INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Effects of Spectral and Temporal Cues to Mandarin Concurrent-vowels Identification for Normal-hearing and Hearing-impaired Listeners

Zhen Fu[1], *Xihong Wu*[1], *Jing Chen*[1]

[1]Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

fuzhen364@pku.edu.cn, wxh@cis.pku.edu.cn, chenj@cis.pku.edu.cn

Abstract

In Mandarin Chinese, lexical Tones are inherently bonded with vowels, making both spectral and temporal cues available for speech perception. Temporal cues provided by Tone contrast have been shown facilitating segregation in Mandarin concurrent-vowels identification (MCVI). The present study investigated the effect of spectral cue measured by vowel contrast within the syllable-pair on MCVI, both for normal-hearing (N-H) and hearing-impaired (HI) listeners. Acoustic cues of duration and mean F0 difference were carefully controlled. Results exhibited that facilitation from vowel contrast existed for NH listeners but was reduced for HI listeners. Identification score positively correlated with the spectral envelope contrast of different vowel-pairs for both groups, but the coefficient for HI listeners was lower. Further analyses based on a power function model revealed more weighting of temporal cues than spectral cues for NH listeners, while the contributions were equal for HI listeners. These results suggested that the spectral cue provided by vowel contrast could facilitate the MCVI, and auditory processing of temporal cues might be more susceptible to hearing loss than that of spectral cues. These findings have instructions for designing speech processing algorithms for Mandarin-speaking HI listeners.

Index Terms: Mandarin concurrent-vowels, spectral envelope, F0 contour, hearing impairment

1. Introduction

Listeners with normal hearing (NH) usually possess a remarkable capability of speech perception in complex auditory scenarios where background noises or competitive speakers exist. One of the most challenging scenarios is the target speech presented simultaneously with the competing speech. A simplified version of such challenging scenarios, in which two vowels are presented simultaneously, is known as concurrent-vowels identification (CVI), and has been widely adopted to investigate which acoustic cues were utilized by listeners to facilitate the speech perception [1, 2, 3, 4]. Besides the spectral envelope cue which contains formant information determining vowel identity, other acoustic cues such as fundamental frequency (F0) difference [1], frequency modulation (FM) [2, 3], onset misalignment [4] and F0 contours difference [3] were also shown beneficial for segregating the two vowels. However, most of these studies were conducted on English vowels, except for [3] which examined the effect of F0 contours using Mandarin vowels. Therefore, the roles of various acoustic cues playing in CVI for tonal languages such as Mandarin, need further investigation.

In Mandarin speech, the direction of vowel's F0 contour defines the category of lexical Tones, which provide crucial features to distinguish the semantic meanings of Chinese words. There are four lexical Tones in Mandarin, described phonetically as high level (T1), high rising (T2), rising falling (T3), and falling (T4), respectively. For identification of Mandarin tonal vowels, cues for vowel and Tone perception are both available and utilized by listeners. Vowel perception relies on spectral cue of spectral envelope, in which peaks manifest formants, and Tone perception depends on temporal cues such as F0 contour, temporal envelope and duration [5, 6]. When duration and mean F0 were controlled, cues of F0 contour and temporal envelope were utilized by listeners in Mandarin CVI (MCVI) [3]. Specifically, MCVI performance increased with both the F0 contour contrast and temporal envelope contrast within the syllable-pair. However, the effect of spectral cue on MCVI was uninvestigated. In other words, whether MCVI performance would increase with the difference between spectral envelopes (spectral envelope contrast) still remains unclear. Besides, a previous research reported similar contributions of vowel and Tone perception to sentence recognition [7]. However, such a relative contribution has not been studied for MCVI task, which is a more complex scenario and without context information.

Unlike NH listeners, hearing impaired (HI) listeners usually have difficulty in speech perception, especially in noisy and multi-talker auditory scenarios. For HI listeners with the type of cochlear hearing loss, such difficulty is considered associated with deficits in suprathreshold auditory functions [8, 9, 10], such as FM detection, amplitude modulation detection and frequency selectivity. It was reported that the assistance provided by temporal cues was weak for HI listeners in MCVI task [3], which was attributed to their degraded ability to utilize F0 contour cue [10]. Besides, reduced frequency selectivity for HI listeners has been well established that it would produce spectral smearing and thus hinder listeners from using spectral cues [8, 9]. Nevertheless, the effect of hearing loss on utilizing of by spectral cues in MCVI remains unclear. Moreover, since both auditory functions to process spectral and temporal cues were degraded for HI listeners, how does hearing loss affect the relative contributions of such two cues in MCVI is another question we attempted to address. A decomposition of these cues could be instructive to develop speech processing algorithms of auditory assistant devices for Mandarin-native users.

In summary, the present study aimed to (1) investigate the effect of spectral cue on MCVI for both NH and HI listeners; and (2) explore the relative contributions of spectral and temporal cues involved in MCVI for the two listener groups. We hypothesized that the spectral cue provided by vowel contrast could facilitate the MCVI prominently for NH listeners, while such a facilitation was weak for HI listeners.

http://dx.doi.org/10.21437/Interspeech.2019-3209

2. Methods

2.1. Subjects

The subject set is an expansion of a previous work [3]. Twelve NH listeners (5 males, mean age = 23.3 years) and thirteen HI listeners (7 males, mean age = 39.8 years) with sensorineural hearing loss participated in the experiment. The mean hearing threshold at 0.5, 1, 2 and 4 kHz ranged from 0 to 13.75 dB H-L and 35 to 80 dB HL for NH and HI subjects, respectively. The latter one corresponds to a mild to severe hearing loss degree. All subjects were Mandarin-native speakers, were paid for their participation and given informed consent approved by the Peking University Institutional Review Board.

2.2. Stimuli

Stimuli in the present study were the same as used in a previous work [3]. A female Mandarin-native speaker produced six Mandarin single-vowels (IPA symbol: [a], [ɔ], [ɤ], [i], [u], [y]) with four Tones for each vowel, resulting twenty four single-vowel syllables in total. The sampling rate of stimuli was 44.1 kHz. To avoid the effect of duration on Tone perception [5, 6], the duration of all syllables was adjusted at 450 ms, including 15-ms raised-cosine rise and fall ramps. To avoid the effect of mean F0 difference between syllables on CVI, all original syllables were processed using Praat [11] to equalize their mean F0 at 210 Hz, while F0 contour shapes remained unchanged. The root-mean-square levels of all syllables were also equated. After these processes, experimenters confirmed the syllable identity was unchanged. The concurrent-vowels pairs were constructed by mixing any two different single-vowel syllables, leading to a total of 276 (C_{24}^2) concurrent-vowels pairs.

2.3. Procedure

A custom graphical user interface written in MATLAB was built to present stimuli and collect responses. Stimuli were presented at the right ear for NH subjects and at the better ear for HI subjects through a Sennheiser HD 650 headphone in a sound-proof booth. For HI subjects, the frequency-dependent gains specified by the "Cambridge" hearing-aid fitting procedure were applied to stimuli to ensure the audibility of stimuli over the frequency range important for speech intelligibility, based on individual audiogram [12]. The sound level was 55 dB SPL and about 30 dB SL for NH and HI subjects, respectively. All concurrent-vowel pairs were presented trial by trial in a random order for each subject. After each presentation, subjects were instructed to identify both syllables (vowel plus Tone) on the interface by clicking the response buttons labeled with vowel or Tone categories. Specifically, for each trial, subjects completed two sets of choices and each set consisted of six vowels and four Tones as candidates. Subjects received a training session to be familiar with the stimuli and procedure.

2.4. Data analysis

2.4.1. Identification score calculation

Identification scores were calculated as the percentage that both syllables (vowel plus Tone), both vowels (regardless of Tone response), or both Tones (regardless of vowel response) were correctly identified. To investigate the effect of vowel contrast on MCVI, syllable identification scores were first counted for each of the 21 vowel-pairs (e.g., [a]-[a], [a]-[ɔ], [a]-[ɤ], etc.), respectively, and then were divided into two groups: same vowel (SV) vs. different vowels (DV). for each vowel category. The percentages were further transformed to rationalized arcsine units (RAU) to make the scores distribution more normal [13].

2.4.2. Spectral envelope contrast measurement

For attributing syllable identification scores to spectral cue, the spectral envelope contrast was calculated for each syllable-pair and arranged as a function of vowel-pair. Specifically, the spectral envelope of each single vowel was calculated using the cepstrum method, and the spectral envelope contrast of a certain concurrent-vowel was measured by calculating the mean-square-error between the two spectral envelopes of the two single vowels. In consideration of the frequencies of first three formants that contributed to vowel identity, the spectral envelopes below 4 kHz were used. Subsequently, the spectral envelope contrast values were collapsed into vowel-pairs, since the difference across Tone-pairs was found little in the present study. Finally, the contrast values were normalized to the range [0,1], and it was assumed that greater contrast would provide more cues for MCVI and thus improve the scores.

2.4.3. Power-function model

Since cues for vowel perception (e.g., spectral envelope) and Tone perception (e.g., F0 contour and temporal envelope [3]) were both utilized in the MCVI task, it is necessary to explore the relative importance between them. The power-function model was proposed to relate the probability of identifying an individual syllable to the identification probabilities of all phonemes constructing it, hence to quantify the contributions of various phonemes to syllable and/or sentence recognition [7, 14]. The present study adopted such a model to disassemble the syllable identification score into vowel and Tone identification scores in power functions:

$$p_s = p_v^{w_v} * p_t^{w_t} \quad (1)$$

where p_s, p_v and p_t represent the identification score of syllable, vowel and Tone, respectively. The parameter w_v and w_t represent the weights for vowel and Tone, respectively. After fitting subjects' identification scores to the model, the weighting parameters were obtained, and this fitting procedure was separately conducted for NH and HI groups. We assumed that greater weight of a certain measurement (e.g., vowel or Tone identification score) meant subjects relied more on the cues dominated in that measurement for MCVI.

3. Results

3.1. Effect of spectral envelope contrast

Syllable identification scores of different vowel-pairs were organized as a function of vowel category and vowel contrast, as shown in Figure 1. SV and DV were considered representing the absence and presence of vowel contrast, respectively. For NH subjects, identification scores of the DV condition were always greater than that of the SV condition. A two-way ANOVA revealed that the main effects of vowel category [F(5,55)=2.547, p=0.038] and vowel contrast [F(1,11)=26.527, p<0.001] were both significant for NH subjects, and their interaction was also significant [F(5,55)=3.486, p=0.008]. Post-hoc analysis indicated that scores of DV were significantly higher than SV for vowel [a], [ɔ] ,[i] and [y] (p<0.029), but not for [ɤ] and [u] (p>0.915). For HI subjects, the main effect was significant for vowel category [F(5,60)=9.273, p<0.001], but not for vowel contrast [F(1,12)=1.412, p=0.258], and their interaction was

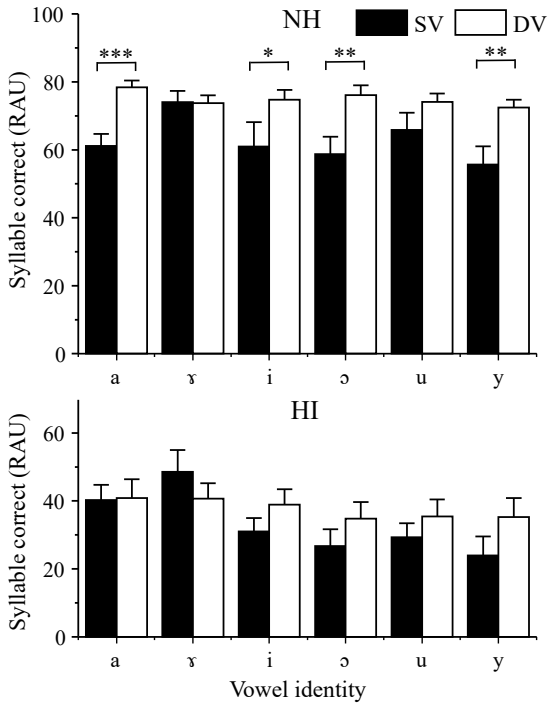

Figure 1: *Averaged syllable identification score as a function of vowel identity and vowel contrast for NH (top) and HI (bottom) subjects. Error bars indicate one standard error,*p <0.05, **p <0.01 and ***p <0.001.*

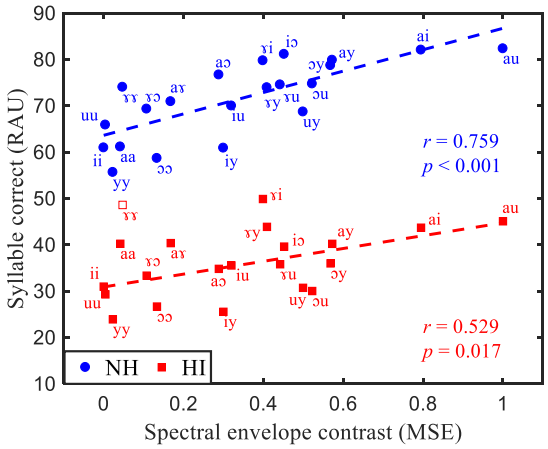

Figure 2: *Averaged syllable identification score as a function of spectral envelope contrast for NH (circles) and HI (squares) subjects. Dashed lines depict the linear regressions. Open symbol is the vowel-pair excluded from the regression analysis.*

significant [F(5,60)=3.343, p=0.028]. Post-hoc analysis indicated that score difference between DV and SV was not significant for any vowel (p>0.069). This result indicated that vowel contrast could significantly improve MCVI performance for NH subjects, but not for HI subjects. It should be noticed that scores for vowel [ɤ] were remarkably high for SV condition, especially for HI subjects. This vowel differs from others because it is an unsteady-state vowel, thus additional perceptual cues besides spectral envelope could be utilized by subjects [15]. Besides, the Mandarin syllable with the highest using frequency (IPA symbols: [tɤ], *pinyin* symbols: /de/, which stands for the possessive particle) is composed of the vowel [ɤ] [16], so listeners probably were more familiar with it. These factors might give rise to the higher syllable identification scores for vowel-pair [ɤ]-[ɤ]. For the complementary result that identification scores as a function of Tone identify, please see Figure 1 in [3].

Although the main effect of vowel contrast on MCVI was not significant for HI subjects, results of two groups still exhibited similar patterns, e.g., scores of DV were higher than SV for most vowels. This implied that it could be more appropriate to arrange MCVI scores according to the spectral envelope contrast of individual vowel-pair, instead of dividing them into two groups roughly. The syllable identification scores were plotted as a function of spectral envelope contrast of individual vowel-pair for NH and HI subjects, separately in Figure 2. Generally, the score improved with increasing spectral envelope contrast for both groups, with the exception of vowel-pair [ɤ]-[ɤ]. Consistent with the Figure 1, the vowel-pair [ɤ]-[ɤ] has a low contrast while with a high score, especially for HI subjects, so it was excluded from further regression analysis for HI subjects (open symbol in Figure 2). The linear regression analysis indicated that the correlation between syllable identification score and spectral envelope contrast was both significant for NH subjects (r=0.759, p<0.001), and HI subjects (r=0.529, p=0.017). Taken together, these results suggested that both NH and HI subjects could utilize the cues provided by spectral envelope contrast. However, since the slope of regression line was shallower for HI than NH subjects, HI subjects' ability to utilize such a cue was degraded. These results are likely due to the spectral smearing caused by the poor frequency selectivity of the impaired auditory system [17, 18].

3.2. Contributions of spectral and temporal cues on MCVI

Figure 3 shows the group-averaged syllable, vowel and Tone identification scores for NH and HI subjects, respectively. A two-way ANOVA (3 measures × 2 hearing status) revealed the main effects of hearing status [F(1,23)=43.707, p<0.001] and measures [F(2,46)=39.898, p<0.001] were both significant, and their interaction was also significant [F(2,46)=4.842, p=0.028]. Post-hoc analysis indicated that scores were significantly lower for syllable than vowel and Tone identification, for both subject groups (p<0.001). Fitting based on the power-function model in equation 1 showed the weights of vowel and Tone were 0.11 and 0.86 (R^2=0.94) for NH subjects, respectively. And the weights of vowel and Tone were 0.48 and 0.44 (R^2=0.78) for HI subjects, respectively. These results demonstrated that NH subjects relied more on Tone perception than vowel perception, while their contributions were similar for HI subjects. In other words, comparing to NH subjects, contribution of spectral cue was increased and that of temporal cues was decreased for HI subjects. This result suggested that the deleterious effect of hearing loss was much greater for auditory processing of temporal cues than spectral cues.

4. Discussion

4.1. Effect of hearing loss on spectral and temporal cues processing

A previous study demonstrated that hearing loss adversely affects auditory processing of temporal cues relevant to Tone per-

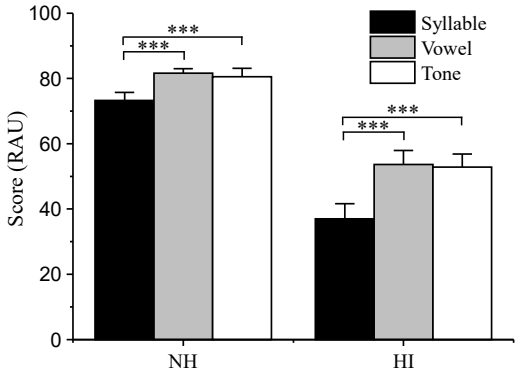

Figure 3: *Averaged syllable, vowel and Tone identification scores for NH (left) and HI (right) subjects. Error bars indicate one standard error,***p <0.001.*

ception [3], which was ascribed to degraded FM detection ability for HI listeners [9, 10]. Such a degradation was interpreted by the perception mechanism of temporal processing, that the short-term pattern of phase-locking to frequency change in auditory neurons was disrupted by hearing loss. Thus, the primary cue for Tone perception (i.e., F0 contour) was diminished, so that the F0 contour contrast within a syllable-pair was reduced.

A major finding of the present study is that the processing of spectral cues was also declined for HI listeners. This is likely because of their reduced frequency selectivity, which is associated with broadened auditory filters consequent to outer hair cell damage [9, 17, 18]. Comparing to normal auditory filters, broader ones would pass more frequency components within each filter and together cause a diffused (or say smeared) internal perception (i.e., the excitation pattern) of the speech spectral envelope. As a consequent, critical acoustic features such as formants are harder to detect and discriminate. A spectral smearing model was employed to process current single vowels as HI simulations [17, 18]. Under the model framework, smearing factors of 3 and 6 were used to be representative of filter bandwidths associated with mild-moderate and moderate-severe hearing loss. Then the aforementioned spectral envelope contrast measurement was performed to calculate the contrast within each syllable-pair after spectral smearing. On average, the spectral envelope contrast decreased by about 20% (*s.d.* = 8%) and 24% (*s.d.* = 8%) for the two factors, respectively. These simulated results were consistent with the conclusion for Figure 2 that for HI listeners, reduced frequency selectivity not only decreased the spectral contrast within each vowel itself, but also diminished the spectral envelope contrast between the two vowels of a syllable-pair, so that their abilities to utilize such a cue was degraded.

Another finding of the current work is that the contribution of temporal cues was greater than that of spectral cue for NH listeners, while was equivalent for HI listeners. This result suggested that utilizing of temporal cues might be more susceptible to hearing loss than that of spectral cues. This finding is similar to the conclusion that cochlear implant users may have been more susceptible to interference in the temporal domain than in the spectral domain [19]. It should be noted that a greater weight of vowel perception for HI than NH listeners does not mean they are better at taking advantage of the spectral envelope contrast, but is a trade-off between two perceptions that are both degraded. The demonstrated relative contributions of these two cues instruct us that, when designing speech processing algorithms of auditory assistant devices for tonal-language users, (1) spectral contrast enhancement algorithm [20] could be incorporated; and (2) the representation strength of these cues should be mediated according to limits of listener's spectral and temporal resolution.

4.2. Identification score counting methods

In the present study, identification scores were separately counted for the situation that both syllables, both vowels, or both Tones were correctly identified, the same as the method in [19]. But this method could be problematic for calculating the scores of vowel identification, since the responses to Tone were regardless, and vice versa. For example, if the syllable pair was [a1]-[u2], and listener's response was [a2]-[u1], this syllable would be considered as falsely-identified, whereas vowel and Tone would both be considered as correctly-identified based on the current criterion. The ratio of such trials in the present result was 0.54% and 0.78% for NH and HI listeners, respectively. Therefore, we consider the influence of this problem was little for our results. In an alternative counting method, each stimulus syllable-pair was divided into two tonal vowels, so was the responsive pair, then the counting was conducted [3]. The advantage of such an approach is that the calculation of confusion matrix of vowel and Tone becomes feasible. However, it could also be problematic for counting the confusions, since multiple matches could exist after uncoupling each pair. For the above example, deciding to whom ([a1] or [u2]) the response [a2] responds would be awkward, since they either coincide with vowel identity ([a1] and [a2]) or Tone identity ([u2] and [a2]). The ratios of such trials for the present result were 1% and 7.61% for NH and HI listeners, respectively, which were greater than the first method. When applying the second counting method, the matching strategy (e.g., higher priority for vowel) could slightly affect the identification and confusion patterns.

5. Conclusions

The present study examined the effect of vowel contrast on MCVI for NH and HI listeners, and the relative contributions of spectral and temporal cues in such a task for the two listener groups. The main findings included: (1) MCVI scores were better for syllable-pairs with different vowel-pair than that with same vowel-pair for NH listeners, but this effect was not significant for HI listeners. (2) When measuring the vowel contrast by the distance between two spectral envelopes, MCVI score positively correlated with vowel contrast for both listener groups. Unsurprisingly, such a facilitation was weakened by hearing loss. (3) NH listeners rely more on temporal cues than spectral cues, while contributions of two cues were equalized for HI listeners.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61771023 and 11590773), and a Medicine-Information research grant (BMU2018MI004) by Peking University.

7. References

[1] P. F. Assmann and Q. Summerfield, "Modeling the perception of concurrent vowels: vowels with different fundamental frequen-

cies," *Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 680–697, 1990.

[2] J. F. Culling and Q. Summerfield, "The role of frequency modulation in the perceptual segregation of concurrent vowels," *Journal of the Acoustical Society of America*, vol. 98, no. 2 Pt 1, pp. 837–846, 1995.

[3] Z. Fu, H. Yang, X. Wu, and J. Chen, "Acoustic cues utilized by normal-hearing and hearing-impaired listeners are different for Mandarin concurrent-vowels identification," *Acta Acustica united with Acustica*, vol. 104, no. 5, pp. 792–795, 2018.

[4] C. J. Darwin, "Perceptual grouping of speech components differing in fundamental frequency and onset-time." *Quarterly Journal of Experimental Psychology A*, vol. 33, no. 2, pp. 185–207, 1981.

[5] D. H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, no. 1, pp. 25–47, 1992.

[6] L. Xu, Y. Tsai, and B. E. Pfingst, "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 247–258, 2002.

[7] Q. J. Fu, F. G. Zeng, R. V. Shannon, and S. D. Soli, "Importance of tonal envelope cues in Chinese speech recognition," *Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 505–510, 1998.

[8] B. C. J. Moore and B. R. Glasberg, "The Relationship Between Frequency Selectivity and Frequency Discrimination for Subjects with Unilateral and Bilateral Cochlear Impairments," in *Auditory Frequency Selectivity*, ser. Nato ASI Series, B. C. J. Moore and R. D. Patterson, Eds. Springer, Boston, MA, 1986, pp. 407–417.

[9] O. Strelcyk and T. Dau, "Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing." *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3328–3345, 2009.

[10] B. C. J. Moore and E. Skrodzka, "Detection of frequency modulation by hearing-impaired listeners: effects of carrier frequency, modulation rate, and added amplitude modulation," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 327–335, 2002.

[11] P. Boersma and D. J. M. Weenink, "Praat : doing phonetics by computer [Computer program]," *http://www.praat.org/*, 2013.

[12] B. C. J. Moore and B. R. Glasberg, "Use of a loudness model for hearing-aid fitting. I. linear hearing aids," *British Journal of Audiology*, vol. 32, no. 5, pp. 317–335, 1998.

[13] G. A. Studebaker, "A "rationalized" arcsine transform." *Journal of Speech & Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.

[14] W. M. Rabinowitz, D. K. Eddington, L. A. Delhorne, and P. A. Cuneo, "Relations among different measures of speech reception in subjects using a cochlear implant," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 1869–1881, 1992.

[15] Z. Wu and M. Lin, *Shiyanyuyinxue Gaiyao [An overview of experimental phonetics]*. Beijing: Higher Education Press, 1989.

[16] H. H. Ho and T. W. Kwan, "Hong Kong, Mainland China & Taiwan: Chinese Character Frequency - A Trans-Regional, Diachronic Survey," 2001. [Online]. Available: http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/

[17] T. Baer and B. C. J. Moore, "Effects of Spectral Smearing on the Intelligibility of Sentences in Noise," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1229–1241, 1993.

[18] T. Baer and B. C. J. Moore, "Effects of Spectral Smearing on the Intelligibility of Sentences in the Presence of Interfering Speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2277–2280, 1994.

[19] X. Luo, Q.-J. Fu, H.-P. Wu, and C.-J. Hsu, "Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users," *Hearing Research*, vol. 256, no. 1, pp. 75–84, 2009.

[20] T. Baer, B. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment - effects on intelligibility, quality, and response-times," *Journal of Rehabilitation Research and Development*, vol. 30, no. 1, pp. 49–72, 1993.