustic cues and they are able to perceive reduced pronunciation variants relying on limited cues. On the other hand, non-native speakers are thought to be less sensitive to certain secondary acoustic cues and they are only able to perceive fully released words. Simply put, making use of secondary cues to perceive unreleased syllable-final voicing requires the understanding of the association between duration and voicing feature. However, because stop consonants never appear in syllable-final position in Mandarin, duration and voicing features usually do not appear together, it is difficult for Mandarin native speakers to be aware of the relationship between the two adjacent acoustic features.

The fact that Mandarin native speakers apply the same the perceptual strategy to syllable-initial contrast and syllable-final contrast makes it difficult for them to perceive those reduced pronunciation variants that are not fully released.

Flege (1989) investigated Mandarin Chinese speakers' ability to perceive English syllable-final /t/ and /d/ contrast. It is found that Mandarin speakers can perceive an unedited English /t/ and /d/ contrast in syllable-final positions very well, but when the release burst and closure voicing part is removed, they performed much worse than English native speakers. The result indicated that Mandarin Chinese speakers tried to rely on syllable-initial cues to identify syllable-final contrast, thus the released burst in pronunciation is crucial for them in the perception of syllable-final stop voicing.

Zhang & Hayashi (2015) also examined Mandarin native speaker's perception of syllable-final consonants by using stimuli with a structure of $C_1V_1C_2C_3V_2$ with the consonant in C_2 position unreleased. In the study, the perceptual ability of Mandarin speakers was compared to that of Japanese speakers. It was found that Mandarin speakers had a worse performance compared to Japanese native speakers in distinguishing voiced and voiceless consonants in C_2 position. Because Japanese has a similar syllable structure to Mandarin Chinese, simply being unfamiliar with stop contrast in syllable-final position could not fully explain the difference in perceptual ability between Mandarin speakers and Japanese speakers.

This brings up another underlying cause to this problem, which is the limited sensitivity of Mandarin native speakers to vowel duration differences. Because there is no phonemic vowel duration contrast in Mandarin, Mandarin native speakers are not sensitive to vowel duration difference.

Vowel duration is one of the important acoustic features of vowels. In some languages vowel duration difference is phonemic, which means long and short vowel

contrast can distinguish meaning. For instance, in Japanese /toke:/ means “clock” while /to:ke:/ means “statistics”. In some other languages, like English and Mandarin, vowel duration difference is not phonemic, which means that lengthening or shortening of a vowel does not change the meaning of a word. It is considered that the native speakers of languages like Japanese show more sensitivity to vowel duration difference changes, and native speakers of English and Mandarin are assumed to be relatively less sensitive to vowel duration differences. Nevertheless, English speakers have been found more sensitive to vowel duration variants than Mandarin speakers. It is considered that this sensitivity is related to the existences of *intrinsic vowel duration* and *extrinsic vowel duration* in English.

In English there is a distinction between lax vowels and tense vowels. These vowels mainly differ in vowel quality, while vowel duration can also serve as an acoustic cue in perception. One characteristic of tense vowels is that they tend to be longer in duration than lax vowels. For instance, English front high vowels /i/ and /ɪ/ differ in both vowel quality and vowel duration, and duration is an effective cue that adult English speakers make use of when discriminating the two vowels. This kind of duration difference of vowel is defined as intrinsic vowel duration, which means that the segment duration is affected by the articulation of the segment itself.

Besides intrinsic vowel duration, there is also extrinsic vowel duration in English. In the case of extrinsic vowel duration, the duration of a vowel is affected by articulations of a preceding or following segment. One major influence comes from syllable-final consonant voicing. The preceding vowel duration tends to be longer before a voiced consonant than a voiceless consonant. This phenomenon can be observed in many languages, while it is most significant in English. It has been found that English native speakers mainly rely on preceding vowel duration cues to perceive

syllable-final stop voicing (Raphael, 1972; Rodgers, 2008).

It is considered that the existences of both intrinsic and extrinsic vowel duration differences makes English native speakers sensitive to vowel duration variables, and they also use it as a secondary cue in perception. This is the main reason why English native speakers can perceive syllable-final stop voicing even when the final consonant is unreleased.

On the other hand, vowel duration barely plays any role in Mandarin perception. Although as a tonal language the vowel duration might change due to different tones, previous studies have shown that Mandarin native speakers rely less on duration for tone distinction compared to L2 learners (Chang, 2011).

It is considered that the insensitivity towards vowel duration could have a potential impact on L2, which make it difficult for Mandarin-speaking learners to associate vowel durational information to syllable-final voicing feature in English. This causes the difficulty in perceiving English syllable-final voicing in reduced production variants.

Some of the evidence is provided by Zhang & Hayashi (2015), which compared the perceptual ability of Mandarin native speakers to Japanese native speakers. Although as an open-syllable language similar to Mandarin Chinese, most words in Japanese have a simple CV structure and syllable-final consonants barely occur, because there is phonemic vowel duration contrast in Japanese, vowel duration does play an important role in Japanese perception. This makes them more sensitive to the durational differences in spoken language.

The results showed that Mandarin Chinese speakers performed better than Japanese speakers in general, but their identification rate was more affected by the position of the consonant in the syllable. At the same time, Japanese speakers have less

difficulty perceiving the voicing in unreleased syllable-final consonants compared to Mandarin Chinese speakers. They also could make better use of vowel duration as a cue to perceive voicing in syllable-final stops. From here, we can see that the sensitivity of Japanese speakers towards vowel duration not only benefits their L1 perception but also their L2 perception.

1.3 Possibility to Improve Perceptual Ability of Foreign Language

Some studies have shown that the perceptual ability of foreign speech sounds could be improved through training.

In the study of Flege & Wang (1989), the performance of English syllable-final /t/ and /d/ by Chinese speakers was improved as the result of feedback training, though the effect is weak compared to Cantonese speakers and Shanghai dialect speakers. However, during feedback training, the subjects were exposed to all acoustic features. The improvement after training is not necessarily related to increased perceptual ability to make use of preceding vowel duration cues in perception.

Because we know that not being able to make full use of vowel duration cue is the main reason, it should be possible to focus on the training of vowel duration cue. The present study investigates the effect of distributional learning, which is a dimension specific training method, which has drawn people's attentions in recent years.

Distributional learning is defined as learning by simply listening to the frequency distributions of the speech sounds in the environment. It is the learning mechanism that human beings use to acquire speech sounds in their early life (Lacerda, 1995; Guenther & Gjaja, 1996).

It is reported that over the first year of life, the perception of infants show a universal pattern and they are able to discriminate both native and non-native phonetics categories; while after that period the perception will narrow to only those contrasts made in their native language (Werker & Tees, 1984). For instance, it is found that Japanese learning infants can discriminate English /r/ and /l/, in spite of the fact that /r/ and /l/ are within one category in Japanese and adults have difficulty telling them apart (Kuhl et al., 2006).

It is possible that during the first year of life, English learning infants and Mandarin learning infants are equally sensitive to vowel duration variations despite of their different importance in adult language perception. Also, the sensitivity of Mandarin learning infants might decrease after perception narrowing to language specific.

In the first year of life, infant's speech sound perception has been observed to change from universal to language-specific, and distributional learning has been hypothesized to contribute to the emergence of language-specific speech perception. It is reported that young infants that are 2- to 3-months old can already acquire phonemic differences through fast distributional learning (Wanrooij et al. 2014a), and also distributional learning is not limited to linguistic materials but also non-linguistic auditory sequences (Saffran et al. 1999).

Some studies show that distributional learning is also effective for adults. For instance, Ong et al. (2015) reveals that non-tone language listeners can also acquire lexical tone categories through distributional learning. Moreover, it is also demonstrated that distributional learning has long-lasting effects (Escudero & Williams, 2014). Meanwhile, some researches also suggest that the effect of distributional learning in adults is less significant than in infants, indicating that it is a

weaker mechanism for learning speech sounds in adults.

1.4 Purpose of the Study

The perception of syllable-final stop voicing has been regarded as one of the most difficult in English learning for Mandarin native speakers.

As mentioned earlier, we can see that the acquisition difficulty comes from the negative transfer from L1, such as syllable structure difference and the lack of vowel length contrast. It is necessary to find an effective training method to help the learners to overcome the difficulty.

As a typical example of secondary acoustic cues, mastering vowel duration conditioned by syllable-final voicing requires L2 learners to associate two adjacent acoustic features. This is the reason why distributional learning comes into the vision. Association is the main component of distributional learning, which is considered to help building up the association between two stimuli that are likely to co-occur. In the case of the present study, the listeners must not only perceive the difference between the long and short vowel (or low and high frequency), but also build the association between the distributions of vowel duration (or F1 offset frequency) with the appearance of syllable-final voicing features.

In the present study, attempts will be made to find out whether or not distributional learning is an effective method for L2 learners to acquire secondary acoustic cues. Most studies investigated distributional learning at a phonemic level and verified its effectiveness, while the preceding vowel duration effect in English is phonological. This study will examine whether or not Mandarin Chinese speakers can obtain sensitivity to secondary acoustic cues to perceive syllable-final voicing through

distributional learning.

Therefore, the goal of this study is to investigate the underlying mechanism of acquiring secondary acoustic cues in L2 speech sounds, and also to find a possible training method (distributional learning) to help learners acquire difficult L2 phonetic contrasts.

In order to achieve the research goal for this the study, a total of four experiments were conducted. First, a production experiment was conducted, and the experiment focused on the production of voiced and voiceless contrast by Mandarin Chinese speakers in comparison to English native speakers.

Experiment 2 examined distributional learning effect of secondary acoustic cue in time dimension, which is preceding vowel duration feature. Experiment 3 examined distributional learning effect of secondary acoustic cues in both time dimension and frequency dimension, which are preceding vowel duration feature and F1 offset frequency feature respectively.

In second language acquisition, the relationship between production and perception has been widely discussed, and some suggest that they are closely connected. As part of experiment 3, the present study also investigates the effect of distributional learning on production, to find out whether production can benefit from perceptual training in a direct way.

2. Previous Studies

In second language acquisition, there are influences from first language in many different aspects, and the influence can occur in both the perception and production of spoken language.

It has been long recognized that the syllable is the basic unit in the perception of spoken language, and some previous researches suggested that L1 syllable structure has a potential impact on L2 speech acquisition. In Kubozono (1996a), 20 pairs of monosyllabic English words CVC and CVC were played to 20 Japanese native listeners. They were told that they would hear two English words, and they were asked to repeat the words and then make a new word by combining the initial part of the first word and the final part of the second word.

The result showed that the participants chose to take the CV part from the first word and C from the second word more frequently, which means they tend to employ a CV/C segmentation rather than C/VC segmentation pattern. It is indicated that the L1 segmentation strategy used by learners makes it difficult for them to process some within-syllable information in L2, including secondary acoustic cues.

2.1. Syllable-final stop voicing perception and production by English speakers

In the production of spoken language, the articulation of each sound is not completely isolated; some indirect information can be stored in nearby context, which can serve as a secondary cue in perception when the dominating cue is missing.

In the perception of English syllable-final stop consonants, an audible release burst and closure voicing usually signal the voicing feature of the final consonant. At the

same time, preceding vowel duration and formant transition can also serve as a secondary cue when the consonant is unreleased.

Some studies have given evidence that preceding vowel duration is an efficient cue to the perception of syllable-final stop voicing.

Denes (1955) indicated that the duration of vowel and the duration of consonant would have influence on the perception of voicing features. He shortened the duration of /s/ sound in use /jus/ and lengthened the duration of /z/ in use /juz/.

The result indicates that the identification of voicing of the consonant increases as the ratio of the duration of final consonant to preceding vowel decreases. This is supported by the fact that when /s/ is pronounced alone it always sounds voiceless, while when it was in a word or after a vowel it may be heard as a voiced sound.

Raphael (1972) tested this effect by varying the duration of vowel duration preceding stop contrasts. The study used a series of CVC words ending in either a voiced stop consonant or a voiceless stop consonant, with the preceding vowel duration ranging from 150ms to 350ms. The stimuli were randomly played to 25 English native speakers, and the subjects were asked to make a choice between the members of a minimal pair of words.

The result showed that final consonants are perceived as voiceless sounds when the preceding vowel duration is short and as voiced sounds when preceding vowel duration is long.

This indicates that preceding vowel duration is a sufficient cue for English native speakers to the perception of the voicing characteristic of word-final consonants and it is more effective before stop consonants compared to fricatives. The result also suggests that the perception cued by vowel duration is continuous rather than categorical.

Moreover, Hillenbrand et al. (1984) conducted two experiments concerning closure voicing, vowel-to-consonant transition, and release burst, to investigate which of these features affect English native speaker's perception the most.

In the experiments, the subjects were asked to identify the voicing of the syllable-final consonants from the edited CVC words ending in stop consonants: /pɛb/, /pɛd/, /pɛg/, /pag/, /pig/ and /pug/. Vowels for all six syllables are of the same duration. In experiment 1, closure voicing and preceding vowel-to-consonant transition were removed while vowel duration and the final release bursts were retained. In experiment 2, release bursts were removed.

It is found that the subjects tended to change from hearing voiced stops to hearing voiceless stops when the closure interval and a portion of the vowel to consonant transition had been removed. And the removal of the final bursts didn't significantly alter the identification. The result indicates that English native speakers do not mainly rely on release burst in perception, and whether or not final consonant is released or unreleased does not affect English native speaker's perception.

Some research focused on whether or not English-speaking children use acoustic cues in the same way as the adults.

Ko et al. (2009) examined 8-month-old and 14-month-old infants' perceptual sensitivity to vowel duration conditioned by post-vocalic consonantal voicing. In the experiment, three CVC minimal pairs /bæg/-/bæk/, /cʌb/-/cʌp/ and /pɪg/-/pɪk/ were used as the stimuli; half the infants heard CVC stimuli with short vowels, and half heard stimuli with long vowels. In both groups, stimuli with voiced and voiceless final consonants were compared. The result showed that older infants are sensitive to mismatching vowel duration and consonant voicing in the short condition, while no significant result was found in younger infants in either condition.

The study suggested that sensitivity to preceding vowel duration conditioned by post-vocalic consonantal voicing might develop over the course of the second half of the first year of life. It is not surprising that English native speakers mainly rely on preceding vowel duration in perception. In daily conversation, reduction of sounds is a normal phenomenon and both vowels and consonants can be reduced under certain circumstances.

It has been stated in some studies that the reduction of consonants is closely related to the position of the sound in a syllable. Consonant reduction happens relatively easier in syllable-final position than in syllable-initial position (ohala & Kawasaki 1984; Recasens, D. 2004). In the case of syllable-final stop consonant reduction, cues like release burst may be absent and the reliability is low. Comparing to other cues, preceding vowel duration is more stable and it is less affected by the context.

Besides, some studies also suggest that children do not base there voicing decisions as strongly on vocalic duration as adults did, but that the frequency of F1 offset have a stronger influence on children's responding compared to that of adults (Krause, 1982; Wardrip-Fruin & Peach, 1984).

When it comes to production, it has been reported that English learning infants start to control vowel durations conditioned by voicing in their speech production before the age of two (Ko, 2007).

Ko (2007) investigated the acquisition of vowel duration in English speaking Children to find out when children begin to produce different vowel durations as a function of syllable-final voicing. The data used in the study comes from the *Providence Corpus,* which contains longitudinal spontaneous child-adult speech interactions of six children from southern New England. A total of 803 samples were extracted and analyzed. Among them, minimal pairs or near-minimal pairs of the form

(C)VC that occur in isolation or in final position of a prosodic phrase were selected for the measurement of the vowel duration.

The findings suggested that children start to make the durational differences between tense and lax vowels before the age of two. At the same time, they also control the vowel duration conditioned by syllable-final voicing in production before the age of two.

2.2. Syllable-final stop voicing perception and production by Mandarin speakers

In Mandarin, there is no contrast between a voiced stop and a voiceless stop, but instead a contrast between a voiceless unaspirated stop and a voiceless aspirated stop. VOT (Voice Onset Time) is defined as the time duration between the release of a stop consonant and the onset of the vocal fold vibration. Voiceless stops have positive VOT values while voiced stops often have negative VOT values. English is known to contrast voiced and voiceless stops in word initial position. Lisker & Abramson (1964) has given two sets of VOT values for English voiced stops /b, d, g/, one is negative (-101ms, -102ms, -88ms) one is positive (1ms, 5ms, 21ms). They are both perceived as voiced stops. In Docherty's study (1992), English voiced stops in utterance initial position generally have a VOT of approximately zero, which means it has a range of -20 ~ 20ms.

In Mandarin Chinese, all stops are voiceless so VOT values are always positive. Previous studies have measured the VOT values for the Chinese voiceless aspirated /p', t', k'/ and voiceless un-aspirated /p, t, k/. It is shown that the Chinese voiceless aspirated /p', t', k'/ have higher values than the English equivalents, while the voiceless un-aspirated /p, t, k/ have similar values with the English voiced stops'

positive values. This is the reason why the same stop sound can be perceived as a voiceless sound in Mandarin Chinese but as a voiced sound in English.

After Lisker and Abramson's study in 1964, VOT has been regarded as the most important cue to identify stop voicing. However, this is only effective for syllable-initial stops.

As we can see from the definition of VOT, both the timing of the release of a stop consonant and the timing of the onset of the vocal fold vibration are essential in calculating the VOT value. In the case of unreleased syllable-final voicing perception, VOT is not an efficient feature anymore.

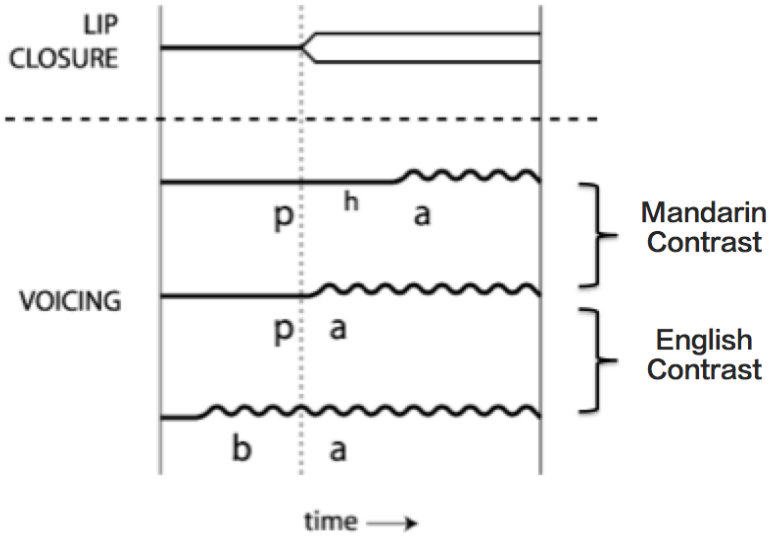


Figure 3 | Stop contrast in Mandarin Chinese and English.

(Based on Gasser, 2006, Consonants in other languages section, para. 6)

Moreover, stop contrasts only occur in syllable-initial positions. Therefore, unlike in English, voicing is not an important feature in the perception of stop consonants in Mandarin. English native speakers usually identify stop categories by voicing

categories, while Mandarin speakers are more used to identifying stop categories by aspiration.

Many previous studies have already focused on the perception of English syllable-final consonants by Mandarin Chinese speakers. By comparing their perceptual abilities with English native speakers, we can see that Mandarin speakers use different strategies to perceive syllable-final stop consonants.

Flege (1989) investigated the perception of syllable-final /t/-/d/ contrast by Mandarin Chinese speakers. It was found that Chinese speakers can perceive an unedited English /t/ and /d/ in syllable-final position very well, but when the release burst and closure voicing were removed, they performed much worse than English native speakers. The result indicated that Chinese speakers are not familiar with syllable-final stops and they attempt to rely on syllable-initial cues to identify syllable-final /t/ and /d/.

Flege & Wang (1989) conducted another experiment to examine Chinese speakers' perception of the syllable-final /t/- /d/ contrast in English, and this time the influence of dialects was also taken into consideration.

The study attempted to find out if the syllable structure of one's native language is related to the learner's sensitivity to non-native sounds appearing in certain positions. Although a contrast between /t/ and /d/ only exists in the initial but not the final position of Mandarin words, Cantonese permits unreleased /p/, /t/, /k/ in word-final position, and in the Shanghai dialect word-final glottal stop is permitted. So the experiment was conducted by dividing all the subjects into three groups based on their language background: Mandarin speakers, Cantonese speakers and Shanghai dialect speakers. Since the number of consonants that are permitted to appear in syllable-final position in their first language is different, the three groups of subjects were presumed

to show a different sensitivity to the perception of stop consonants in a second language.

The stimuli used in this experiment were derived from an audio recording of an adult native English speaker's reading of "beat", "bead", "bet" and "bed" in the phrase "Now I will say ---". Then the release bursts and closure voicing were removed by editing. In total, there are twenty-seven Chinese subjects who participated in the experiment.

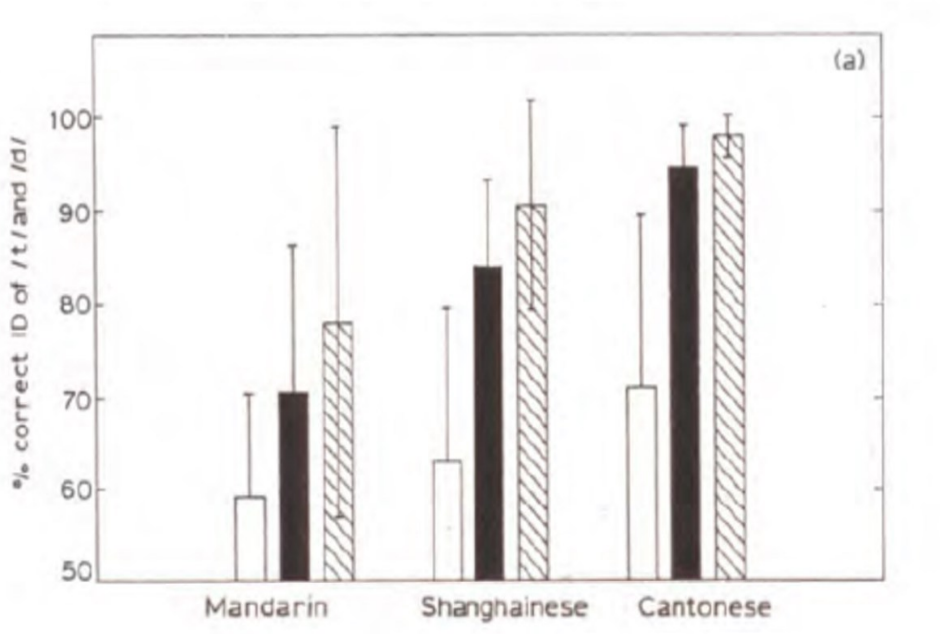


Figure 4 | Mean correct identification of /t/-/d/ by Mandarin, Shanghainese and Cantonese speakers before □, during ■, and after ▨ feedback training. (Flege & Wang, 1989，page 308)

The experiment contained three phases: the "pre-training" phase, the "feedback" phase and the "post-training" phase. In each phase, the subjects were asked to select the target sounds /d/ or /t/ by pushing a button.

During the second “feedback” phase, the subjects were informed of the correct answer immediately after pushing the "t" or "d" button. The identification rate for each final stop consonant was calculated.

The result showed that the correct identification rate of native speakers of Cantonese, as predicted, was the highest at 88%, and Shanghai dialect speakers had a higher correct identification rate than native Mandarin speakers (79% and 69% respectively).

The study also investigated the effect of feedback training on perception. During the training phase, the subjects were informed of the correct answer soon after they made the choice between /t/ and /d/. As the result of training, the performance of each group was improved. But the difference between pre-training phase and post-training phase was greater for Cantonese and Shanghai dialect speakers (26% and 28%) than for Mandarin speakers (19%).

The difference between the three groups supports the hypothesis that if there is more syllable-final stops consonants permitted in first language, the learners will have more sensitivity towards syllable-final voicing consonants in foreign language. This result also indicates that the effectiveness of a training method varies with learner’s language background.

2.3. Phonemic vowel duration contrast for Mandarin speakers

It was proved by previous studies that vowel duration is the most effective cue to the perception of stop consonant’s voicing features for English native speakers. Whether or not Mandarin speakers can make good use of this cue becomes the next question to investigate. The key point is considered to be their sensitivity to the change

of vowel duration and by reviewing some previous studies, we can see there is a big difference between Mandarin speakers and English native speakers in this aspect.

As mentioned above, vowel duration difference is not phonemic in Mandarin. Some might argue that this is not entirely true, because some tonal differences are also reflected in durational differences. It is known that there are in total five tones (including neutral tone) in Mandarin, and tone 2 and tone 3 have been considered to be the most perceptually confusable for non-native speakers. Simply put, Mandarin tone 2 is characterized as having a high-rising pitch contour and tone 3 is characterized as having a low falling-rising pitch contour. Some studies found that tone 3 is consistently longer than tone 2, and thus duration is suggested to be an effective acoustic cue for tone distinction (Blicher et al. 1990).

Chang (2011) investigated whether or not vowel duration serves as a perceptual cue and if L1 and L2 listeners weigh the cue in the same manner in distinguishing the two tones.

Six Mandarin native speakers were recruited to record the stimuli, which were two syllables /ma/ and /da/ carrying four Mandarin tones. Duration normalization was performed using PSOLA function in Praat. Through this procedure, ma2 and ma3 were all duration-normalized to 470ms while da2 and da3 to 370ms.

Eight Mandarin native speakers and eight non-native speakers participated in the experiment. The experiment consisted of two forced choice discrimination tasks. Task 1 used the stimuli where tone 3 was naturally longer than tone 2, while Task 2 used duration-normalized stimuli in which tone 2 and tone 3 have the same duration. All stimuli were presented to the participants in E-Prime, and the participants were asked to press the button Y when the tones they heard were the same and to press N for different tones.

The result showed that L2 listeners exhibited significantly lower accuracy when the duration cue is absent. This suggests that L2 listeners give more perceptual weight to duration.

On the other hand, native speakers do not heavily rely on duration for tone discrimination, and their accuracy did not suffer in discriminating two duration-normalized tones.

At the same time, efforts have been made to examine how Mandarin speakers acquire a foreign language where phonemic vowel duration differences exist.

Kurihara (2004) conducted a listening test to investigate how Chinese native speakers perceive long vowels by vowel duration. In the test, the stimuli were two Japanese minimal pairs 「かと(過渡)/かとう(加藤)」 and 「とこ(床)/とこう(渡航)」. The words were inserted into the carrier sentence [これを～と読みます] and were recorded by a Japanese native speaker.

The recordings were then edited in software PICOLA plus 2. The subjects were 30 Mandarin native speakers who were studying in Japan. They were presented with the stimuli through a headphone.

The results showed that Chinese speakers perceive vowels with short duration as long vowels, and also among Chinese speakers it is hard to see a similar tendency of categorical perception compared to Japanese native speakers. The result was compared to their Japanese proficiencies and no relations were found.

He also investigated how Mandarin-speaking Japanese learners produce Japanese long and short vowels. In Kurihara (2005), he asked the 30 Mandarin speaking Japanese learners to pronounce the same minimal pairs as in Kurihara (2004). Four Japanese native speakers were recruited to evaluate their pronunciations. Acoustic analyses were also conducted by using SUGI Speech Analyzer, the total duration and

final vowel duration values were measured. The result was also compared to their Japanese proficiency levels.

It was revealed that subjects with relatively low Japanese proficiency tend to produce long vowels shorter, and subjects with relatively high Japanese proficiency tend to produce short vowels longer. This suggests that at the beginning of their language studies, the pronunciation of long vowels did not receive enough attention. Even for advanced learners, producing both long and short vowels correctly still remains a difficult task.

The lack of vowel duration difference in Mandarin also reflects in prosody production. Minagawa-Kawai (2001) compared the production of nonsense words by Mandarin native speakers and Japanese native speakers. As is well known, Japanese is a mora-timed language, which means the duration of every mora is equal. That is to say, the duration of the consonant and the vowel within a mora usually have a compensating relationship. It was hypothesized in the study that Japanese speakers would show a stronger tendency to keep CV length equal than speakers of other languages, because CV must not be too long in Japanese where CV and CVC/CVV have a durational contrast.

Previous studies already examined the existence of temporal compensation while Minagawa-Kawai focused on the issue of how differently temporal compensation is performed depending on language. The study tested the hypothesis by comparing Japanese with Mandarin, which has a different rhythm pattern from Japanese. In the pronunciation test, two nonsense words “maCVka” and “maCVnka” were made, where the CV is /ra/, /ba/, /sa/, /ri/, /bi/ and /si/, and these words are inserted in a carrier sentence.

The recordings were analyzed on computer and segmental durations of the

recorded utterances were measured. The duration of C_2 (e.g. /b/ in /mabaka/) and the duration of its following vowel V_2 in each utterance were plotted.

After comparing the tendency of temporal compensation effect across languages, it is clear that when the consonant duration increases, the vowel duration decreases. This is true for both Japanese and Chinese speakers but the extent of temporal compensation is larger for Japanese than for Chinese. Japanese speakers show stronger temporal compensation because CV and CVV have a durational contrast, while the contrast does not exist in Mandarin.

Although this study examined temporal compensation in production, the result also gave some evidence that Mandarin speakers pay less attention to vowel duration in spoken language compared to Japanese speakers.

2.4. Vowel duration as a secondary acoustic cue for Mandarin speakers

Zhang & Hayashi (2015) conducted two experiments to examine the perception of English consonants in syllable-final position by Mandarin native speakers and Japanese native speakers, and to also find out to what extent they rely on the vowel duration cues to perceive syllable-final stop voicing.

The stimuli used in this experiment were nonsense English words containing 13 American English consonants /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /m/, /n/ and /l/. These sounds were chosen because they represent consonants with different manners of articulation and also the contrast of voiced and voiceless are concluded. All the words have the same structure of $C_1V_1C_2C_3V_2$, with /e/ and /i/ in V_1 and V_2 position and /p/ in C_1 position. The 13 target consonants appear in C_2 and C_3 position

alternately. Among them, words with the same consonant in C_2 and C_3 positions were removed from the list.

An American English native speaker recorded the stimuli in a soundproof room. The speaker pronounced all the words naturally and the consonants in C_2 position were all unreleased. The subjects were told to listen to a series of nonsense English words with a structure of /peC_2C_3i/ and they were asked to concentrate on the two consonants in C_2 and C_3 positions, and then to write down the sounds they had identified on the answer sheet. The identification rate for each consonant in C_2 and C_3 position was calculated and the error patterns in each position were also recorded.

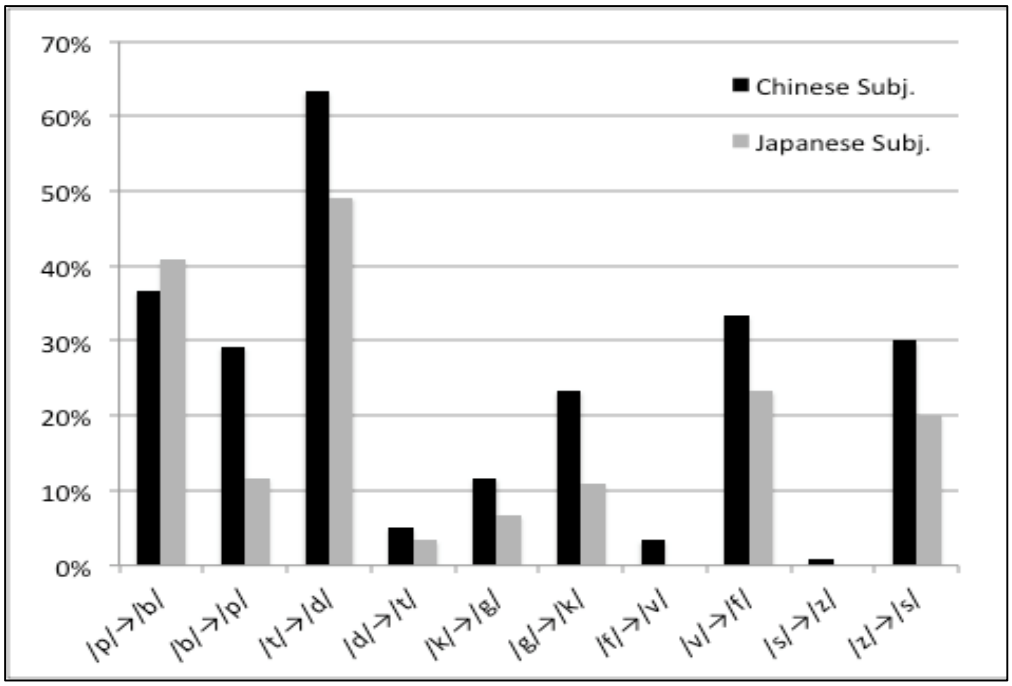


Figure 5 | Percentage of errors occurring within voiced-voiceless contrast in syllable-final position. (Zhang & Hayashi, 2015, page 2)

The results showed that although Chinese subjects have a higher average score than Japanese subjects in general, the perception of Chinese subjects is more affected

by syllable position. It seemed that when a consonant appears in the onset position, Chinese speakers do not have any problems identifying it, but when it appears in the syllable-final position, the identification rate could be quite low. For example, Chinese speakers have an average identification rate of 95.83% for syllable-initial /t/ while for syllable-final /t/ this number drops to only 15%.

As can be seen in Figure 5, when it comes to syllable-final stop consonants and fricative consonants, Chinese speakers made more mistakes than Japanese speakers in perceiving the voicing feature. This difference suggests that Chinese subjects must be less sensitive to certain acoustic features cueing voicing than the Japanese subjects, which was further investigated in the second experiment.

In the second experiment of the study, six English words with a simple CVC structure /cɑp/, /cɑb/, /bɛt/, /bɛd/, /pɪk/ and /pɪg/ were recorded by an American English native speaker. For each word, the vowel durations were measured. The stimuli were made by editing the words ending with voiceless stops. The release burst was eliminated and the duration of the preceding vowel was lengthened in five steps. In the final step, the duration is equal to that of the preceding vowel of the words ending with voiced consonants. The stimuli are made based on the voiceless sounds because there is no closure voicing following the vowel. So after removing the release burst from each word, the vowel duration will be the only parameter in this experiment, because vowel-to-consonant transition is the same for every word in each group.

The subjects were presented with the stimuli and they were asked to identify if the word ended with a voiced or a voiceless consonant.

The results showed a significant difference between Mandarin and Japanese speakers (Figure 6). In Figure 6, the percentage judged as voiceless by Japanese speakers becomes lower if the preceding vowel duration becomes longer. There was a

significant difference between groups of different vowel durations. This suggested that for the Japanese subjects, the final consonant was likely to be perceived as voiceless when the preceding vowel duration is shorter. This is the same result as reported for English native speakers by Raphael (1972). Thus Japanese speakers show a tendency to use vowel duration as a voicing cue in a similar way to English speakers. No significant difference was found as a function of vowel duration in the perception of Mandarin native speakers.

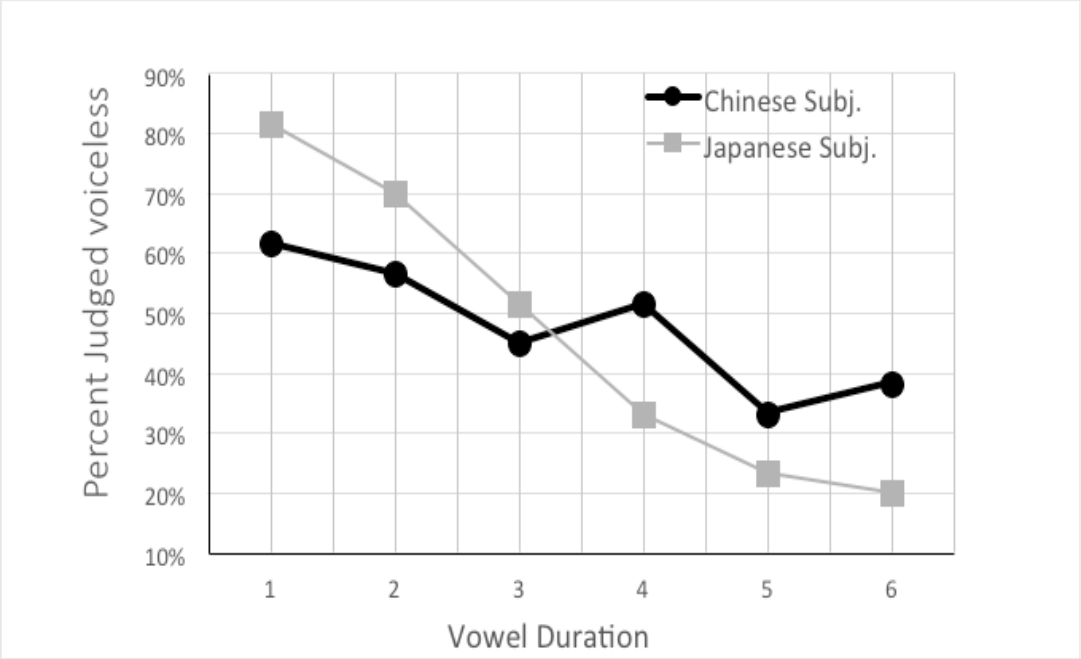


Figure 6 | Percentage judged as voiceless by Chinese and Japanese speakers with the change of vowel duration. (Zhang & Hayashi, 2015, page 3)

It was found in the interviews after the experiment that most Japanese speakers can notice the differences between each stimulus and they have clear criteria to help them to make a choice. Some said they have a clear standard for making the decision, as they chose the voiceless sound when the vowel sounded shorter and chose the

voiced sound when the vowel sounded longer. This identification strategy is thought to come from their native language. In Japanese there is a contrast between long vowels and short vowels, and also a contrast between words with geminate consonants and words without geminate consonants. It is known that geminates can change the duration of the preceding vowel and even the posterior vowel. This is thought to be the reason why Japanese speakers can notice the change of vowel duration easily.

On the other hand, most Chinese speakers said that they cannot tell the difference between the stimuli in each group when they were asked how they made their choice. Some said that they could perceive the difference to some extent but they could not tell exactly what part was different. This suggests that Chinese speakers are less sensitive to vowel duration and it is difficult for them to find the connection between vowel duration and the voicing feature of the consonant.

This finding may account for their relatively poor performance in Experiment 1 for identifying voicing features of syllable-final stops. As discussed earlier, vowel duration difference is phonemic in Japanese but not in Mandarin Chinese. This is considered to be a negative transfer from their first language.

As mentioned before, it is clear that both in the perception of Mandarin and foreign languages, Mandarin native speakers do not regard vowel duration as an effective cue. Besides the difficulties in perception, Mandarin speakers show difficulties in production as well. It is reported that Mandarin native speakers often omit an English final consonant or insert a vowel after the final consonant (Flege, 1989).

Flege et al. (1992) examined the production of word-final English /t/ -/d/ contrast by native speakers of English, Mandarin and Spanish. In the study, English native speakers and non-native speakers produced English minimally paired CVC words.

Each talker read four lists in which 34 randomized CVC words occurred at the end of the carrier phrase " I will say _."

Then English speaking listeners were asked to identify the voicing feature in syllable-final stops produced by speakers in five groups: English native speakers, experienced and inexperienced Spanish-speaking English learners, and experienced and inexperienced native Mandarin speaking English learners. All of the nonnative subjects have learned English as an L2 in adulthood.

Acoustic analyses showed that both experienced L2 groups differed from the native English speakers; and the final stops produced by the experienced L2 learners were identified no better than those produced by relatively inexperienced L2 learners.

As a result, there is a significant difference between the identification rates of stops produced by experienced second language learners compared to stops produced by the native English speaker. Meanwhile, no significant difference in identification rate was found between the stops produced by experienced second language learners and the stops produced by the inexperienced L2 learners.

Moreover, it is also revealed that the productions of native English speakers show some significant acoustic features. For instance, they produced vowels significantly longer before /d/ than /t/. Although the same kinds of acoustic differences between /t/ and /d/ were also found in the production of the L2 learners, the difference is relatively small. Finally, there is a strong correlation between vowel duration differences and the listeners' identification under the edited pronunciation condition (where post-vocalic cues were removed) than under the full pronunciation condition.

This study not only supports that English native speakers rely on vowel duration to perceive syllable-final voicing, but also reveals that the preceding-vowel duration

effect in Mandarin speaker's production is smaller compared to English native speakers.

In second language acquisition, production and perception are closely connected. The fact that Chinese speakers could not make use of the vowel duration cue efficiently and also could not produce vowel duration cues in their own pronunciation added another piece of evidence for the statement. This will potentially cause misunderstanding in their communication with English native speakers. It is important to find an effective training method.

2.5. Previous Studies on Distribution Learning

Although it is still not clear how the sensitivity is obtained, it is possible that distributional learning is also involved in the process of acquisition. Distributional learning is defined as learning by simply listening to the frequency distributions of the speech sounds in the environment. It is the learning mechanism that human beings use to acquire speech sounds in their early life.

Wanrooij et al. (2014a) investigated the effect of fast distributional learning in the lavatory. In the study, Dutch infants aged 2 to 3 months were presented with either a unimodal or a bimodal vowel distribution based on the English /ɛ/-/æ/ contrast for only 12 minutes. Because this contrast does not exist in Dutch, Dutch adults often find it difficult to master. The two vowels have a major difference in their formant frequency.

In the experiment, half of the infants were exposed to a unimodal distribution while the other half of the infants were exposed to a bimodal distribution. The stimuli were synthesized in Praat and they only varied in the values of F1 and F2. F3, F4 and

F5 were set to the same value for each stimulus. The duration of each stimulus was also kept at 100ms.

In unimodal distribution, a large number of different vowel tokens with F1 values centering around 11.47 ERB were presented to the infants. Meanwhile, in bimodal distribution, most of the tokens have F1 values typical of English [æ] and [ɛ], which center around 10.44 and 12.50 ERB. It is presumed that the bimodal distribution suggests the existence of a contrast between [æ] and [ɛ], while the unimodal distribution does not.

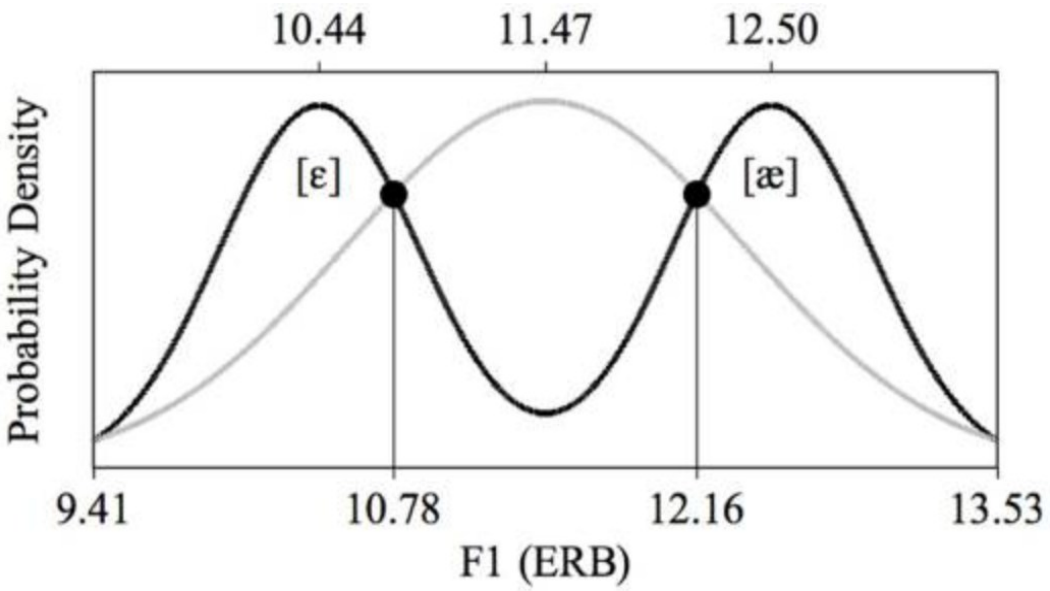


Figure 7 | Unimodal (gray) and bimodal (black) training distributions of the first vowel formant (F1). The values of the test stimuli lie at the intersections of the two distributions. (Wanrooij et al., 2014a, page 2)

After the training, their ability to discriminate [æ] from [ɛ] was tested in an ERP oddball paradigm. If distributional training is effective, infants trained bimodally should have a better performance in discriminating the two test stimuli. The result

showed that the infants exposed to bimodal distribution can better discriminate /ɛ/-/æ/ contrast after the training than the infants exposed to unimodal distribution.

Wanrooij et al. (2014b) further examined whether the capacity for using the mechanism is different in adults than in infants. In the experiment, the participants were Dutch native speakers that had been raised monolingually. The subjects were exposed to either a bimodal distribution that suggested the existence of the two English vowels, or to a unimodal distribution that did not. Before and after exposure the participants were tested on their discrimination of a representative [æ] and a representative [e], by measuring mismatch responses (MMRs). As a result, the significant effect of bimodal distribution in infants observed by Wanrooij et al. (2014a) was not shown in the adult subjects.

Ong et al. (2015) examined whether or not non-tone language listeners can acquire lexical tone categories through distributional learning.

In the study, native Australian English listeners were trained on a Thai lexical tone minimal pair. None spoke a tone language. The stimuli were produced by four native Thai speakers, which were two syllables /kha/ and /na/, each produced with two different tone types (tone 33 and tone 241). After that, a series of sounds morphing from tone 33 to tone 241 was created.

During Training, the subjects were randomly assigned to one of the distribution conditions. Each condition group further consists of four sub-groups based on the training minimal pairs. It is suggested that a unimodal distribution would induce a single central category, while a bimodal distribution would induce two separate categories.

A discrimination task was conducted before and after training. It is found that the bimodal group outperformed the unimodal group at post-test relative to at pre-test.

These results suggest that distributional learning is also an effective training method for adults, and non-tone language listeners are able to learn lexical tones through distributional learning.

Furthermore, Escudero & Williams (2014) conducted a longitudinal study to examine both short-term and long-term effects of distributional learning of phonetic categories on non-native sound discrimination over a 12-month period. 79 Spanish native speakers who were living in the Netherlands participated the experiment, and they were assigned to one of the three groups: bimodal group, enhanced groups or music group.

During the exposure phase, the bimodal group heard a bimodal vowel continuum, which consisted of eight synthetic vowel tokens generated based on the Dutch /a:/-/ɑ/ contrast. In the enhanced group, the participants were presented with Dutch vowels with exaggerated quality differences.

It was also presumed that a larger spectral difference between the vowels in the Enhanced distributions will be perceptually more salient for non-native listeners and may lead to a larger effect of training than the smaller difference in the Bimodal distribution.

During the experiment, the participants were exposed to vowel distributions lasting for two minutes, whereas the Music groups listened to classical music for the same amount of time. Before and after the exposure, the participants received an identical categorization test.

The categorization test was a two-alternative forced choice categorization accuracy of the vowels in the Dutch /a: - ɑ/ contrast. On each trial, they listened to three vowel sounds and were asked to decide if the first sound was more like the second or the third.

The participants attended the experiment for three times over 12 months, separated by a period of six months. In the last session, they performed a categorization test once and did not receive further exposure to distributions.

The result showed that discrimination by listeners in the two distribution groups improved immediately after the short exposure. In the first session, the Enhanced group's improvement scores were significantly higher than those of the Music group. For session 2 improvement, the Bimodal group improved more than the Enhanced group and the Music group.

Moreover, by comparing the difference between improvement after six and 12 months, it showed that improvement was maintained after six and 12 months, demonstrating that distributional learning has long-lasting effects.

Pająk & Levy (2012) conducted an experiment on adults with different language backgrounds. In the study, native speakers of Korean and Mandarin were tested. In Korean, duration is a cue to distinguish phonetic categories (e.g., [pul] 'fire' vs. [pu:l] 'blow'). Therefore, they are better at discriminating duration contrasts than speakers of Mandarin, in which there are no length contrasts. While Mandarin speakers are considered to be better at distinguishing place of articulation contrast between alveolo-palatal and retroflex sibilants.

Mandarin speakers and Korean speakers participating in the experiment were exposed to the sound distribution in a hypothetical language that could be interpreted as either a place distinction (alveolo palatal vs. retroflex) or a duration distinction (short vs. long).

The results showed that the effect of training on length is better for Korean speakers, which suggests that speakers of Korean were biased toward inferring a length-based category distinction and against inferring a place-based category

distinction. On the other hand, no effect of training was found in Mandarin native speakers for both length-based category distinction and length-based category distinction.

3. Experiment 1 – Production of Syllable-final Voicing

In second language acquisition, perception and production are closely related. Therefore, before diving into perception, it is necessary to check the production of syllable-final stop voicing by Mandarin native speakers.

In this experiment, Mandarin speakers were compared to English native speakers in their production of English syllable-final stop contrasts, with a main focus being on the two acoustic features of preceding vowel, which are vowel duration and F1 offset frequency.

3.1. Reading Material

Six English words were used as the reading material in this experiment. They are simple CVC words beginning with /læ/ and ending with six stop consonants /p/, /b/, /t/, /d/, /k/ and /g/.

These words were used so that the phonemic environments are the same, leaving only the contrast between voiced and voiceless stops.

3.2. Participants

Ten Mandarin Chinese native speakers (age M= 28.8, SD = 2.39) and ten English native speakers (age M= 29, SD = 2.94), including six native speakers of American English and four native speakers of Canadian English, participated in this experiment. All Mandarin speaking participants have learned English as a second language after the age of 14.

3.3. Procedures and Measurements

The participants were asked to read the six words in the order of / læp/, /læb/, /læt/, /læd/, /læk/ and /læg/, and each word was read for four times. Their voices were recorded, and in total, there are 24 samples from each participant. Acoustic analyses were conducted by using Praat (Boersma & Weenink, 2006), and the two important acoustic features of the preceding vowel were measured, which are duration value and F1 offset frequency value.

3.4. Results

The mean value of vowel duration produced by each participant is calculated, and the vowel duration values in voiced condition are compared to the duration values in voiceless condition.

In the production by English native speakers, paired sample t-tests show significant differences in duration between voiced conditions and voiceless conditions, and this is the case for all three consonant types ($t(9) = 9.5$, $p < .0001$, for /p/-/b/, $t(9) = 7.12$, $p < .0001$, for /t/-/d/, and $t(9) = 8.16$, $p < .0001$, for /k/-/g/).

In the case of Mandarin native speakers, however, paired sample t-tests show no significant differences in duration between voiced conditions and voiceless conditions, and this is the case for all three consonant types ($t(9) = 1.7$, $p = 0.12$, for /p/-/b/, $t(9) = 0.68$, $p = 0.5$, for /t/-/d/, and $t(9) = 2.18$, $p = 0.06$, for /k/-/g/).

This suggests that most of the participants produce vowels of a similar length no matter whether or not the following consonant is voiced or voiceless. In some cases, the vowel duration is even longer for voiceless stops than voiced stops.

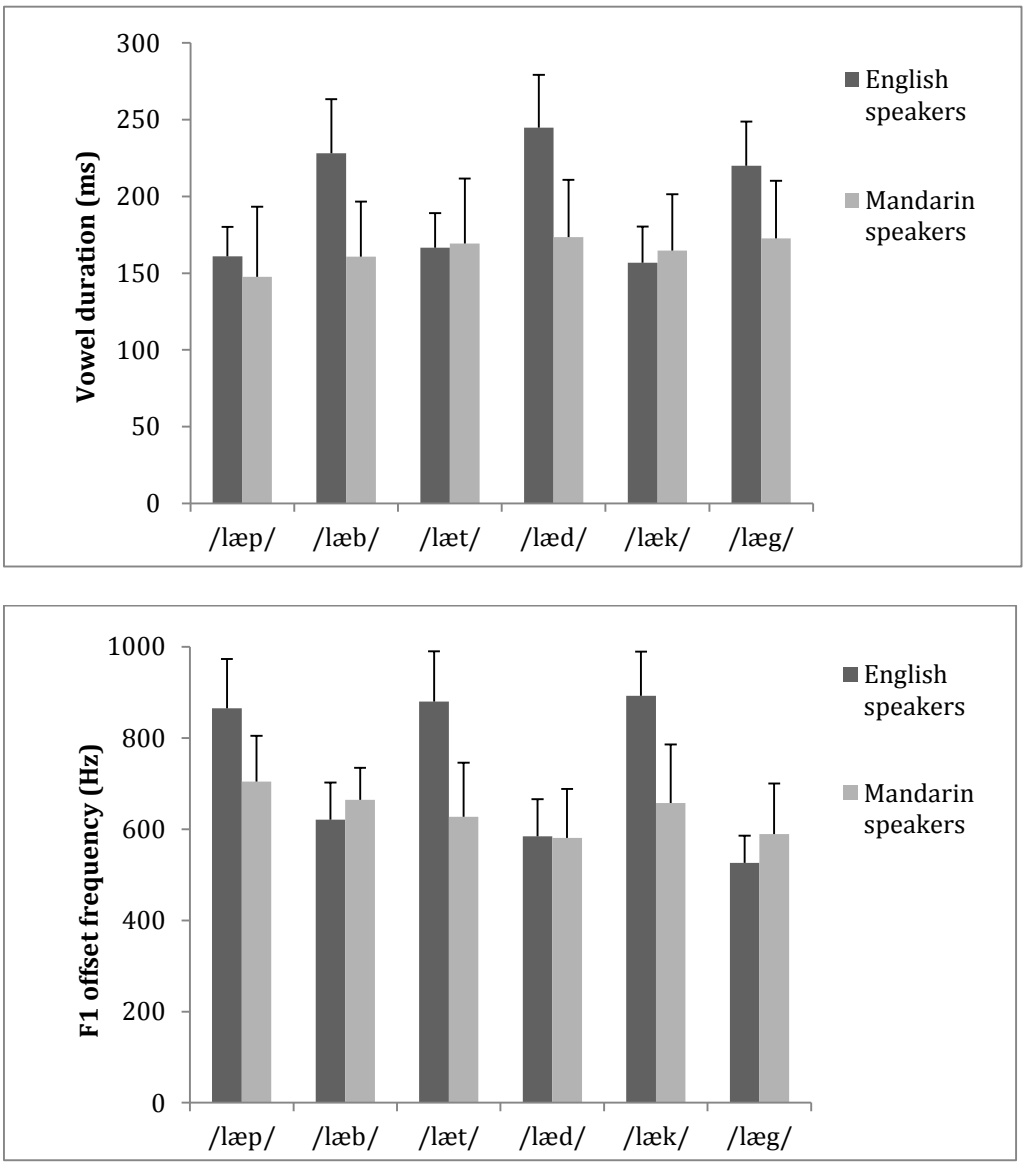


Figure 8 | Mean values of vowel duration (top) and F1 offset frequency (bottom) produced by English native speakers and Mandarin native speakers.

Two-way ANOVA tests were then conducted to all vowel duration data from English native speakers and Mandarin native speakers，where voicing (voiced vs. voiceless) and L1 (English vs. Mandarin) were the variables. The result showed a significant interaction between the two factors voicing (voiced vs. voiceless) and L1 (English vs. Mandarin), and this applies to all consonant types ($F(1, 36) = 5.7$, $p =$

0.02, for /p/-/b/, F (1, 36) = 10.8, p = 0.002, for /t/-/d/, and F (1, 36) = 7.1, p = 0.01, for /k/-/g/). This shows that the difference in vowel duration between voiced conditions and voiceless conditions is significantly larger for English native speakers, which indicates that English native speakers produce significantly more salient vowel duration cues conditioned by syllable-final voicing.

When it comes to F1 offset frequency features produced by English native speakers, there is a significant difference between voiced condition and voiceless conditions, and it applies to all stop consonant types (t (9) = -9.63, p < .0001, for /p/-/b/, t (9) = -10.3, p < .0001, for /t/-/d/, and t (9) = -11.3, p < .0001, for /k/-/g/).

In the production by Mandarin native speakers, there is also a significant difference between voiced condition and voiceless conditions, and it applies to all stop consonant types (t (9) = -2.95, p = 0.016, for /p/-/b/, t (9) = -5.19, p = 0.0005, for /t/-/d/, and t (9) = -2.42, p = 0.038, for /k/-/g/). This suggests that in the production of Mandarin native speakers, the offset frequency value has been proven to be significantly lower in the voiced conditions.

However, two-way ANOVA tests conducted on both English native speakers and Mandarin native speakers showed a significant interaction between the factors *voicing* (voiced vs. voiceless) and *L1* (English vs. Mandarin) for all three consonant types (F (1, 36) = 12.4, p =0.001, for /p/-/b/, F (1, 36) = 13.9, p =0.0006, for /t/-/d/, and F (1, 36) = 21.5, p =0.00005, for /k/-/g/). This again indicates that although Mandarin speakers produce lower F1 offset frequencies in voiced conditions; the difference between voiced conditions and voiceless conditions is still quite small compared to English native speakers. It seems that English native speakers produce more salient F1 offset cues in production.

In the production of voiced stop consonants by Mandarin native speakers, it can be

seen that the closure voicing (which can be indicated by a voice bar) is less salient compared to English native speakers. On the other hand, the aspiration is much stronger in the case of voiceless stops compared to English native speakers. This suggests the possibility that Mandarin speakers produce stop contrast in syllable-final positions in the same way as their counterparts in syllable-initial positions.

3.5. Discussion

It is reported that English native speakers make use of vowel duration cues and F1 offset cues to perceive syllable-final stop voicing (Hillenbrand et al., 1984; Raphael, L. J., 1971); English learning infants also start to control vowel durations conditioned by voicing in their speech production before the age of two (Ko, 2007). It is considered that being able to produce these cues is crucial during the communication with English native speakers.

Nevertheless, the result of the present experiment has shown that in the production of syllable-final stops by Mandarin native speakers, preceding vowel durations tend to remain the same for both voiceless and voiced sounds. In this aspect, we can say that Mandarin native speakers did not acquire a native-like pronunciation.

It has been reported in the previous study (Zhang & Hayashi 2015) that Mandarin Chinese speakers show less sensitivity to the preceding vowel duration cues when perceiving syllable final voicing. This is probably due to the fact that there is no long and short vowel contrast in Mandarin Chinese. The result of this production experiment also suggests that the influence of their first language also appears in the production of a second language. In second language acquisition, production and perception are closely connected. The fact that Chines speakers could not make use of

vowel duration cues efficiently and also they could not produce vowel duration cues in their own pronunciation added another evidence for the statement.

When it comes to F1 offset frequency, the result shows that there is a significant difference between F1 offset values in voiced conditions and voiceless conditions, which suggests that the participants tend to produce a lower F1 offset value when the final consonant is voiced. However, after the comparison with English native speakers, we can see that the difference in F1 value might be too small to serve as an efficient acoustic cue.

The results of the present experiment aligns with Flege et al. (1992), which states that Mandarin speaking English learners also shows acoustic features in their speech, while the difference is much less salient when compared to English native speakers. This suggests a possibility that the participants tend to pronounce the same vowel in both cases; the low F1 offset value in voiced conditions might just be due to the potential influence of articulation.

It was reported that Mandarin speaking Japanese learners have difficulties producing phonemic long and short vowels (Kurihara, 2005). It is considered that controlling duration in production is a difficult task for Mandarin native speakers in second language studies.

From the result, we can also see that Mandarin speakers tend to produce stop contrast in syllable-final position the same way as their counterparts in syllable-initial position. Namely, they produce no closure voicing or less salient closure voicing in the case of voiced consonants. This aligns with the report from Flege (1995), which states that Mandarin native speakers usually could not produce perceptible stop voiced contrast. It is evident from some previous studies that Chinese learners of English devoice /b, d, g/, delete both voiced and voiceless stops, or add schwa-like vowels

following word-final stops, which is similar to Japanese learners.

That is to say, when English native speakers listen to a minimal pair ending with a stop contrast produced by mandarin native speakers, the vowel duration, which is the primary cue expected by them, is usually missing. Not mentioning other vague voicing features. This will potentially cause misunderstanding in their communication with English native speakers.

Meanwhile, some studies have suggested that the perception ability of foreign sounds, as well as the learners' pronunciation of foreign sounds can be improved through training. Since secondary cues like vowel duration and F1 offset value are important in perceiving final stop voicing, especially in unreleased cases. It is necessary to examine whether or not the sensibility of Chinese speakers towards these secondary acoustic cues can be obtained by training, and also if this can have positive influence on the production.

Therefore, the next part of the study will focus on the perception, and it will investigate the effect of distributional learning of vowel duration cues to syllable-final stop voicing.

The aim of the experiment is to find out whether or not distributional learning can help mandarin native speakers to not only gain sensitivity to vowel duration differences but also to build up the association between durational information and syllable-final voicing features.

4. Experiment 2: One - dimensional Distributional Learning

This perceptual experiment was conducted to examine whether or not Mandarin speakers can acquire secondary acoustic cues through distributional learning, with a focus on preceding vowel duration conditioned by syllable-final voicing. In order to achieve this research goal, the stimuli used in the study will be a continuum changing in the dimension of duration only.

4.1. Participants

Forty Mandarin Chinese native speakers (age M = 28, SD = 2.79) participated in the experiment. All participants reported normal hearing and all had learned English as an L2 after the age of 14. Cantonese speakers and Shanghai dialect speakers were not included. All subjects received the same discrimination test in the pre-test phase and post-test phase. In the training phase, the participants were assigned randomly to one of the four groups: released bimodal group, released unimodal group, unreleased bimodal group and unreleased unimodal group. Thus, half of the participants were exposed to a bimodal distribution and the other half were exposed to a unimodal distribution. Likewise, half of the participants were presented with released stimuli, and the other half were exposed to unreleased stimuli.

4.2. Stimuli

Two CVC words /mat/ and /mad/ were recorded by a male native speaker of English in a soundproof laboratory, and they were used as the sources to create the

stimuli. The recordings were then edited with the software Praat (Boersma & Weenink, 2006; version 6.0.21). Recordings with a steady pitch contour were chosen as the released stimuli. The unreleased stimuli were made by eliminating the syllable-final stop consonants /t/ and /d/.

Vowel duration was the parameter in this experiment, and it should be shorter when preceding a voiceless stop consonant and longer when preceding a voiced stop consonant. The duration was controlled by shortening the duration of the vowel in /mad/ by several steps. In order to keep a smooth pitch transition after the editing, every step one period from each steady part of the vowel was cut off. At the same time, cutting two adjacent periods was avoided during the editing.

Stimulus	Vowel Duration (ms)
S1	345
S2	335
S3	325
S4	315
S5	305
S6	295
S7	285
S8	275
S9	265
S10	255
S11	245
S12	235
S13	225
S14	215
S15	205
S16	195
S17	185

Table 1 | Vowel duration of stimuli in the *mat-mad* continuum.

Table 1 shows the vowel duration of each stimulus in the experiment. Stimuli used

in the unreleased bimodal group and the unreleased unimodal group were from the [mat̚] - [ma:d̚] continuum with vowel duration ranging from 185ms to 345ms. Stimuli used in the released bimodal group and released unimodal group were from the [mat̚_t] - [ma:d̚_d] continuum with vowel duration ranging from 185ms to 345ms and /C/ changing from fully released /t/ to fully released /d/ when the vowel duration increased to 265ms (stimulus S9).

The stimuli were created in such a way that voiceless final /t/ always occurs with relatively shorter vowels and voiced final /d/ always occurs with relatively longer vowels. As the most controversial duration, stimulus S9 was chosen as the stimulus where the consonant changed from voiced sound to voiceless stop, based on the advice of two native English speakers. Therefore, a total of 18 released stimuli and 17 unreleased stimuli were created.

4.3. Procedures

The experiment consists of three phases: pre-test phase, training phase and post-test phase. During the experiment, the Stimuli were presented to the subjects in Praat ExperimentMFC.

Pre-test Phase and Post-test Phase

In the pre-test and post-test phase, all subjects received the same discrimination test, which was conducted by using the unreleased version of two training stimuli.

In the test, the subjects were presented with two stimuli /ma/ S7 and S11 with different vowel durations (245ms and 285ms). They were asked to listen to two words from a new language, and then to determine whether the two words they heard were

the same. After the subjects made the judgment, the next pair of stimuli was presented automatically. This procedure was repeated 120 times; each time the two stimuli were presented in a randomly generated order. There is a total of four possible orders: S7-S11, S11-S7, S7-S7, S11-S11.

The two test stimuli S7 and S11 were chosen because they had been presented in the same frequency in both bimodal distribution and unimodal distribution during the training phase. Therefore, the four groups of subjects came to the test phase with equal prior exposure in the training phase to sounds in the region of the test stimuli, so that any difference between the groups observed in the test could not be attributed to differences in familiarity with the test stimuli.

Training Phase

In the training phase, the subjects were divided into four groups: released bimodal group, released unimodal group, unreleased bimodal group and unreleased unimodal group. For the released bimodal group and released unimodal group, the stimuli were from a continuum changing from [mat̚_t] to [ma:d̚_d]. For the unreleased bimodal group and unreleased unimodal group, the continuum changing from [mat̚] to [ma:d̚] was used. The only difference between the two sets of stimuli was whether the final stop consonant is released or not.

Before the training, participants were instructed to listen to the vowels carefully, because they would perform a second task (post-test) similar to the first one (pre-test) after the training.

During the training, each stimulus was repeated multiple times. The total number of presentations of stimuli was 472 for all groups, while the frequency distribution of the stimuli was different in the bimodal group and the unimodal group. It is presumed

that the unimodal distribution induces a single central category, while the bimodal distribution induces two separated categories.

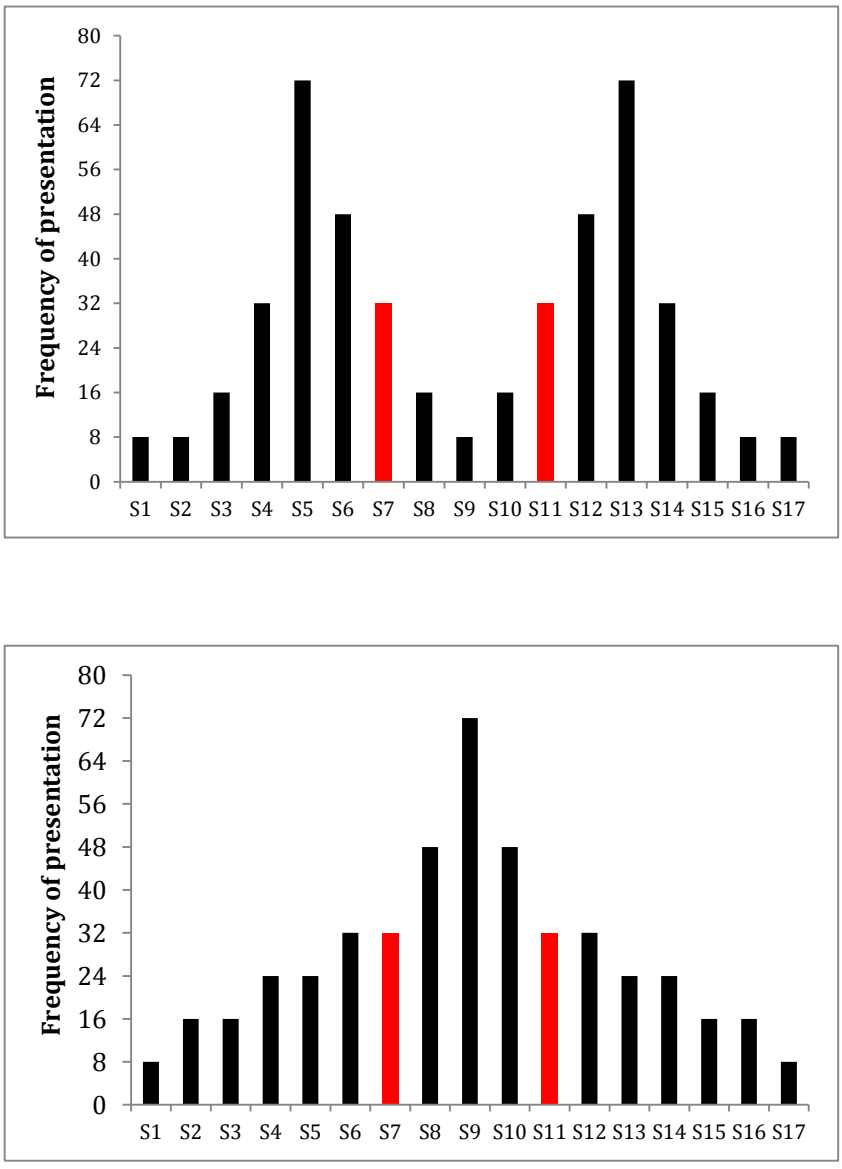


Figure 9 | Frequency of presentation of stimuli in bimodal groups (top) and unimodal groups (bottom). The values of the test stimuli are marked in red color.

Figure 9 shows the frequency of presentations of stimuli in bimodal groups (top) and unimodal groups (bottom). In bimodal distribution, stimuli S5 and S13 occur most

frequently, which is 72 times each. The stimuli S1, S2, S9, S16 and S17 occur the least frequently, which is 8 times each. By contrast, the unimodal mean value lays exactly in the middle of the range of vowel duration values and in between the two bimodal mean values. The most frequently occurring stimulus was the S9 and the least frequently occurring ones were S1 and S17.

The stimuli were randomized and presented with Praat based on two different distribution scripts. The order of presentation of the stimuli in the training was randomized separately for each subject. The subjects were told to listen to a series of words, and during the training they were not required to make any operations on the computer.

The scores of each subject in pre-tests and post-tests were calculated. Comparing only the result of bimodal groups and unimodal groups could not lead to a conclusion on whether or not the training would be effective for extrinsic vowel duration difference acquisition. The score difference does not necessarily suggest that the subjects can better associate the vowel duration differences with syllable-final voicing. It might simply be an indication that the training phase increased the subjects' sensitivity to vowel duration. In order to exclude that possibility, the result of the groups trained with released stimuli will be compared to that of the groups trained with the unreleased stimuli. In those two groups, the training only aims at increasing sensitivity to phonemic vowel duration differences. This is why the two unreleased groups are necessary in this experiment.

Therefore, instead of training on acquiring secondary acoustic cues directly, this experiment is designed to examine whether or not distributional learning on vowel duration differences is more effective when syllable-final voicing features are also available in the stimuli.

By comparing the unimodal groups and bimodal groups, we can see if there is a bimodality effect; and by comparing the unreleased groups and released groups, we can see if the training is more effective when the final voicing feature also appears.

In the experiment, the bimodality effect in both released and unreleased conditions were examined. The main point we want to investigate is whether or not the bimodality effect is stronger when training with released syllable-final consonants compared to training with unreleased syllable-final consonants. Therefore, besides the difference between pre-test and post-test and the difference between bimodal and unimodal, the difference between released and unreleased is also an indispensable factor in this study.

We expected to see a statistically significant triple interaction among *test condition (pre-test vs. post-test), distribution type (bimodal vs. unimodal)* and *stimulus type (released vs. unreleased)*. Assuming that the *distribution type (bimodal vs. unimodal)* has an effect on the score across different *test type (pre-test vs. post-test)*, this interaction could reveal that *stimulus type (released vs. unreleased)* has an effect on this *distribution type effect*. Only when this interaction is significant, we can say that training for duration is more effective with released stimuli.

4.4. Results

Figure 10 gives the pre-test and post-test percentages accuracy (i.e., the percentage of correct judgment of the 120 test stimuli) for the participants in all four groups. A one-way ANOVA test on all pre-test scores did not display a significant difference between the four groups, $F(3, 36) = 0.29$, $p = 0.82$. This suggests the equality of the groups before training.

The difference between pre-test and post-test score is a measure of improvement after training. Paired sample t-test was conducted to the pre-test scores and post-test scores for all four groups. The results showed that the subjects in released bimodal group see a significant improvement after the training phase, t (9)= 3.95, p = 0.003, and so are the subjects in the unreleased bimodal group, t (9) = 3.34, p = 0.008.

On contrary, there is no significant improvement in the subjects of the released unimodal group, t (9) = 0.51, p = 0.62, and the unreleased unimodal group, t (9) = 0.12, p = 0.90.

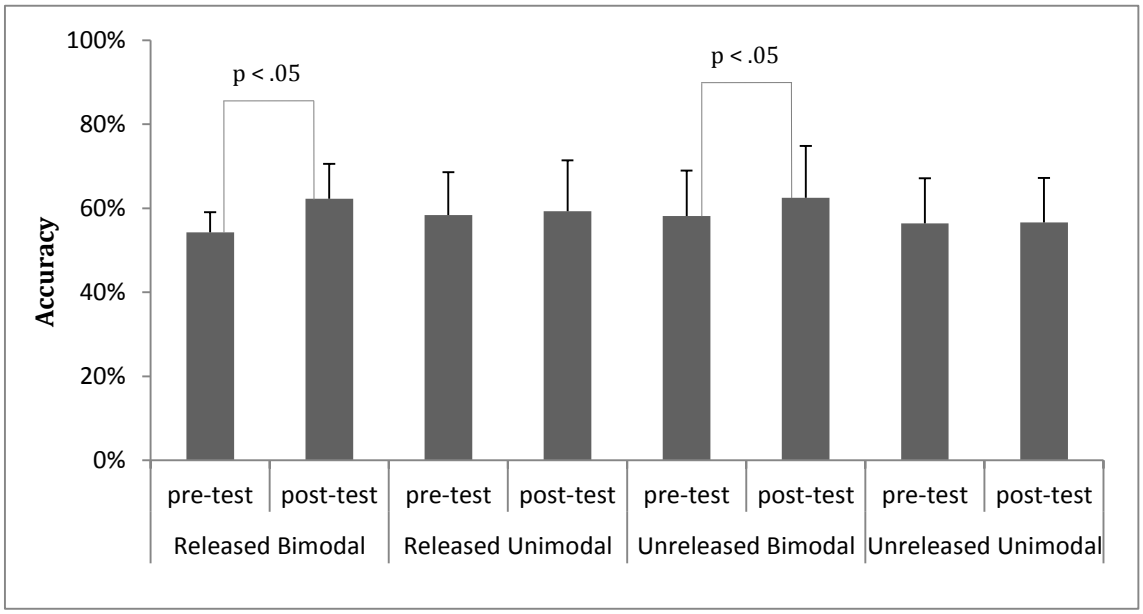


Figure 10 | Pre-test and post-test percentages correct for all four groups.

A three-way ANOVA test was then conducted on all data in the four groups, where *test condition (pre-test vs. post-test), distribution type (bimodal vs. unimodal) and stimulus type (released vs. unreleased)* were the variables. The results revealed a main effect of *test type (pre-test vs. post-test)*, and the percentages of post-test people score was improved by 14 percent compared to the odds of pre-test people, F (1, 72) = 10.7,

p = 0.0016. There is no significant effect of *released type (released vs. unreleased)*, F (1, 72) = 10.7, p = 0.85, or *distribution type (bimodal vs. unimodal)*, F (1, 72) = 10.7, p = 0.65.

There was also an interaction effect of *test type (pre-test vs. post-test)* and *distribution type (bimodal vs. unimodal),* and the average of pre-test unimodal people and post-test bimodal people scored significantly better than the average of pre-test bimodal people and post-test unimodal people, F (1, 72) = 8.2, p = 0.0055. This suggests that the participants trained bimodally made more improvements compared to those who were trained unimodally.

However, no significant triple interaction was found among *distribution type (bimodal vs. unimodal)*, *test type (pre-test vs. post-test)* and *stimulus type (released vs. unreleased)*, F (1,72) = 0.046, p = 0.83. For that reason, although bimodality effect is confirmed, the possibility that the score improvement observed in released bimodal group is only in the phonemic level cannot be excluded. Therefore, we cannot draw a conclusion that training with released stimuli is more effective.

4.5. Discussion

In order to investigate the effectiveness of distributional learning of extrinsic vowel duration differences, an experiment was conducted on Mandarin Chinese native speakers. In the experiment, the subjects were divided into four different groups: released bimodal group, released unimodal group, unreleased bimodal group and unreleased unimodal group. A significant three-way interaction among *test type (pre-test vs. post-test), distribution type (bimodal vs. unimodal)* and *stimulus type (released vs. unreleased)* was expected in order to show that the training was effective.

The result showed significant improvement before and after training in bimodal groups compared to unimodal groups. However, training with released stimuli did not significantly improve the subjects' perception compared to training with unreleased stimuli.

As mentioned above, the bimodal released group is the main group investigated. The result showed that the subjects trained in this group had indeed a better performance in post-tests. However, the improvement observed in released bimodal group could not lead to the conclusion that the training is effective for acquiring extrinsic vowel duration differences. The possibility that the training only improved the sensitivity to phonemic vowel duration difference must be excluded.

For the subjects in unreleased groups, the subjects trained with bimodally distributed stimuli were also significantly better in identifying vowel duration differences after training than the subjects trained with unimodal distributed stimuli. However, the result revealed no significant effect of the factor stimulus type. It still remains to be seen whether or not training with released stimuli is easier for the subject to associate the durational change with the final consonant compared to training with unreleased stimuli.

Most subjects reported the experiment as a difficult task. In the pre-test phase, some Mandarin Chinese speakers reported that duration difference was not taken into consideration in discriminating the two words; some reported that they could not accurately perceive the durational difference. Although after the training they were able to notice that the difference lies in duration, making the correct judgment remains a difficult task for them.

This might partially be the result of the selected test stimuli. In the discrimination test, stimuli with vowel duration 245ms and 285ms were chosen as the test stimuli. It

is possible that the durational difference between the two stimuli is too small while the effect of training is limited.

Wanrooij et al. (2014b) has indicated that distributional learning is less effective in adults than infants. As a weaker learning mechanism, it is possible that more time exposed to the stimuli is needed to see a training effect on adults. It also remains to be seen whether or not the same type of training can help infants at an early age to acquire extrinsic vowel duration differences.

The experiment design also has potential influence on the training effect, because a pre-test would have been an additional distributional training and this could have distorted the intended training distributions. This also needs to be taken into consideration in further experiment design. Moreover, it can be seen from the result that there is a big personal difference between the participants; more participants are needed in order to achieve a more accurate result.

4.6. Follow-up Experiment

A follow-up experiment was conducted in order to achieve a more accurate result. Because bimodality effect has been confirmed in the previous experiment results, as a follow-up to experiment 2, this experiment continues to investigate whether or not distributional learning is more effective when the syllable-final voicing feature is also presented in the stimuli. Therefore, the participants were randomly assigned to a bimodal released group and a bimodal unreleased group.

Ten additional Mandarin native speakers were recruited for this experiment. Combining the original 20 participants in experiment 2, there was a total of 30 participants (age M = 27.7, SD = 3.15). All have reported normal hearing and have

learned English as a second language since the age of 14.

The experiment was conducted following the same procedures as experiment 2.

4.7. Results and Discussions

Ten Mandarin native speakers participated in this follow- up experiment. Adding the original 20 participants in experiment 2. In total, two sets of result data from 15 Mandarin speakers each was gathered for bimodal released group and bimodal unreleased group respectively. A two-way ANOVA test was conducted on the data, and a significant interaction between *test-type (pre-test vs. post-test)* and stimulus type *(released vs. unreleased)* was expected in order to confirm that distributional learning of extrinsic vowel duration is effective.

First, the result showed significant difference between pre-test scores and post-test scores, and the score of post-test was improved by 6.9 percent compared to the odds of pre-test, $F(1, 56) = 8.19$, $p = 0.006$. This means the accuracy scores were significantly improved after the training phase. And also, there is significant function of *stimuli type (released vs. unreleased)*, and the score under released condition was improved by 5.1 percent compared to the score under unreleased condition, $F(1,56) = 4.52$, $p = 0.038$.

Moreover, there is a significant interaction between the two factors *test type (pre-test vs. post-test)* and *stimuli type (released vs. unreleased)*, $F(1, 56) = 4.06$, $p = 0.048$. That is to say, the participants trained with released stimuli see a significantly bigger improvement in post-test compared to pre-test.

In experiment 2, it was confirmed that training is more effective for bimodal groups than unimodal groups. Combining with the results achieved in this follow-up experiment, we can say that training is also more effective for released groups

compared to unreleased groups. Therefore, distributional learning is an effective method in increasing Mandarin speakers' sensitivity to secondary acoustic cues.

This result indicates that distributional leaning is not only effective in the phonemic level but also on the phonological level. The subjects can integrate the acoustic feathers in speech sounds and find out the underlining phonological rule.

The next step of the study will investigate whether the effect of distributional learning can also expand to multiple dimensions. As was mentioned above, in the perception of syllable-final stop voicing, in addition to preceding vowel duration formant transition can also serve as an acoustic cue.

In recent years, the transition pattern of the first formant (F1) in preceding vowel also starts to draw people's attention. The first formant (F1) of vowels has been seen to vary with the voicing of the following consonant, i.e. F1 offset frequency of the preceding vowel usually has a lower value when it is followed by a voiced stop consonant.

Previous studies have shown that English native speakers can also rely on vowel to consonant transition to perceive syllable-final voicing (Hillenbrand et al., 1984). Since vowel duration feature happens on the time domain, while F1 offset value in on the frequency domain. Our next research goal is to find out will training with stimuli changing in both time domain and frequency domain be still effective in acquiring secondary acoustic features.

5. Experiment 3: Two-dimensional Distributional Learning

This experiment is conducted to find out if the effectiveness of distributional learning can extend to multiple dimensions. In order to achieve the research goal, stimuli used in the experiment differ in both vowel duration (time dimension) and F1 offset frequency (frequency dimension).

It is predicted that the training in the vowel duration dimension will not be as effective as experiment 2, since F1 offset frequency will also be a parameter in the stimuli, the changes happening in vowel duration dimension will be less clear to the participants.

5.1. Participants

20 Mandarin Chinese native speakers (age M = 26.5, SD = 3.3) participated in the experiment. There are only two groups in this experiment, namely bimodal released group and unimodal released group. All subjects reported normal hearing and all had learned English as an L2 after the age of 14. Cantonese dialect speakers and Shanghai dialect speakers were not included.

5.2. Stimuli

In this experiment, both the F1 offset frequency values and the vowel duration values varied among the stimuli. In order to control both of the parameters, synthesized words were created.

There were two types of stimuli in the experiment, stimuli used in the training

phase were from a continuum changing from [lat̚_t] to [la:d̚_d]. In pre-test phase and post-test phase, unreleased stimuli [lat̚] to [la:d̚] were used.

First, a continuum of words /la/ were synthesized by using KlattGrid in Praat (Boersma & Weenink 2006, version 6.0.21). F0 was set to start at 180Hz and fell linearly throughout the stimulus to an offset frequency of 150Hz in order to achieve a more natural sound.

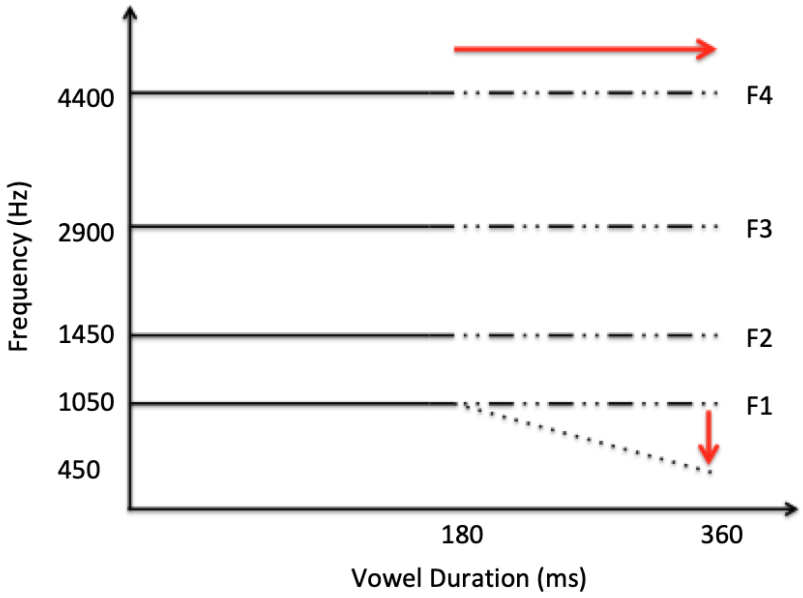


Figure 11 | Synthesized vowel in duration dimension and frequency dimension.

As is shown in Figure 11, the stimuli change in both F1 frequency and duration. In the frequency dimension, the F1 frequency was set to 1050Hz at steady state and the transition pattern was controlled by setting the F1 offset parameter to 10 different values ranging from 450Hz to 1050Hz. F2 was set to 1450Hz, F3 was set to 2900Hz and F4 was set to constant 4400Hz. In order to synthesize the transition from /l/ to /a/, the F1 value ramped up from 460Hz to 1050Hz while F3 ramped down from 3500Hz to 2900Hz during the period of 80ms and 100ms.

In the duration dimension, the duration was controlled by shortening the duration of the vowel in /lad/ by several steps. The duration of the vowel /a/ changes from 360ms to 180ms in 9 steps. In order to keep a smooth pitch transition after the editing, every step one period from each steady part of the vowel was cut off. At the same time, cutting two adjacent periods was avoided during the editing.

Table 2 shows the duration value and F1 offset frequency value for the stimuli used in the experiment. The stimuli are designed in such a way that the vowel preceding a voiced consonant has a relatively long duration and a relatively low F1 offset frequency, and the vowel preceding a voiceless consonant has a relatively short duration and a relatively high F1 offset frequency.

Stimulus	Vowel Duration (ms)	F1 Offset Freqeuncy (Hz)
S1	360	450
S2	340	517
S3	320	583
S4	300	650
S5	280	717
S6	260	783
S7	240	850
S8	220	917
S9	200	983
S10	180	1050

Table 2 | Vowel duration and F1 offset frequency value of each stimulus synthesized.

In order to create the released stimuli, the synthesized /la/ continuum were then connected with /t/ or /d/ waveform, which were retrieved from a natural pronunciation of /lat/-/lad/ contrast by an English native speaker.

5.3. Procedures

There are five phases in the experiment: pre-production test phase, pre-perception test phase, training phase, post-perception test phase, and post-production test phase.

Training Phase

In the training phase, the subjects were divided into two groups: released bimodal group, released unimodal group.

During the training, each stimulus was repeated multiple times. The total number of presentations of stimuli was 400 for all groups, while the frequency distribution of the stimuli was different in the bimodal group and the unimodal group. It was suggested that the unimodal groups induced a single central category, while the bimodal groups induced two separated categories.

Figure 12 shows the frequency of presentations of stimuli in bimodal group (top) and unimodal group (bottom).

In bimodal distribution, the stimuli S3 and S8 occur the most frequently, which is 80 times each. The stimuli S5 and S6 occur the least frequently, which is 8 times each. By contrast, the unimodal peak lay exactly in the middle of the range of vowel duration values and F1 offset frequency values. The most frequently occurring stimulus was S5 and S6 and the least frequently occurring ones were S1 and S10. The stimuli were randomized and presented with Praat based on two different distribution scripts. The order of presentation of the stimuli in the training was randomized separately for each subject.

In this phase, the subjects were exposed to a series of words passively, and during the training they were not required to make any operations on the computer.

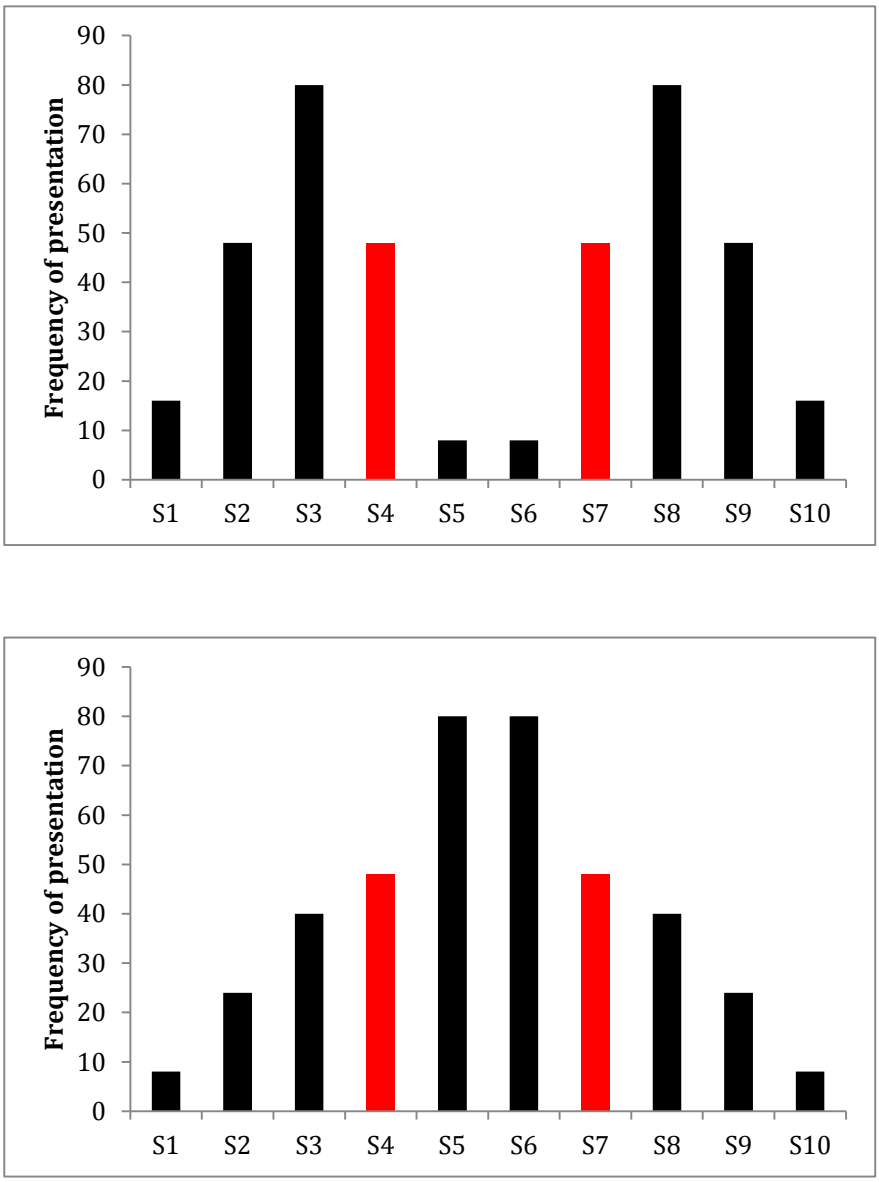


Figure 12 | Frequency of presentation of stimuli in bimodal groups (top) and unimodal groups (bottom). The values of the test stimuli are marked in red color.

Perception pre-test phase and post-test

During pre-perception test phase and post-perception test phase, all subjects received the same discrimination test. In the test, the subjects were presented with a pair of stimuli /la/-/la/ each time, which were a pair of unreleased version of stimuli S4 and S7.

In order to examine the effect of training in both the duration dimension and the frequency dimension, in this discrimination test, each task consists of three blocks. The test stimuli used in block 1 differ in both vowel duration and F1 offset frequency, the test stimuli used in block 2 only differ in F1 offset frequency and the test stimuli in block 3 only differ in vowel duration.

The subjects were then told to listen to two words from a new language, and to determine whether the two words they heard were the same. After the subjects made the judgment, the next pair of stimuli was presented automatically. This procedure was repeated 240 times (80 times in each block), each time in a randomly generated order (combinations of two same words are also included). There is a total of four possible orders: S4-S7, S7-S4, S4-S4, S7-S7.

The two test stimuli S4 and S7 were chosen because they had been presented in the same frequency in both bimodal distribution and unimodal distribution during the training phase. Therefore, the two groups of subjects came to the test phase with equal prior exposure in the training phase to sounds in the region of the test stimuli, so that any difference between the groups observed in the test could not be attributed to differences in familiarity with the test stimuli.

Production pre-test phase and post-test phase

This experiment also checks the effect of distributional learning on production. Therefore, before and after the training phase, each participant's production of syllable-final voicing was also recorded.

The participants were also asked to read six English words printed on a paper, each word was read for four times. The same reading material in experiment 1 is used, which are six words begin with /læ/ and ended with six stop consonants /p/, /b/, /t/, /d/,

/k/ and /g/.

5.4. Results

The participants scored highest (84.1%) in block 1 (frequency dimension and duration dimension). Their score in block 2 (frequency dimension) is a bit lower than block 1 (83.1%), and they have the worst performance (71.4%) in block 3 (duration domain). Moreover, a one-way ANOVA shows that the participants scored significantly lower in block 3 compared to block 2, $F(1, 78) = 10.46$, $p = 0.002$.

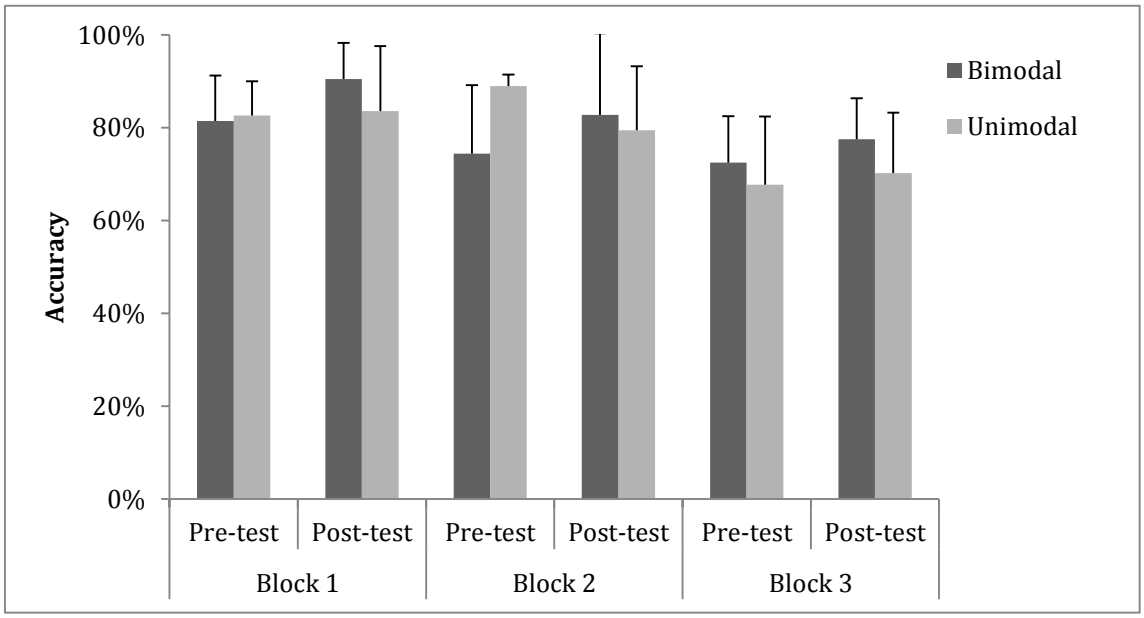


Figure 13 | Mean accuracy scores in discrimination tasks.

Two-way ANOVA tests were conducted where *test-type (pre-test vs. post-test)* and *distribution type (bimodal vs. unimodal)* were the variables. The results revealed no main effect of test-type *(pre-test vs. post-test)* or distribution type *(bimodal vs. unimodal)* in all three blocks. Meanwhile, there was a significant interaction between

test type (pre-test vs. post-test) and *distribution type (bimodal vs. unimodal)* in block 2, F (1, 36) = 4.43, p = 0.04. That is to say, the training is more effective for the bimodal group compared to the unimodal group, which means that bimodality effect is confirmed in frequency dimension.

However, this kind of interaction between the two factors was not significant in block 1, F (1, 36) = 1.60, p > .05, and block 3, F (1, 36) = 0.11, p > .05. This suggests that the effect of two-dimensional distributional training is only effective in the frequency dimension but not in the duration dimension.

When is comes to production, however, no significant difference was found in vowel duration conditioned by voicing after the training (t (9) = 0.74, p = 0.49, for /p/-/b/, t (9) = 1.1, p = 0.29, for /t/-/d/, and t (9) = 2.13, p = 0.76, for /k/-/g/. Two-way ANOVA tests also revealed no significant interaction between voicing (voiced vs. voiceless) and test type (pre-test vs. post-test) (F (1,36) = 0.32, p = 0.57, for /p/-/b/, F (1,36) = 0.08, p = 0.76, for /t/-/d/ and F (1,36) = 0.03, p = 0.86, for /k/-/g/).

Although paired t-tests revealed significant differences in F1 frequency conditioned by voicing after the training (t (9) = -4.13, p = 0.002, for /p/-/b/, t (9) = -3.25, p = 0.009, for /t/-/d/ and t (9) = -3.65, p = 0.005, for /k/-/g/), ANOVA tests again revealed no significant interaction between voicing (voiced vs. voiceless) and test type (pre-test vs. post-test), F (1,36) = 0.03, p = 0.84, for /p/-/b/, F (1,36) = 0.01, p = 0.9, for /t/-/d/ and F (1,36) = 0.17, p = 0.67, for /k/-/g/.

The result indicates that the subjects did not produce a more salient secondary acoustic cue (vowel duration or F1 offset frequency) after exposed to bimodally distributed stimuli.

5.5. Discussion

This experiment examined distributional leaning effect of secondary acoustic cues for syllable-final stop voicing perception in both duration dimension and frequency dimension.

As a result, the participants have a higher score in block 2 (frequency dimension) compared to block 3 (duration dimension). This suggests that Mandarin speakers are more sensitive to frequency information compared to durational information.

After the experiment, most of the participants also reported that discrimination task is relatively easy in block 2, and some of them said they could hear another vowel following /a/ though it is quite difficult to perceive due to the short duration. This suggests that they are able to perceive the change in formant frequency. It is possible that this sensitivity is due to the existence of diphthongs and triphthongs in Mandarin. Since in the case of Mandarin diphthongs and triphthongs, the change of frequency usually indicates a second vowel, and it is crucial to perceive the transition of formant. This is considered to be a positive transfer from L1.

We can also see from the results that the effect of distributional training is significant for block 2 (frequency dimension) while is not significant in block 3 (duration dimension). This indicates that in distributional learning, the participants tend to focus on the information in the dimension that is easier for them.

It is clear that compared with experiment 2, the training effect of vowel duration is weaker in two-dimensional learning compared to one-dimensional learning and it does not seem to have helped building up the association between vowel duration and consonant voicing feature. A possible explanation is that, the durational information becomes less salient in input when appearing together with frequency information. In

this case, it is more difficult for the participants to focus on the changes of vowel durational information. This could also explain why no training effect is observed in block 1, in which the test stimuli differ in both duration and F1 offset frequency. It is possible that frequency feature also becomes less salient when appearing together with vowel duration feature.

Finally, the experiment didn't show the effect of perceptual training on production. The relationship between speech perception and production has always been discussed. Just like a language learner does not need to know how to produce a sound in order to perceive it correctly, it is assumed that production is a process requiring articulation organs to work along, and the knowledge in perception does not seem to be able to expand to production in this study.

6. General Discussion

Zhang & Hayashi (2015) reported that Mandarin speakers have difficulties perceiving unreleased syllable-final stop voicing, and also they are not able to make use of preceding vowel duration cue in perception. The results of experiment 1 have shown that Mandarin native speakers also do not produce different vowel duration values conditioned by syllable-final voicing.

Concerning the nature of vowel length effect in English, whether or not it is a universal process is still under debate. The auditory explanations claim that speakers in a language community produce the vowel length effect as a strategy to perceptually enhance the voicing feature of the consonants by indicating contrast effects between the vowel duration and the closure-duration of consonants (Kluender et al., 1988). The fact that vowel length effect can be observed in many languages while it is most significant in English no doubt suggests the existence of language-specific enhancement. Therefore, in English learning it is also important for the learners to enhance the vowel duration differences in their production. And the precondition is that the leaners are aware of the association between the vowel duration feature and the final voicing features.

In the case of F1 offset frequency value, it can be seen that Mandarin native speakers produce lower F1 offset value when it is followed by a voiced consonant, though the difference is significantly small compared to English native speakers. In experiment 3, Mandarin speakers also had a good performance in differentiating vowels with different F1 offset value, and most participants also reported that the block of frequency is easier than the block of duration.

Language transfer is one of the earliest theories of second language acquisition. In

the study of phonological acquisition, the effect of this transfer is divided into two categories. *Positive transfer* occurs when the L1 and L2 patterns are similar, and habits from the L1 are carried over to the L2. *Negative transfer* occurs when the L1 and L2 patterns are different, and habits carried over interfere with the L2. In the acquisition of vowel duration length effect by Mandarin speakers, negative transfer from L1 occurs. Comparatively, their relatively high sensitivity to vowel frequency information may indicate a positive transfer from Mandarin.

The speech learning model theory (Flege, 1995) has indicated that the phonetic system of native and foreign sounds can affect each other. The hypothesis stated that so long as the learner can perceive part of the different features of a foreign sound, there would be more chances for a new category to be formed, which does not belong to any native sound category. The more different a foreign language sound is from native sounds, the easier it is for learners to identify it. And if two sounds are not in a phonemic contrast with each other in one's native language, it will be very difficult for the listener to tell them apart, even in a foreign language. This hypothesis is also supported by the report by Aoyama et al. (2004) who stated that it is easier for Japanese speakers to improve through perceptual training on /r/ sounds than /l/ sounds, since /r/ sounds do not exist in Japanese and they have more specific features.

This provides a possible explanation to the limited training effect of phonemic vowel duration on Mandarin speakers. Since long vowel and short vowel is not a phonemic contrast in L1 for Mandarin speakers, training on durational differences brings more difficulties than training on complete novel sounds. In spite of the difficulty, we still saw significant improvement in Mandarin native speakers' sensitivity to extrinsic vowel duration after distributional learning in experiment 2. However, the result of experiment 3 further suggests that the training effect might only

appear when vowel duration serves as the only parameter in the stimuli. When another acoustic feature (like F1 frequency) also comes into play, the training effect becomes very limited.

Although many studies have reported distributional learning as an effective way for adults to master foreign speech sounds (Ong et al., 2015; Escudero & Williams, 2014), there are also studies that reported limited effect in adults. For instance, Wanrooij et al. (2014b) conducted the same distributional leaning experiment on Dutch adults, and it turned out that distributional learning is a less effective mechanism for them.

By comparing the present study and other precious studies, it is suggested that the effect of training could be affected by many factors. For instance, some studies suggest that the training effectiveness can be limited by the stimuli, and a bimodal distribution in which the relevant acoustic cue is artificially exaggerated would lead to a better training effect (Escudero, 2011).

If we compare the Mandarin speaker's performance in discriminating vowel duration in experiment 2 and experiment 3, we can see that the average accuracy is slightly higher in experiment 3. This difference might be related to the stimuli used in the tasks. In experiment 2, the duration difference between the test stimuli is 40ms, while in experiment 3 this number is 60ms. Moreover, the difference between the peak values of the stimuli in bimodal distribution is also slightly larger in experiment 3 (100ms) compared to experiment 2 (80ms).

It is possible that a larger difference in the peak value of the stimuli will be perceptually more salient for non-native listeners to capture the acoustic cue that signals the contrast between a minimal pair.

In the present study, the participants were exposed to the stimuli passively, and

during the training they were not required to conduct any operations on the computer. However, this type of passive training might not result in an effective training when the listener's attention is absent. Some studies in second language acquisition have claimed that attention is necessary for the process of language learning to take place. It is suggest that distributional learning is more effective when attention is given to the to-be-learned items (Toro et al., 2005; Turk-Browne et al., 2005). Therefore, it is assumed that the participants might have a better performance if they were told to focus on the vowel duration feature ahead of time instead of listening to the stimuli passively.

Experiment 3 also attempted to investigate the relationship between perception and production in second language acquisition. It turned out that the training effect of perception did not show in production.

The relationship between perception and production in second language learning has been widely discussed. Many researches have stated that perception and production are closely related and they also share one common underlying mechanism. For instance, Motor theory (Liberman et al., 1967) claims that when it comes to speech perception and production, perceiving speech is to perceive gestures, and the motor system is recruited for perceiving speech. Basically, it means that that listeners perceive speech using a phonetic module that represents speech in terms of neuromoteor commands to the articulators.

Direct realist theory (Best, 1995; Fowler, 1986) also states that listeners perceive speech using a general perceptual system, which detects the actual articulatory gestures of the speaker's vocal tract, which means that speech perception is part of learning to use the vocal tract, and a listener can only perceive the sound correctly when he knows how to produce it himself.

If these hypotheses are true, we can make suppose that if L2 learners who are good at perceiving L2 speech sounds are also good at producing the sounds. And also, the production should benefit from perceptual trainings. Nevertheless, the result of the present study shows that distributional leaning did improve the speaker's perception of F1 offset frequency, but it did not improve the speaker's production significantly, which indicates that perception might be somewhat independent from production, and perceptual knowledge does not benefit production directly.

7. Conclusion

This study was undertaken in order to investigate the perception and production of secondary acoustic cues to syllable-final stop voicing by Mandarin native speakers, also to examine the effect of distributional learning on the acquisition of secondary acoustic cues.

The perception and production of unreleased syllable-final stop voicing is one of the most common difficulties that Mandarin native speakers encounter during English learning, and it is considered to be due to negative transfer from their first language. Besides the reason that voicing contrast in syllable-final position is unfamiliar to Mandarin speakers, the fact that Mandarin lacks intrinsic and extrinsic vowel duration differences also makes it difficult for Mandarin native speakers to perceive durational differences in a foreign language. Therefore, Mandarin speakers have more difficulties in perceiving syllable-final stop voicing when the final consonant is unreleased.

In order to investigate this issue, four experiments were conducted. Experiment 1 focused on the production of syllable-final stop voicing by Mandarin native speakers in comparison with English native speakers. The result indicates that mandarin speakers do not produce different vowel durations conditioned by voicing. They produce lower F1 offset value under voiced conditions than voiceless conditions, but the difference is significantly smaller compared to English native speakers.

This indicates that Mandarin speakers could not provide salient vowel duration cue and F1 frequency cue in their production.

Experiment 2 investigated the effect of distributional learning of secondary acoustic cue in durational domain. In experiment 2, the participants were divided into four groups based on whether the stimuli they exposed to are in a bimodal distribution

or a unimodal distribution, and also whether or not the final consonant of the stimuli are released or unreleased.

The result reveals significant interaction between test type and distribution type. At the same time, no significant triple interaction was found among distribution type, test type and stimulus type. In other words, although bimodality effect is confirmed in the result, we cannot draw a conclusion that training is more effective with released stimuli. Further research needs to be done to confirm that the improvement in the released bimodal group is due to the factor of syllable-final voicing.

The follow-up experiment further confirmed a significant two-way interaction between test type and stimulus type, which indicates that distributional leaning is not only effective in the phonemic level but also on the phonological level. The subjects can integrate the acoustic feathers in speech sounds and find out the underlining association between the features.

Based on the result, experiment 3 was conducted to find out whether the effect of distributional learning could expand to multi-dimensions. The result shows that the participants scored significantly higher in block 2 (frequency dimension) compared to block 3 (duration dimension). And also, there was a significant interaction between test type (pre-test vs. post-test) and distribution type (bimodal vs. unimodal) in block 2 but no in block 1 and block 3. This suggested that the effect of distributional training only showed in the frequency dimension while it did not improve the subjects' sensitivity to vowel duration significantly. From here, we can see that in distributional learning, the participants tend to focus on the information in the dimension that they are most familiar with, which leads to a better training effect.

By comparing the training effect in vowel duration in experiment 2 and experiment 3, we can further see that distributional learning of difficult phonetic

contrasts is more effective in the case of one-dimensional training, in which the difference between the contrasts is more salient for the listeners to capture.

As part of experiment 3, the present study also investigates whether training on perception could have a positive influence on production. The result shows that distributional leaning did improve the speaker's perception of F1 offset frequency, but it did not improve the speaker's production significantly, which indicates that perception might be somewhat independent from production.

Foreign language skill has always been seen as an important part in education in China. Every year new foreign language schools are being established. But due to the lack of phonetic studies, the pronunciation training is still insufficient. As one of the problems that haven't been widely noticed, the perception and production of syllable-final voicing contrasts have confused many English learns in China.

As a typical example of acquiring secondary acoustic cues in L2 speech sounds, by investigating the acquisition of English syllable-final stop voicing by Mandarin native speakers, we tried to find a possible training method (distributional learning) to help learners' to built up associations between acoustic features.

The results revealed that distributional learning is effective to help Mandarin speaking English learners to build up the association between preceding vowel's acoustic features with the syllable-final voicing feature, while it is more effective when the training is one-dimensional. It is considered that the effect of distributional learning can also be applied to the acquisition of other difficult L2 secondary acoustic cues.

The present study will hopefully draw more people's attentions to the potential of distributional learning in English education. Nowadays, most of the training methods in foreign language education have been on staying on sentence level or word level. It

is hoped that this study can provide another possibility to training speech perception and production on feature level.

8. Further Studies

In this study, Mandarin native speakers showed higher sensitivity to frequency information compared to durational information. Whether or not mandarin native speakers rely on F1 frequency value to perceive syllable-final voicing still remains unclear. Further perceptual experiment is needed in order to see how Mandarin native speakers weigh the cues in perception.

In order to have a clearer vision of Mandarin speaker's perceptual strategy, it is also necessary to see how Mandarin native speakers perceive each other's English production of unreleased syllable-final stop contrasts. Moreover, in this experiment all subjects were exposed to the stimuli passively. It still remains to be seen how listener's attention influences the training effect.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Prof. Ryoko Hayashi for the precious comments, remarks and engagement through the whole process of this thesis, and to Prof. Reiko Yamada and Prof. Toshiyuki Sadanobu for being the examiners for my thesis defense and giving me very valuable comments.

I also want to thank the participants who have willingly taken their precious time to take part in the experiments. I express my warm thanks to Michaël van Rijsoort for reviewing this thesis in a very short notice. Also many thanks go to all the people who have made valuable comment suggestions on this thesis. Finally thanks to my parents for their continuous and unconditional supports. This thesis would not have been possible without the support and helps of the people that mentioned above.

References

Anderson, J. I. (1983). The difficulties of English syllable structure for Chinese ESL learners. Language Learning & Communication, 2, 53-61.

Aoyama, K. et al. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. Journal of Phonetics, 32, 233–250.

Benkí, J. R. (2005). Perception of VOT and First Formant Onset by Spanish and English Speakers. In ISB4: Proceedings of the 4th International Symposium on Bilingualism, edited by James Cohen, Kara T. MacAlister, Kellie Rolstad, and Jeff MacSwan, 240-248.

Best, C. T. (1995). Learning to perceive the sound pattern of English. In C. Rovee-Collier & L. Lipsitt, Advances in infancy research, 217-304.

Best, C. T. (1995). A direct realist view of cross-language speech perception. in Speech Perception and Linguistic Experience: Issues in Cross-Language Research, edited by W. Strange (York, Baltimore), 171-204.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: a perceptual assimilation model. In J. Goodman & H.C. Nusbaum (Eds.), The development of speech perception: The transition from speech to spoken words, 167–224.

Best, C.T., McRoberts, G.W., & Sithole, N.M. (1988). Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants. Journal of Experimental Psychology: Human Perception and Performance, 14, 345–360.

Boersma, P. (1998). Functional phonology: formalizing the interaction between articulatory and perceptual drives. Ph.D. dissertation, University of Amsterdam.

Boersma, P. (2007a). Some listener-oriented accounts of h-aspiré in French. Lingua, 117, 1989–2054.

Boersma, P. (2007b). Cue constraints and their interactions in phonological perception and production. Rutgers Optimality Archive 944.

Boersma, P., Weenink, D. (2006). Praat: Doing Phonetics by Computer. (Version 6.0.21). [computer software]. Retrieved from http://www.praat.org/.

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2 / Tone 3 distinction: Evidence of auditory enhancement. Journal of Phonetics, 18(1), 37-49.

Broselow, E., & Xu, Z. (2004). Differential difficulties in the acquisition of second language phonology. International Journal of English Studies, 4, 135 -163.

Burnham, D. K., Earnshaw, L.J., & Clark, J.E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. Journal of Child Language, 18, 231-260.

Chang, Y. S. (2011). Distinction Between Mandarin Tones 2 and 3 for L1 and L2 Listeners. Proceedings of the 23rd North American Conference on Chinese Lingusitics (NACCL-23), 2011, 84-96.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. Phonetica, 22, 129-159.

Denes, P. (1955). Effect of duration on the perception of voicing. JASA, 27, 761-64.

Eilers, R. (1977). Context-sensitive perception of naturally produced stop and fricative consonants by infants. Journal of the Acoustical Society of America, 61, 1321-1336.

Eilers, R., et al. (1984). The discrimination of vowel duration by infants. Journal of the Acoustical Society of America, 75, 1213-1218.

Eimas, P. D. (1975). Speech perception in early infancy. In L.B. Cohen & P. Salapatek

(Eds.), Infant perception: From sensation to cognition, Vol. II, 193–231. New York: Academic Press.

Enomoto, A. (2005). Perception of coda consonants by Japanese learners of English. Journal of the Phonetic Society of Japan, Vol.9, No.1, 73-84.

Engstrand, O., & Krull, D. (2001). Segment and syllable reduction: preliminary observations. Working Papers Lund University, Departement of Linguistics, 49, 26-29.

Escudero, P, Benders, T, Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. Journal of the Acoustical Society of America [Internet], 130 (4), EL206–12.

Escudero, P., Simon, E., Mitterer, H. (2012). The perception of English front vowels by North Holland and Flemish listeners: acoustic similarity predicts and explains cross-linguistic and L2 perception. Journal of Phonetic, 40, 280–288.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German and Spanish listeners. Journal of Phonetics, 37, 452–465.

Escudero, P., Williams, D. (2014). Distributional learning has immediate and long-lasting effects. Cognition, 133(2), 408-13.

Eimas, P. D., and Miller, J. L. (1980). Discrimination of information for manner of articulation. Infant Behav. Dev. 3, 367–375.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realist perspective. Journal of Phonetic, 14, 3-28.

Flege, J. E. (1979). Phonetic interference in second language acquisition. Unpublished doctoral dissertation, Indiana University.

Flege, J. E. (1993). Production and perception of a novel second-language phonetic

contrast. Journal of the Acoustical Society of America, 93, 1589-1608.

Flege, J. E. (1995). Second Language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience, 233-273.

Flege, J. E. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. Journal of the Acoustical Society of America, 86, 1684–1697.

Flege, J. E., Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast. Journal of Phonetics, 17. 299-315.

Flege, J. E., Munro, M. J., Skelton, L. (1992). Production of the word-final English /t/ - /d/ contrast by native speakers of English, Mandarin, and Spanish. Journal of the Acoustical Society of America, 92, 128-142.

Flege, J. E., Takagi, N., and Mann, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. Language and Speech, 38, 25-55.

Flege, J. E., Bohn, O-S., and Jang, S. (1997). The production and perception of English vowels by native speakers of German, Korean, Mandarin, and Spanish. Journal of Phonetic, 25, 437-470.

Flexner, S. B. (Ed.) (1987). The Random House Dictionary of the English Language, 2nd edition. New York: Random House.

Fischer, R. M., and Ohde, R. N. (1990). Spectral and duration properties of front vowels as cues to final stopconsonant voicing. Journal of the Acoustical Society of America, 88, 1250–1259.

Gasser, M. (2006). How Language Works: The Cognitive Science of Linguistics. Indiana University. Free Software Foundation.

Goudbeek, Martijn, Anne Cutler, & Roel Smits. (2008). Supervised and unsupervised

learning of multidimensionally varying non-native speech categories. Speech Communication, 50. 109–125.

Guenther, F. H., Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. Journal of the Acoustical Society of America, 100, 1111–1121.

Gulian, M., Escudero, P., Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. Proc. 16th ICPhS Saarbrücken, 1893–1896.

Hillenbrand, J., Ingrisano, D.R., Smith, B.L., Flege, J. E. (1984). Perception of the voiced-voiceless contrast in syllable-final stops. The Journal of the Acoustical Society of America, 76 (1), 18-26.

House, A. S. (1961). On Vowel Duration in English. Journal of f e co sffc Socze of menca, 33, 1174-1178.

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

Kessler, Brett, Treiman, Rebecca. (1997). Syllable structure and the distribution of phonemes in English syllables. Journal of Memory and Language, 37, 295-311.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. Journal of the Acoustical Society of America 72, 379–389.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87, 820–857.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. Journa1of the Acoustlc Socle of America, 54 (4), 1102-1104.

Kluender, K. R., Diehl, R. L., & Wright. B. A. (1988). Vowel-length differences before voiced and voiceless consonants: an auditory explanation. Journal of Phonetic, 16, 153-169.

Ko, E-S. (2007). Acquisition of Vowel Duration in Children Speaking American English, Proceedings of Interspeech, 2007, 1881-1884.

Ko, E., Soderstrom, M., Morgan, J. (2009). Development of perceptual sensitivity to extrinsic vowel duration in infants learning American English. Journal of the Acoustical Society of America, 126(5), EL134–EL139.

Kurihara, M. (2004). Perceptual characteristics of Japanese long vowels by Mandarin Chinese speakers. Journal of Linguistic Science (Tohoku University), 8, 1-12.

Kurihara, M., Chan, W. M., Chin, K. N., Martin-Lau, P., Nagami, M., Suthiwan, T., Suzuki, M. (2006). The Identification of Japanese Long and Short Vowels by Mandarin Chinese Speakers. CLaSIC 2006: Processes and Process-Orientation in Foreign Language Teaching and Learning. The 2nd CLS International Conference, CD-ROM. National University of Singapore, 451-460.

Kubozono, H. (1995). Perceptual evidence for the mora in Japanese. In B. Connell and A. Arvaniti (eds.) Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV. (pp.141-156). Cambridge: Cambridge University Press.

Kubozono, Haruo. (1996a). Speech segmentation and phonological structure. In Phonological structure and language processing: Cross-linguistic studies, 1–12.

Kubozono, Haruo. (1996b). Syllable and accent in Japanese: Evidence from loanword accentuation. Onsei Kaihoo, 211, 71–82.

Kuhl, P.K. (1993). Early linguistic experience and phonetic perception: implications for theories of developmental speech perception. Journal of Phonetics, 21, 125–139.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet

effect". In Speech perception and linguistic experience: issues in cross-language research, ed. by Winifred Strange, 121–154.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Dev. Sci, 9, F13–F21.

Krause, S. E. (1982). Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. Journal of the Acoustical Society of America, 71, 990–995.

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: Flexible yet constrained. Frontiers in Psychology, 3, 598.

Lisker, L. (1986) “Voicing” in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. Language and Speech, 29, 3–11.

Lisker, L, Abramson, A. S. (1964). A cross-linguistic of voicing in initial stops: Acoustical measurements. Word, 20, 384-422.

Lisker, L. (1974). On 'explaining' vowel duration variation. G1ossa, 8, 233-246.

Lacerda F. (1995). The perceptual-magnet effect: an emergent consequence of exemplar-based phonetic memory. Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, (Eds. K. Elenius and P. Branderyd), Vol.2, 140-147.

Ladefoged, P., Johnson, K. (2010). A Course in Phonetics 6th edition. Boston: Thomson Wadsworth.

Lehiste, Ilse. (1970). Temporal organization of spoken language. In Working papers in linguistics, 4, 96–114.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of

Experimental Psychology, 54, 358–368.

Liberman, A. M., Cooper, F. S., Shanksweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of speech code. Psychological Review, 74, 431-461.

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

Liberman, A. M., and Mattingly, I. G. (1989). A specialization of speech perception. Science, 243, 489-494.

Maye, J., Weiss, D., Aslin, R. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. Dev. Sci., 11, 122–134.

Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? In A. H.-J. Do, L. Dominquez, & A. Johansen (Eds.), BUCLD 25 Proceedings, 480–490.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), BUCLD24 Proceedings, 522–533.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82, B101–111.

Minagawa-Kawai, Y. (2001). Preciseness of temporal compensation in Japanese mora timing. Phonetica, 58, 1-25.

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J., and Amano, S. (2009). Perception of vowel Length by Japanese- and English-learning infants. Dev. Psychol. 45, 236–247.

National Bureau of Statistics, China statistical yearbook 2001.

Naeser, M. (1970). The American Child's Acquisition of Differential Vowel Duration. Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, Madison, Tech. Rep. 144.

Ong, J.H., Burnham, D., Escudero, P. (2015). Distributional Learning of Lexical Tones: A Comparison of Attended vs. Unattended Listening. PLoS ONE 10(7): e0133446.

Ohala, J. J., & Kawasaki, H. (1984). Prosodic phonology and phonetics. Phonology, 1, 113–128.

Pająk, B., Levy, R. (2012). Distributional learning of L2 phonological categories by listeners with different language backgrounds. In A. K. Biller, E. Y. Chung, & A. E. Kimball (Eds.), Proceedings of the 36th Boston University Conference on Language Development, 400-413.

Pisoni, D. B., Remez, R. E. (2005). The handbook of speech perception. Malden, MA: Blackwell Pub.

Raphael, L. J. (1972). Vowel Duration as a Cue to the Perceptual Separation of Cognate Sounds in American English. Journal of the Acoustical Society of America, 51, 1296-1303.

Raphael LJ, Dorman MF, Freeman F, Tobin C. (1975). Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies. Journal of Speech, Language, and Hearing Research, 18, 389–400.

Raphael, L. J., Dorman, M.F., Liberman, A.M. (1980). On defining the vowel duration that cues voicing in final position. Lang Speech, 23, 297–307.

Recasens, D. (2004). The effect of syllable position on consonant reduction (evidence fromCatalan consonant clusters). Journal of Phonetics, 32 (2004), 435–453.

Repp, B. (1982). Phonetics trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

Rodgers, B., Purnell, T., Salmons, J. (2008). Expanding the set of acoustic features of the post-vocalic voicing contrast in English. Journal of the Acoustical Society of America, 123, 3888.

Saffran, J. R., Johnson, E. K., Aslin, R.N., Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70 (1), 27-52.

Saffran, J.R. (2003). Statistical language learning: Mechanisms and constraints. Current Directions in Psychological Science, 12, 110–114.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. Journal of Memory and Language, 35, 606–621.

Stevens, K. N. & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55, 653-659.

Summers, W. V. (1988). F1 structure provides information for final-consonant voicing. Journal of the Acoustical Society of America, 84, 485.

Tauberer, J. 1. and K. Evanini (1999). Intrinsic vowel duration and the post-vocalic voicing effect: Some evidence from dialects of North American English. Eurospeech 1999 [CD-ROM].

Toro, J. M., Sinnett, S., Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. Cognition, 97 (2), B25 - B34.

Treiman, R. (1984). On the status of final consonant clusters in English syllables. Journal of verbal learning and verbal behavior 23, 343-356.

Turk-Browne, N. B., Jungé, J., Scholl, B. J. (2005). The automaticity of visual statistical learning. Journal of Experimental Psychology: General, 134, 552–564.

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. Journal of Experimental Psychology: Human Perception and Performance, 35, 195–202.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without

awareness. Journal of Cognitive Neuroscience, 21, 1934–1945.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. The Journal of Neuroscience, 30, 11177–11187.

Uchida, T. (1993). Characteristics of auditory cognition of long vowels and double consonants for Chinese students in learning Japanese language. Japanese Journal of Educational Psychology, 41, 414-423.

Wanrooij, K., Boersma, P., & Van Zuijen, T. L. (2014a). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. Frontiers in Psychology. Language Sciences, 5, Article 77, 1-12.

Wanrooij, K., Boersma, P., & Van Zuijen, T. L. (2014b). Distributional Vowel Training Is Less Effective for Adults than for Infants. A Study Using the Mismatch Response. PloS One 9, 1-13.

Wanrooij, K., Boersma, P. (2013). Distributional training of speech sounds can be done with continuous distributions. Journal of the Acoustical Society of America, 133, EL398-EL404.

Wanrooij, K., Escudero, P., Raijmakers, M.E.J. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. J. Phon, 41, 307-319.

Wardrip-Fruin, C., Peach, S. (1984). Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants. Lang Speech, 27, 367-379.

Weenink, D. (2016). Speech Signal Processing with Praat. Retrieved from http://www.speechminded.com.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for

perceptual reorganization during the first year of life. Infant Behavior and Development, 7, 49-63.

Winn, M., Edwards, J., & Litovsky, R. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear and Hearing, 36, e153-e165.

Xu, N., & Demuth, K. (2012). Acquisition of English codas by Mandarin-speaking adult learners of English. Proceedings of the 14th Australasian International Conference on Speech Science and Technology, 41-44.

Zamuner, T. S. (2001). Input-based phonological acquisition. Unpublished doctoral dissertation, University of Arizona, Tucson.

Zhang, Y., Hayashi, R. (2015). Perception of syllable-final consonants by Chinese speakers and Japanese speakers, Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4.

Appendix

Appendix 1 – Praat script for discrimination task in experiment 2

"ooTextFile"

"ExperimentMFC 7"

blank while playing? <no>

stimuli are sounds? <yes>

"sounds/" ".wav"

carrier phrase "" ""

initial silence duration 1 seconds

medial silence duration 0.8 seconds ! inter-stimulus interval

final silence duration 0 seconds

4 different stimuli

"S4,S7" ""

"S7,S4" ""

"S4,S4" ""

"S7,S7" ""

20 replications per stimulus

break after every 0 stimuli

<PermuteBalancedNoDoublets>

"Click to start."

"Say whether these words were the same or different."

"You can have a short break if you like. Click to proceed."

"The experiment has finished. Call the experimenter."

0 replays

replay button 0 0 0 0 "" ""

ok button 0 0 0 0 "" ""

oops button 0 0 0 0 "" ""

responses are sounds? <no> "" "" "" "" 0 0 0

2 response categories

0.1 0.4 0.35 0.65 "same" 24 "" "same"

0.6 0.9 0.35 0.65 "different" 24 "" "different"

0 goodness categories

Appendix 2 – Praat script for stimuli distribution (bimodal) in experiment 2

"ooTextFile"

"Distributions"

1 "frequency"

17

"a-4" 8

"a-5" 8

"a-6" 16

"a-7" 32

"a-8" 72

"a-9" 48

"a-10" 32

"a-11" 16

"a-12t" 4

"a-12d" 4

"a-13" 16

"a-14" 32

"a-15" 48

"a-16" 72

"a-17" 32

"a-18" 16

"a-19" 8

"a-20" 8

Appendix 3 – Praat script for stimuli distribution (bimodal) in experiment 3

"ooTextFile"

"Distributions"

1 "frequency"

10

"S1" 16

"S2" 48

"S3" 80

"S4" 48

"S5" 8

"S6" 8

"S7" 48

"S8" 80

"S9" 48

"S10" 16

Appendix 4 – Praat script for stimuli presentation in experiment 3

```
dist = Read from file: "r.bimodal.Distributions"
trials = To Strings (exact): 1
numberOfTrials = Get number of strings
for trial to numberOfTrials
        selectObject: trials
        stimulus$ = Get string: trial
        sound = Read from file: "stimuli/released/" + stimulus$ + ".wav"
        Play
        Remove
endfor
removeObject: dist, trials
```

Appendix 5 – Praat script for synthesizing stimuli /la/ continuum

```
pitchTier = Create PitchTier: "f0", 0.00, 0.50
Add point: 0.00, 180
Add point: 0.50, 150
pointProcess = To PointProcess
Remove points between: 0.00, 0.05
Remove points between: 0.45, 0.50
source = To Sound (phonation): 44100, 0.6, 0.05, 0.7, 0.03, 3.0, 4.0
intensity = Create IntensityTier: "intens", 0.00, 0.50
Add point: 0.05, 60
Add point: 0.10, 80
Add point: 0.45, 70
for i from 1 to 10
    fn = 450 + ((1050 - 450)/9) * (i - 1)
    formantGrid = Create FormantGrid: "filter", 0.0, 0.5, 9, 800, 1000, 60, 80
    # it is better to set the formant values explicitly:
    # first remove all points in the tiers we want to set
    Remove formant points between: 1, 0.0, 0.5
    Remove formant points between: 2, 0.0, 0.5
    Remove formant points between: 3, 0.0, 0.5
    Remove formant points between: 4, 0.0, 0.5
    Add formant point: 1, 0.05, 450
    Add bandwidth point: 1, 0.05, 50
    Add formant point: 1, 0.08, 450
```

Add bandwidth point: 1, 0.05, 50

Add formant point: 1, 0.10, 650

Add bandwidth point: 1, 0.05, 50

Add formant point: 1, 0.20, 1050

Add bandwidth point: 1, 0.05, 50

Add formant point: 1, 0.30, 1050

Add bandwidth point: 1, 0.05, 50

Add formant point: 1, 0.45, fn

Add bandwidth point: 1, 0.05, 50

Add formant point: 2, 0.05, 1450

Add bandwidth point: 2, 0.05, 100

Add formant point: 2, 0.08, 1450

Add bandwidth point: 2, 0.05, 100

Add formant point: 2, 0.10, 1450

Add bandwidth point: 2, 0.05, 100

Add formant point: 3, 0.05, 3500

Add bandwidth point: 3, 0.05, 150

Add formant point: 3, 0.08, 3500

Add bandwidth point: 3, 0.05, 150

Add formant point: 3, 0.10, 2900

Add bandwidth point: 3, 0.05, 150

Add formant point: 4, 0.05, 4400

Add bandwidth point: 4, 0.05, 200

plusObject: source

filtered = Filter

```
    plusObject: intensity
    scaled = Multiply: "yes"
    Rename: "VtVd2_" + string$ (i)
    removeObject: formantGrid, filtered
endfor
removeObject: source, intensity, pointProcess, pitchTier
```