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

Contributions of Consonant-vowel Transitions to Mandarin Tone Identification in Simulated Electric-acoustic Hearing

Fei Chen

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
fchen@sustech.edu.cn

Abstract

For hearing-impaired listeners fitted with cochlear implants (CIs), they rely on electric (E) stimulation with primarily slow-varying temporal information but limited spectral information for their speech perception. Many recent studies showed that for those implanted listeners with residual low-frequency hearing, the combined electric-acoustic (E+A) stimulation could significantly improve their speech perception in adverse listening conditions. The present work assessed the contributions of consonant-vowel transitions to Mandarin tone identification via a vocoder based simulation of E+A stimulation. Isolated Mandarin words were processed to preserve full consonants and vowel onsets across consonant-vowel transitions, and replace the rest with noise. The two types of vocoded stimuli, simulating E and E+A stimulations, were presented to normal-hearing Mandarin-speaking listeners to identify lexical tones. Results consistently showed the advantage of E+A stimulation over E-only stimulation when full consonants and the same amount of vowel onset segments were preserved for lexical tone identification. In addition, compared with E stimulation with full vowel segments, the combined-stimulation advantage was observed even when only a small portion of vowel onset segments were presented. Results in this work suggested that in E+A stimulation, segmental contributions were able to provide tone identification benefit relative to E stimulation with the entire Mandarin words.

Index Terms: combined electric-acoustic stimulation, consonant-vowel transitions, noise-replacement paradigm, combined stimulation advantage

1. Introduction

For listeners with profound-to-severe hearing loss, cochlear implants (CIs) are so far the only way for them to restore partial hearing [1]. CI speech processors extract important acoustic cues from the sound signals collected, use them to modulate electric pulse trains, and subsequently stimulate residual auditory nerves to evoke sound perception. Nowadays, implanted listeners enjoy a satisfactory sound perception and speech communication in quiet. However, there are still a number of challenges for CI-based speech communication, including speech perception in noise, music appreciation, etc. Studies to understand the mechanism of electric (E) stimulation underlying the CI technology are still active ongoing. Among many, the combined electric-acoustic (E+A) stimulation is widely recognized as an effective means for improving the speech perception of implanted patients [e.g., 2-11].

The combined E+A stimulation is particularly designed for patients whose low-frequency residual hearing is preserved after implantation. For those patients, E stimulation is employed to restore their perception of high-frequency sounds, and acoustic (A) stimulation provides them low-frequency acoustic information. Compared with E-only stimulation, E+A stimulation is able to make use of the extra acoustic information encoded in the low-frequency region, e.g., fundamental frequency (F0) and its harmonics. Hence, speech understanding under E+A stimulation is significantly improved compared with that based on E-only stimulation, particularly in noise and for tonal languages [e.g., 2-3].

Although the E+A benefit or combined stimulation advantage has been widely established, the accounts for this advantage are still on debate [2-11]. Incerti et al. investigated the effect of varying cross-over frequency (CF) settings for E+A stimulation in one ear combined with acoustic hearing in the opposite ear on binaural speech perception, localization and functional performance in real life, and suggested that implanted patents who used E+A and A stimulation may benefit from access to different CF settings to achieve maximal device usage [8]. Xu et al. examined the advantage of combined E+A stimulation over E-only hearing using an odd-ball paradigm based event-related potential (ERP) experiment and vocoder simulation (simulating CI speech processing, see more in Section 2.2) [12], and demonstrated that compared with the ERP response elicited in the E-only condition, the response in the combined stimulation condition was much closer to that elicited by the full-spectrum stimulus, yielding neurophysiological evidence on the combined-stimulation advantage [9]. Fu et al. evaluated factors that affect the integration efficiency (i.e., the ratio between the observed and predicted performance for acoustic-electric hearing) in E+A and bimodal listening, and their simulation results suggested that acoustic and electric hearing may be more effectively and efficiently combined within rather than across ears, and that tonotopic mismatch should be minimized to maximize the benefit of acoustic-electric hearing, especially for E+A [10]. Chen et al. recently studied the contribution of F0 contour for combined stimulation advantage for Mandarin sentence understanding, and found that the E+A advantage was absent when the F0 contour was manipulated to have a flat trajectory [11].

While early work mainly focused on the spectral contributions, little was done to study the roles of temporal segments for the combined stimulation advantage. More specifically, the segmental or subsegmental contributions for speech perception in E+A stimulation are unclear. Studying segmental contributions has been a common means for us to understanding the perceptual impacts of various speech segments (e.g., consonant-vowel transitions). Noise-

http://dx.doi.org/10.21437/Interspeech.2019-1124

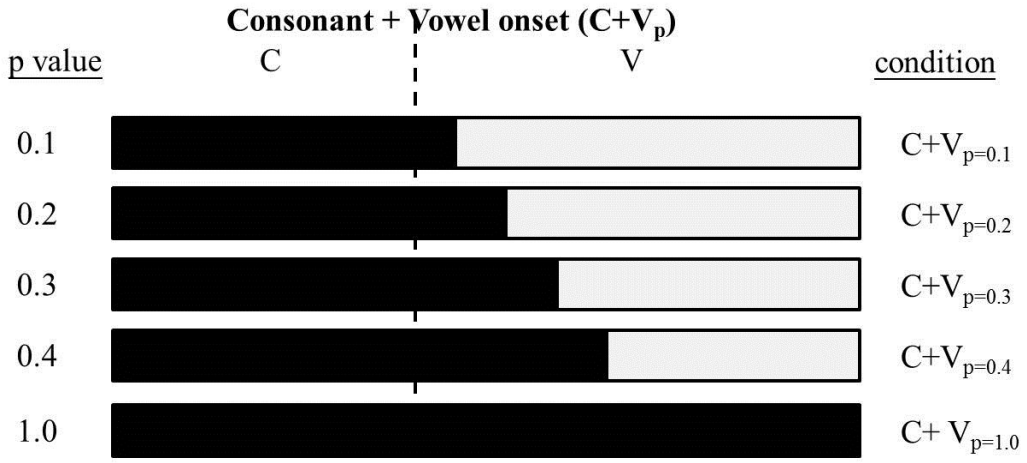

Figure 1: *Schema of the subsegmental conditions created. Dashed line represents the C-V boundary. Black bars represent the speech segments preserved, while white bars represent the speech segments replaced by noise.*

replacement paradigm was long used to examine the segmental contribution to speech perception [13-17]. That is, the segments of interest were preserved, and the rest segments were replaced with either noise or silence. Using the noise-replacement paradigm, many important findings were revealed in early work. For instance, Cole et al. suggested that in English sentences, vowel segments contained much more intelligibility information than consonant segments, rendering an intelligibility ratio of 2:1 [13]. Similarly, the relative perceptual importance of vowels vs. consonants was also found with sentences and isolated words in Mandarin [16-17].

The aim of this work was to study the combined stimulation advantage at the subsegmental level. Specifically, this work examined whether the combined stimulation advantage could be received when full consonants and a portion of vowel onsets (or consonant-vowel transitions) were presented for lexical tone identification in Mandarin. Fundamental frequency contour is the primary cue for lexical tone identification in Mandarin, and it is mainly included in the low-frequency region or acoustic stimulation. This work hypothesized that even if a portion of vowel onsets (across consonant-vowel transitions) were preserved for tone identification, the combined E+A stimulation could still demonstrate its perceptual advantage over E-only stimulation.

2. Methods

2.1. Subjects and materials

Fourteen (eight male and six female) normal-hearing native Mandarin-Chinese listeners participated in the experiment. The subjects' age ranged from 23 to 35 years, and the majority of subjects were undergraduate students at Southern University of Science and Technology. All subjects were paid for their participation. The experimental procedure involving human subjects was approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

Isolated Mandarin words were taken from a database of 1128 isolated Mandarin (monosyllabic) words, covering almost all daily-used words in Mandarin Chinese. All the words were spoken in isolation by a female native-Mandarin talker at a normal speaking rate and with broadcaster's voice quality. The fundamental frequency of recorded words ranged from 130 to 330 Hz [17].

2.2. Signal processing

The stimuli were presented in two different processing conditions. The first processing condition was designed to simulate the effect of six-channel E-only stimulation, and used a six-channel noise vocoder [12]. Signals were first processed through a pre-emphasis highpass filter (2000 Hz cutoff) with a 3 dB/octave rolloff, and then bandpassed into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. The equivalent rectangular bandwidth scale was used to allocate the six channels with the specified bandwidth [18]. The cutoff frequencies for the channel allocation of bandpass filters were (in Hz): 80, 281, 682, 1158, 2060, 3547 and 6000. The envelope of the signal was extracted by full-wave rectification and low-pass filtering using a second-order Butterworth filter (400 Hz cutoff). A white noise was used as the carrier signal, and amplitude-modulated by the extracted envelope. Output from each band was further band-limited with the same bandpass filter at that band. All amplitude-modulated noises (with band-limiting processing) were summed to generate the noise-vocoded stimulus, with its amplitude adjusted to have the same root-mean-square power as the original signal.

The second processing condition simulated the combined E+A stimulation. A LP stimulus was generated by low-pass (LP) filtering the original signal to 600 Hz using a sixth-order Butterworth filter. The 600 Hz cutoff was chosen as it closely mimicked the situation with E+A patients who had residual hearing up to approximately 500–750 Hz and precipitous hearing loss thereafter [e.g., 19-20]. To simulate the effect of E+A with residual hearing below 600 Hz, the LP stimulus was combined with the upper four channels of the above-mentioned six-channel noise vocoder.

Using the above two types of stimulations (i.e., E and E+A), this study further generated consonant plus vowel-onset stimuli, i.e., preserving the whole consonant segments and vowel onset segments while replacing the rest with noise. The C-V boundaries (defined based on traditional segmental boundaries) were labeled manually by an experienced

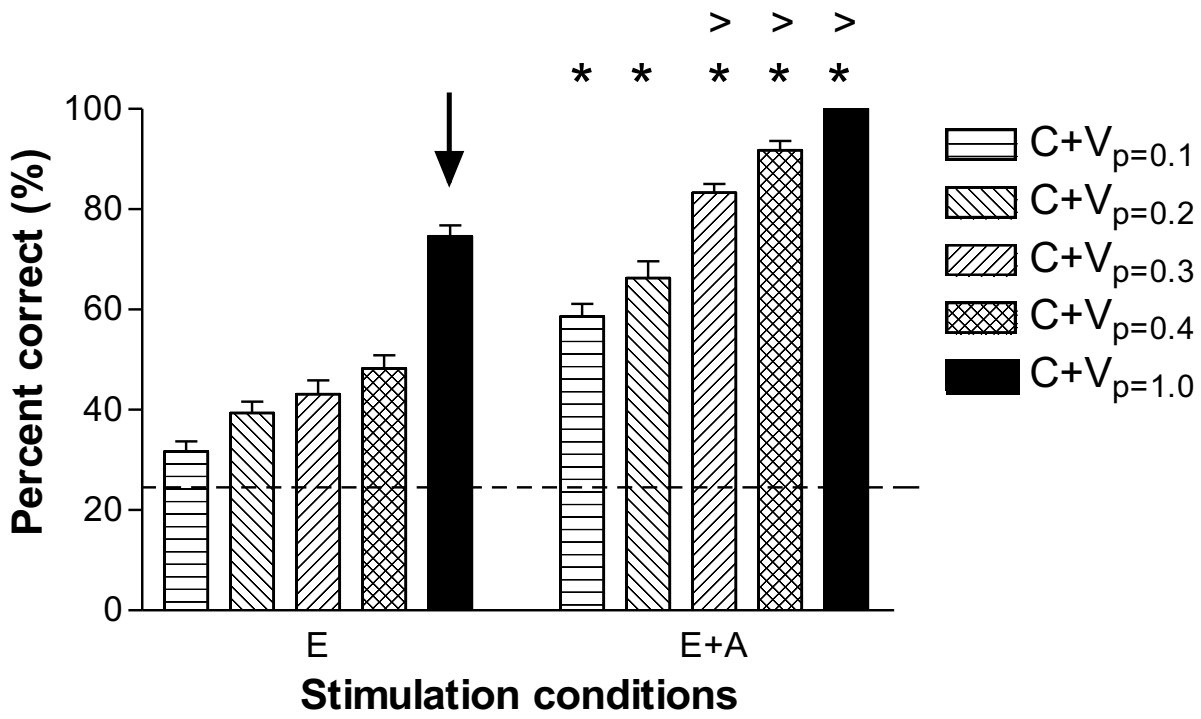

Figure 2: *Mandarin tone identification scores for all conditions. The dashed line indicates chance level (i.e., 25%) for Mandarin tone indentation. Error bar denotes ±1 standard error of the mean. Asterisk denotes that the identification score at the E+A condition is significantly ($p<0.005$) larger than its paired score at the E condition with same proportion factor p. '>' denotes that the score is significantly ($p<0.005$) larger than that at the E condition of $C+V_{p=1.0}$.*

phonetician, and later verified by another experienced phonetician. All final nasal consonants were counted as part of their preceding vowels. The average duration of the words was 468 ms (consonants: 118 ms, vowels: 350 ms). The words include 21 consonants ([p], [ph], [m], [f], [t], [th], [n], [l], [k], [kh], [x], [tɛ], [tɛh], [ɛ], [tʂ], [tʂh], [ʂ], [ʐ], [ts], [tsh], [s]), 35 vowels ([a], [o], [ɤ], [i], [u], [y], [aɪ], [eɪ], [ɑʊ], [oʊ], [ia], [iɛ], [iaʊ], [iəʊ], [ua], [uo], [uaɪ], [ueɪ], [yɛ], [an], [ən], [ɑŋ], [əŋ], [oŋ], [iɛn], [in], [iɑŋ], [iŋ], [iʊŋ], [uan], [uən], [uɑŋ], [ʊəŋ], [yɛn], [yn]), and 4 tones (high level, rising, falling and rising, and falling) [21]. In terms of Mandarin vowels, the 35 vowels in this study consist of 6 simple vowels, 13 complex vowels, and 16 compound nasal vowels [21]. The vowel nuclei with a final nasal were grouped into vowels as many earlier studies classified them as part of vowels in Mandarin [e.g., 21-22].

A proportion factor p was used to denote the proportional duration of vowel onset preserved, and the rest of the segments were replaced by a speech-shaped noise scaled to 16 dB below the level of the intact speech waveform [15]. For instance, p=0.1 meant that 10% vowel onset were preserved and the rest 90% vowel segments were replaced by noise (as illustrated in Fig. 1). This work chose five p values, i.e., p=0.1, 0.2, 0.3, 0.4 and 1.0. Figure 1 demonstrates the 5 subsegmental conditions created, noted as $C+V_{p=0.1}$, $C+V_{p=0.2}$, $C+V_{p=0.3}$, $C+V_{p=0.4}$ and $C+V_{p=1.0}$, respectively.

2.3. Procedure

The experiment was conducted in a sound-proof booth. Participants listened to the stimuli through a circumaural headphone, and the stimuli were binaurally played at a comfortable listening level. Before the experiment, all participants listened to 40 words at the E+A condition of $C+V_{p=0.4}$, and 40 words at the E condition of $C+V_{p=0.4}$ as practice to familiarize them with the noise-replaced stimuli and experimental procedure. Feedback was given in the practice. Each participant attended a total of 10 test conditions [= 2 simulation conditions (E, and E+A) × 5 values of proportion factor (p=0.1, 0.2, 0.3, 0.4 and 1.0)], and each condition had 60 randomly selected isolated Mandarin words (15 words for each lexical tone). Each participant listened to all ten conditions which were presented as blocks in a random order to minimize the effects of practice across participants. Participants were allowed to listen to each stimulus at most three times, and they were asked to choose the Mandarin tone from a custom-designed MATLAB interface The identification score for each condition was computed as the ratio between the number of the correctly identified tones and the total number (i.e., 60) of stimuli tested in each condition. Note that only the recognition of the tones within the words presented was scored. A 5-minute break was given in every 30 minutes. The whole experiment took about one hour per participant.

3. Results

Mean tone identification scores for all conditions are shown in Fig. 2. Statistical significance was determined by using the identification score as the dependent variable, and stimulation condition and vowel onset duration as the two within-subject factors. The scores were first converted to rational arcsine units using the rationalized arcsine transform [23]. Two-way

analysis of variance with repeated measures indicated a significant effect ($F_{1,13}$=481.17, p<.001) of stimulation condition, vowel onset duration ($F_{4,52}$=209.53, p<.001), and a significant interaction ($F_{4,52}$=19.06, p<.001) between stimulation condition and vowel onset duration.

For identification scores grouped at the same value of proportion factor p, statistical analysis showed that all identification scores at the E+A condition are significantly (p<0.005) larger than their paired scores at the E condition, as noted with '*' in Fig. 2. Further analysis compared the scores at the E+A condition with the score at the E condition of $C+V_{p=1.0}$ (marked by arrow in Fig. 2), and results showed that for the E+A conditions of $C+V_{p=0.3}$, $C+V_{p=0.4}$ and $C+V_{p=1.0}$, their scores are significantly (p<0.005) larger than the score at the E condition of $C+V_{p=1.0}$, as noted with '>' in Fig. 2.

4. Discussion and conclusions

The present work studied Mandarin tone identification in the combined E+A stimulation when consonants and only a portion of vowel onset segments were presented. Early work with intact Mandarin isolated words showed that while Mandarin consonant segments contained little information for lexical tone identification, listeners could make use of vowel onsets and vowel-consonant transitions for lexical tone identification [17]. Consistent with early work [17], this study also showed that in the scenarios of E and E+A stimulations, consonants plus a portion of vowel onsets across consonant-vowel transitions contained much information for tone identification. The identification scores at the E and E+A conditions of $C+V_{p=0.1}$ are 31.7% and 58.6%, respectively, and the combined stimulation advantage is clearly seen in this work. In addition, as shown in Fig. 2, for all subsegmental conditions (i.e., p=0.1, 0.2, 0.3 and 0.4), the score at the E+A condition is markedly larger than its counterpart at the E condition. Hence, the combined stimulation advantage is also present under conditions studying the subsegmental contributions to lexical tone identification. It is reasonable to believe the combined E+A stimulation is able to increase the subsegmental contribution for Mandarin tone identification. This advantage may be largely attributed to the added F0 information in A stimulation, although only a portion of F0 contour in A stimulation is added. Note that this work did not test the A stimulation condition, which warrants further investigation.

Second, this work also showed that the subsegmental contribution at the combined E+A stimulation may cause a large perceptual benefit surpassing the performance at E stimulation with the entire word segments. As shown in Fig. 2, the identification score at the E condition of $C+V_{p=1.0}$ (i.e., including the whole vowel segments or the whole isolated Mandarin word) is 74.6%, while the score at the E+A condition of $C+V_{p=0.3}$ (i.e., including 30% of vowel onset segments) is 83.3%. There is a significant difference between the scores of these two conditions. Hence, the advantage of the combined E+A stimulation over E stimulation could be achieved in the absence of full speech segments. In other words, the perceptual benefit of adding A stimulation could be represented by glimpsing a portion of perceptual information (i.e., vowel onsets) contained in the low-frequency region.

In conclusion, the present work studied the combined stimulation advantage in the context of Mandarin tone identification with subsegmental cues in isolated words. The noise-replacement paradigm was used to preserve full consonants and a portion of vowel onsets across consonant-vowel transitions and replace the rest vowel segments with noise. Experimental results showed that, compared with E stimulation, the combined E+A stimulation yielded significant improvement for correctly identifying Mandarin tones, manifesting the combined stimulation advantage at the subsegmental level. In addition, the combined E+A stimulation with a small portion of vowel onsets (e.g., 30%) was able to provide perceptual benefit relative to E-only stimulation with the entire Mandarin words.

5. Acknowledgements

This work was supported by the Basic Research Foundation of Shenzhen (Grant No. JCYJ 20170817110841907), the Research Foundation of Department of Science and Technology of Guangdong Province (Grant No. 2018A050501001), and the Shenzhen High-level Overseas Talent Program (Peacock Plan) (Grant No. KQJSCX20180319114453986).

6. References

[1] Chen, F., Ni, W. L., Li, W. Y., and Li, H. W. (2019). "Cochlear implantation and rehabilitation," in *Hearing Loss: Mechanisms, Prevention and Cure*, Eds: Huawei Li, and Renjie Chai, pp. 129–144.

[2] Chang, J. E., Bai, J. Y., and Zeng, F. G. (2006). "Unintelligible low-frequency sound enhances simulated cochlear-implant speech recognition in noise," *IEEE Trans. Biomed. Eng.* 53, pp. 2598–2601.

[3] Luo, X., and Fu, Q. J. (2006). "Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations," *J. Acoust. Soc. Am.* 120, pp. 2260–2266.

[4] Chen, F., and Loizou, P. C. (2010). "Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing," *Ear Hear.* 31, pp. 259–267.

[5] Oh, S., Donaldson, G., and Kong, Y. Y. (2016). "Top-down processes in simulated electric-acoustic hearing: The effect of linguistic context on bimodal benefit for temporally interrupted speech," *Ear Hear.*, 37, pp. 582–592.

[6] Qin, M., and Oxenham, A. (2006). "Effects of introducing unprocessed low-frequency information on the reception of the envelope-vocoder processed speech," *J. Acoust. Soc. Am.* 119, pp. 2417–2426.

[7] Stilp, C., Donaldson, G., Oh, S., and Kong, Y. Y. (2016). "Influences of noise-interruption and information-bearing acoustic changes on understanding simulated electric-acoustic speech," *J. Acoust. Soc. Am.* 140, pp. 3971–3979.

[8] Incerti, P. V., Ching, T. Y., and Cowan, R. (2019). "The effect of cross-over frequency on binaural hearing performance of adults using electric-acoustic stimulation," *Cochlear Implants Int.* 18, pp. 1–17.

[9] Xu, D., Wang, L., and Chen, F. (2018). "An ERP study on the combined-stimulation advantage in vocoder simulations," *Proceedings of 40th Annual International Conference of the IEEE-EMBS*, pp. 2442–2445

[10] Fu, Q. J., Galvin, J. J. 3rd, and Wang, X. (2017). "Integration of acoustic and electric hearing is better in the same ear than across ears," *Scientific Reports* 7(1), 12500.

[11] Chen, F., and Chen, J. (2018). "Effects of fundamental frequency contour on understanding Mandarin sentences in bimodal hearing simulations," *J. Acoust. Soc. Am.* 143, pp. EL354–EL360.

[12] Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* 270, pp. 303–304.

[13] Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (1996). "The contribution of consonants versus vowels to word recognition in fluent speech," *Proceedings of the IEEE*

International Conference on Acoustics, Speech, and Signal Processing, pp. 853–856.
[14] Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.* 122, pp. 2365–2375.
[15] Fogerty, D., and Kewley-Port, D. (2009). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.* 126, pp. 847–857.
[16] Chen, F., Wong, L. L. N., and Wong, Y. W. (2013). "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," *J. Acoust. Soc. Am.* 134, pp. EL178–EL184.
[17] Chen, F., Wong, M. L. Y., Zhu, S. F., and Wong, L. L. N. (2015). "Relative contributions of vowels and consonants in recognizing isolated Mandarin words," *J. Phonetics* 52, pp. 26–34.
[18] Glasberg, B., and Moore, B. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear Res.* 47, pp. 103–138.
[19] Kiefer, J., Pok, M., Adunka, O., et al. (2005). "Combined electric and acoustic stimulation of the auditory system: Results of a clinical study," *Audiol Neurootol*, 10, pp. 134–144.
[20] Gantz, B. J., Turner, C., and Gfeller, K. E. (2006). "Acoustic plus electric speech processing: Preliminary results of a multicenter clinical trial of the Iowa/Nucleus Hybrid implant," *Audiol Neurootol*, 11, pp. 63–68.
[21] Yin, B., and Felley, M. (1990). *Chinese Romanization: Pronunciation and Orthography* (Sinolingua, Beijing, China).
[22] Fu, Q. J., Zhu, M., and Wang, X. S. (2011). "Development and validation of the Mandarin speech perception test," *J. Acoust. Soc. Am.* 129, pp. EL267–EL273.
[23] Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hearing Research* 28, pp. 455–462.