INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Segment and Tone Production in Continuous Speech of Hearing and Hearing-impaired Children

Shu-Chuan Tseng[1], *Yi-Fen Liu*[2]

[1]Institute of Linguistics, Academia Sinica, Taiwan
[2] Department of Information Engineering and Computer Science, Feng-Chia University, Taiwan
tsengsc@gate.sinica.edu.tw, yifenliu@gmail.com

Abstract

Verbal communication in daily use is conducted in the form of continuous speech that theoretically is the ideal data format for assessing oral language ability in educational and clinical domains. But as phonetic reduction and particularly lexical tones in Chinese are greatly affected by discourse context, it is a challenging task for automatic systems to evaluate continuous speech only by acoustic features. This study analyzed repetitive and storytelling speech produced by selected Chinese-speaking hearing and hearing-impaired children with distinctively high and low speech intelligibility levels. Word-based reduction types are derived by phonological properties that characterize contraction degrees of automatically generated surface forms of disyllabic words. F0-based tonal contours are visualized using the centroid-nearest data points in the major clusters computed for tonal syllables. Our results show that primary speech characteristics across different groups of children can be differentiated by means of reduction type and tone production.

Index Terms: Child speech, continuous speech, segment reduction, tonal contour

1. Introduction

1.1. Spoken words and oral language assessment

Spoken word production and perception are widely studied in interdisciplinary research studies. Hearing ability, topic familiarity, word frequency, and contextual information are related to the efficacy of speech communication [1], [2], [3], [4], [5], [6]. Within an integrated cognitive-linguistic knowledge system that enables verbal communication, words connect linguistic forms with cognitive components, so quantitative descriptions are often deduced from properties of spoken words [7], [8]. In the contemporary use of Mandarin Chinese, mono- and disyllabic words are the majority [9]. For pronunciation variants of disyllabic words, canonical form and syllable merger are the two most frequent forms found in adult conversation [10]. Nevertheless, given a sufficient degree of phonetic and semantic predictability, extremely reduced words in real speech can be understood without difficulties [10], [11].

In studies of language acquisition and speech pathology, oral language ability is normally assessed by impressionistic judgments conducted by experts [12], [13], [14], [15], [16]. Automatic aids that utilize computational models enhanced with acoustic information accordingly reflecting phonological contrasts have proved helpful, e.g., spectral features of speech sounds [17], [18], duration and fundamental frequency (F0) [19], [20]. For a tone language like Mandarin Chinese, F0 contour is related to the meaning associated with syllables having the same syllable structure. For instance, the F0 contour pattern is accordingly different in the use of 那/na/ as a demonstrative filler, a determiner, and a connective in adult conversation [21]. Most studies of these kinds require signal-aligned data. But unlike adults' data, children's data, especially comparative data produced by hearing and hearing-impaired children are rare [22], [23].

One of the main goals of oral language assessment for children is to identify the level of speech intelligibility and then further specify their difficulties [17]. While carrying out the task of impressionistic judgments, we need to deal with the issue whether to apply criteria of accuracy, appropriateness, or acceptability for annotating variants of spoken words [24]. For children with hearing impairment, whose access to acoustic information is limited, it is even more difficult to handle individual differences [25]. Hearing loss degree, age of diagnosis and intervention, type of hearing aids, and implantation period etc. can all play a role [14], [26], [27], [28]. This paper proposes two types of quantitative measures reflecting word-level reduction and tone production to illustrate that they may serve as useful acoustics-based indicators for early diagnosis of child speech problems.

1.2. Tone production in child speech

Mandarin Chinese has four lexical tones: High-level (T1), low-rising (T2), low-falling-rising (T3), high-falling (T4), and one unstressed, neutral tone [29], [30]. They constitute a system of contrastive pitch contours explicitly associated with lexical meaning. Tonal variants are context-dependent. T3 is seldom a full-fledged dipping tone in continuous speech, but often a half T3, in which the rising part of T3 is omitted, resulting in a short, low-falling contour [30]. T3 can also be pronounced as a T2-like sandhi T3, when it is immediately followed by another T3.

Developing children acquire tones earlier than speech sounds [13], [15], usually before the age of three. T1 and T4 are acquired earlier than T2 and T3 [12]. Stimuli used for tone-related production, recognition, and discrimination experiments are either mono- or disyllables [13], [32], [33]. Tone pairs contrasting with T4 were found easier to identify than those contrasting with the other tones [34]. It seems that high-registered tones (T1/T4) are less problematic for hearing and hearing-impaired children to acquire than low ones (T2/T3). Prelingually deaf children with cochlear implants were better at producing T1/T4 than T2/T3 in mono- and disyllabic words as well [34]. T2 is the most severely impaired tone in children with cochlear implants, followed by T3 and T4 [35]. As mentioned earlier, tone production varies by context, results and conclusions of previous studies are dependent on the tasks and the stimuli of the conducted experiments. Mostly, spoken words examined are produced in isolation. In this study, we will show that automatically derived word-level segment reduction and

http://dx.doi.org/10.21437/Interspeech.2021-757

tonal contours provide empirical cues to speech characteristics of continuous speech that can show differences in groups of hearing and hearing-impaired children with high and low speech intelligibility levels.

1.3. Sinica Child Speech Corpus

The *Sinica Child Speech Corpus* consists of continuous speech recordings of 79 preschool children with normal hearing (NH) and 45 children with hearing impairment. Among them, 30 wore traditional hearing aids with mild to profound degrees of hearing loss (HA), and 15 were fitted with a cochlear implant with severe to profound degrees of hearing loss (CI) [9]. For the repetitive data collection, eighteen phonemically balanced sentences recorded by a female adult were played to the children one by one for them to repeat. Pictures illustrating the contents of the sentences were made available on the computer screen. Speech intelligibility were scored by three phoneticians according to the degree of fluency and the intelligibility of segments at a scale from 1 to 5. The sum was used as the final intelligibility score. For the storytelling data collection, the children were instructed to tell the story *The Hare and the Tortoise* with picture cards that were presented to them in an arranged order. For the current study, the speech content was orthographically transcribed in Chinese characters and the *ILAS phone aligner* [9] was applied to obtain signal-aligned boundary information for phonemes, syllables, and words with no human intervention.

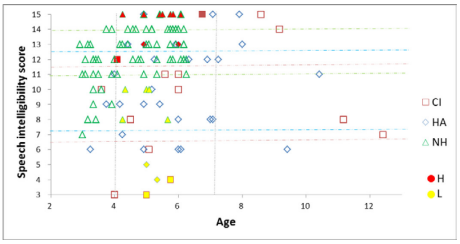


Figure 1: *Subject data, color-matched horizontal lines designate the 1st and 3rd quartiles for each group.*

Distribution of intelligibility score and age of the subjects in the *Sinica Child Speech Corpus* is given in Figure 1. Recent works of deep models on this dataset performed very different results in some of the subjects [17]. We formed a more homogeneous subject group with children with distinctively high and low speech intelligibility levels in this study. Gender-balanced children aged from 4 to 7 were selected, whose intelligibility levels are higher than the 3rd quartile as high (H), and those lower than the 1st quartile as low (L) intelligibility subgroups in NH, HA and CI children. CI_H, HA_H, NH_H consists of 2, 2, and 10 children, and CI_L, HA_L, NH_L 2, 2, and 5 children.

1.4. Categorical reduction types

Segment deletion is indicative of the degree of phonetic reduction. As spoken words in continuous speech are full of variability, we adopted a system of categorical reduction types to obtain preliminary, but effective diagnosis for phonetic reduction. Four reduction types are distinguished based on the presence of word-internal syllable boundary and non-vocalic segments [6]. Table 1 shows that Canonical Form (CAN) retains all cross-syllable non-vocalic segments, while Marginal Segment Deletion (MSD) still has a clear syllable boundary but with some of the existing non-vocalic segments omitted. The syllable boundary in Nucleus Merger (NUM) is blurred; the two nuclei are partially merged. In the case of Syllable Merger (SYM), the disyllabic word is merged into a monosyllable.

Table 1: *Categorization principles of reduction types.*

	Word-internal syllable boundary	Cross-boundary segment deletion
CAN	+	-
MSD	+	+
NUM	+/-	+
SYM	-	+

Surface forms of disyllabic words in the repetitive and the storytelling speech produced by the 23 children are generated by free phone recognition using acoustic models trained by adults' speech [9]. No human intervention was involved in the entire process. The categorization principles in Table 1 were applied by taking into accounts of individual syllable types, as proposed in [10]. The *ILAS phone aligner* and the reduction type algorithm designed for adults' speech [10] seem to perform well for our child speech data, as shown in Figure 2.

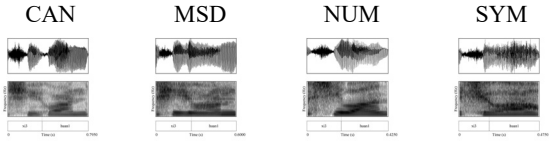


Figure 2: *Four reduction types of xīhuān (like).*

1.5. Tone visualization approach

Different from conventional approaches that examine single F0 values for tonal properties, we suggest to examine the contours by an approach that applies features reflecting phonological contrast of tones in linearized F0 contours and selects the centroid-nearest data points from the major clusters to visualize tonal contours [21], [36]. Making use of the syllable boundary information obtained by the *ILAS phone aligner*, F0 values are extracted by PRAAT first [31]. F0 ranges and normalization are calculated for each speaker, respectively [21]. Five acoustic-phonetic features are implemented in the clustering experiment: (1) the normalized onset F0 value, (2) the slope of the first fitted line, (3) the slope of the second fitted line (if any), (4) the (normalized) time of the turning point in the case of two fitted lines and (5) the normalized offset F0 value. Fuzzy *c*-mean (FCM) is an unsupervised learning clustering [37], [38] that has been successfully applied to a number of problems involving clustering on small feature sets with a limited sample size. We treated data from each tone category as unlabeled data. For a given set of data, we aim to partition all tonal syllables into clusters in which data points from the same cluster are as similar to each other as possible by minimizing an objective function J_m [39], based on a norm and clusters prototype, which is defined as follows:

$$J_m(\boldsymbol{U},\boldsymbol{V};X)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^{m}\,\|x_k-v_i\|_A^2,\quad 1<\mathrm{m}<\infty, \qquad (1)$$

where $V = (v_1,v_2,\ldots,v_c)$ is a vector of unknown cluster prototypes. The value of u_{ik} represents the degree of the membership of data point x_k to the ith cluster. The inner product defined by a norm matrix A shows the measure of similarity between a data point and the cluster prototypes. Techniques based on a fuzzy set decomposition minimize the weighted sum of squared errors within groups as defined in Eq. (1), all the parameters of the clustering method are fixed when J_m is

converged to a local minimum or a saddle point, except for the number of cluster c, which often unknown *a priori*. Thus, when the FCM is used for finding a partition of data under a fixed number of clusters (objects), we still need a cluster validity procedure to determine automatically the optimal number of clusters (the plausible tonal varieties for each tone category in our task). As desired, a number of validation indices for the fuzzy c-mean arose to find the optimal number of data clusters based on extrema (the minima) of these validation values for all possible $c_i \in \{2, 3,\ldots, c_{\max}\}$. One validation index, V_{CWB}, proposed by Rezaee *et al.* [39] is a validation functional for FCM method in consideration of both the *compactness* and the *separation* of a fuzzy c-partition.

$$V_{CWB}(\boldsymbol{U},\boldsymbol{V}) = \alpha \mathrm{Scat}(c) + \mathrm{Dis}(c), \qquad (2)$$

where Scat(c) indicates the average scattering (*compactness*) *within* the clusters. Dis(c) indicates the total scattering (*separation*) *between* the clusters. As well, a weighting factor α, set as Dis($c_{\max}$), is used for counterbalancing both terms in a proper way. Operated under such an FCM validation algorithm proposed in [39], an optimal number of clusters c of a tone category can be obtained and the data point closest to the centroid in the major cluster then is chosen for visualization. For the current study, we only selected syllables with a C(G)V structure to exclude contextual influences from the nasal coda. To evaluate the reliability of the proposed tone visualization approach, we applied it on a dataset consisting of all words appearing in the *Sinica Core Vocabulary* [9] read by a female adult. Figure 3 illustrates the centroid-nearest tonal contours of the four lexical tones, *shū* (book), *lái* (come), *bǐ* (pen), *zài* (in). The contours generated from our tone visualization approach clearly conform to the standard phonological descriptions of the four lexical tone categories in Chinese.

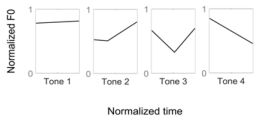


Figure 3: *Canonical contours as visualized by our approach*

2. Reduction in continuous speech

2.1. More reduced speech in hearing-impaired children

For adult conversation, high-frequency disyllabic words are mostly assigned with the extreme reduction type SYM [10]. From a production-based perspective, this may be empirical evidence supporting the notion that syllable merger is phonologically represented in the Chinese mental lexicon [5]. In the current study of child speech, reduction type is primarily used to categorize the degree of phonetic reduction. Figure 4 shows results of 1,355 disyllabic words produced in both continuous speech styles. The two edge forms, the canonical form CAN and the syllable merger form SYM, are produced more often than MSD and NUM, quite similar to adults' results [10]. In general, SYM appears more often in HA and CI children than in NH children, while the trend of CAN production is the opposite. The preference for SYM production is an essential cue for a larger degree of incompleteness of phonetic forms. For the repetitive speech, children scored higher in speech intelligibility (NH_H, HA_H, CI_H) preferred CAN production, suggesting that segment reduction are less severe in high-intelligibility hearing as well as hearing-impaired children.

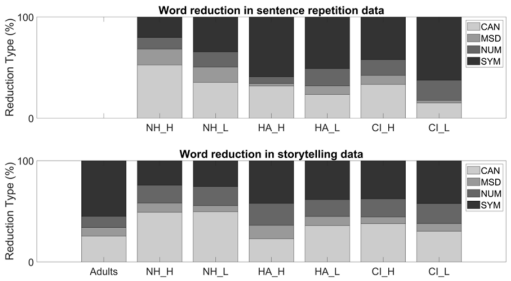


Figure 4: *Disyllabic word reduction in CI, HA, and NH, with adults' data adopted from [10].*

In the repetitive speech, differences in high versus low intelligibility as well as across NH, HA, and CI subgroups are also well represented by reduction types, in particular by the occurrences of CAN, as the Kendall tau correlation tests of reduction type distribution in ratios show no significance across the subject groups. For the storytelling speech, NH_H and NH_L are correlated, Spearman's ρ=0.995, p<0.01 as well as CI_H and HA_L, Spearman's ρ=0.994, p<0.01. Basically, hearing children, both high and low intelligibility subgroups, behave more similarly to each other compared to the hearing-impaired children. Interestingly, for hearing children, results of the repetitive speech are distinctive in high and low intelligibility children, but those of the storytelling speech do not.

Moreover, CI children's reduction in the storytelling speech is similar to that in the repetitive speech. But HA children, on the other hand, show greater discrepancies in the two data types, viewed from the production of CAN and SYM. This may be due to individual differences of the children, as our sample is really small. Nevertheless, according to our current results, spontaneous speech might not be the ideal speech format for all tasks of oral language ability evaluation, as more controlled data such as repetitive speech seem to be more conclusive for segment production assessment.

2.2. Segment reduction and duration

Duration is one of the most studied prosodic features, which is often used for speech tempo, rhythm, boundary effect, and phonetic reduction [7], [11]. Reduction type, as proposed in this paper, is a categorical variable. If it is properly defined, it should reflect the property of duration, as theoretically, when a disyllabic word is classified as a more reduced type, the surface form that is generated from the free phone recognition results would contain fewer phones and be produced with a shorter duration. Please note here that the reduction type is solely determined by the number of nuclei and the cross-syllable non-vocalic segment omission. Exact match on blurred, high-order vocalic structure across syllable boundary is requisite, but no direct comparison of the surface and the canonical phone sequences is involved in the derivation procedure. As reduction may be related to word morphology and within-utterance position [10], we conducted linear mixed-effect models separately for the repetitive and storytelling speech data with REDUCTIONTYPE as a fixed effect and IPUPOSITION, SPEAKER, and WORD as random effects. IPUPOSITION distinguishes initial, medial, final, and isolated positions; WORD includes disyllabic words only. The results in Table 2 confirm the expected tendency. SYM is significantly shorter than CAN in both speech styles. Figure 5 illustrates a clear correlation between

the categorically defined reduction types and the trend of gradient quantity of duration in the storytelling data.

Table 2: *Linear mixed-effect models of duration (***=p<0.001)*

Repetitive	REDUCTIONTYPE	Estimate	Std. error	t-value	p-value
Tokens: 525	(Intercept)-CAN	0.660	0.034	19.634	***
DF: 521	MSD	-0.029	0.025	-1.155	0.249
	NUM	-0.016	0.022	-0.721	0.471
	SYM	-0.102	0.017	-6.110	***
Storytelling	REDUCTIONTYPE	Estimate	Std. error	t-value	p-value
Tokens: 830	(Intercept)-CAN	0.599	0.027	22.504	***
DF: 826	MSD	-0.022	0.022	-0.995	0.320
	NUM	-0.077	0.015	-5.018	***
	SYM	-0.121	0.013	-9.074	***

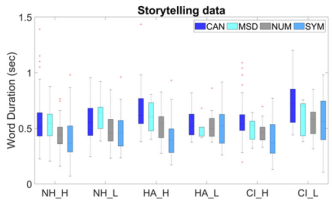


Figure 5: *Duration patterns in terms of reduction types*

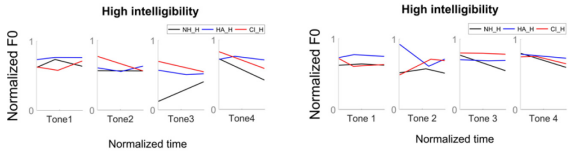
3. Tones in continuous speech

The proposed tone visualization approach was conducted to 1,981 syllables in the repetitive speech and 2,936 syllables in the storytelling speech. Phonological properties of canonical tones are summarized in Table 3.

Table 3: *Phonological description of canonical tones*

	T1	T2	T3	T4
Onset register	high	low	low	high
Contour	flat	rising	dipping	falling

Because the contours selected by our tone visualization approach are too diverse to obtain conclusive results in the low intelligibility subjects, we will only discuss results of the high intelligibility subjects as illustrated in Figure 6a and 6b by focusing on the level of onset register and contour contrast.



(a) (b)

Figure 6: *(a) Repetitive speech; (b) Storytelling speech.*

3.1. Repetitive speech: T2/T3 more problematic

Previous works suggest that hearing and hearing-impaired children acquire the high-registered T1 and T4 earlier than T2 and T3. Results of our tone visualization approach have shown that T1 and T4 have a high-pitched onset register in all three high intelligibility subgroups. In the NH_H children, the T4 contour clearly conforms to the canonical tone. Although the HA_H and CI_H children produced T4 with different slopes, the falling trend retains. T1 contours produced by the three subgroups all have a relatively high onset register level, though the contours are flat and slightly rising or falling. Moreover, T2 and T3 produced by the NH_H and HA_H children are either low- or mid-registered, which conform to the canonical tonal properties. T3 contours produced by the NH_H children is low-rising, similar to a sandhi-T3, while the T2 contour is rather flat, not exactly conforming to the canonical rising contour of T2, but probably expected for continuous speech. It is also noteworthy that the CI_H children produced T2 and T3 with very similar tonal properties, both having a very high onset level and a clearly falling contour. As a whole, for all three high intelligibility subgroups, T1 and T4 produced in the repetitive speech to some extent conform to the canonical tonal properties as those produced in isolation. Table 4 shows that hearing-impaired children produced T2 and T3 with more deviations.

Table 4: *Tone production of high-intelligibility children*

Onset register (√ canonical tonal property)				
	T1	T2	T3	T4
NH_H	high (√)	mid (√)	low (√)	high (√)
HA_H	high (√)	mid (√)	mid	high (√)
CI_H	high (√)	high	high	high (√)
Contour (+ similar to canonical tonal contour)				
	T1	T2	T3	T4
NH_H	slightly falling (+)	flat	rising (+)	falling (+)
HA_H	flat (+)	slightly rising (+)	falling	falling (+)
CI_H	slightly rising (+)	falling	falling	falling (+)

3.2. Storytelling speech: Higher level of variability

A high degree of variability is expected for tones produced in the storytelling data. The production of tones may be affected by lexical choices and intonation declination in continuous speech. But still, T1 and T4 remain quite similar to the canonical contours in Figure 6b. T4 has a high onset level and a falling contour. T1 is basically also high-registered and flat. But it is not the case for T2 and T3. Even for the high-intelligibility hearing subgroup, where we have ten children, their T2 and T3 deviate from the canonical forms. As mentioned earlier, tone varies by context. Thus, we would suggest that for evaluating tone production in continuous speech, repeating or reading out loud sentences with a well-designed vocabulary is probably more suitable than spontaneous speech. So far, we have only used F0 for the tonal contour presentation. Perceptual experiments with extensive parameters [40] may be necessary for studies of more acoustic features involved in the production and perception of tones in continuous speech.

4. Conclusions

Most works on children's oral language ability assessment are conducted by experts, teachers, and therapists [12], [33], [34], [41]. This paper proposed two quantitative measures by applying speech processing techniques and linguistically motivated computational models. Categorical reduction type and tone visualization approach seem to be useful for identifying differences of speech characteristics in hearing and hearing-impaired children, i.e., more SYM production and tonal contour deviation of T2 and T3 production in CI and HA children than in NH children. We have presented limited evidence with a small number of subjects. Large-scale normative datasets, quantitative measures, and subject groups are needed for comparative studies on children with different types of speech impairment, so that efficient automatic systems and related training programs can be accordingly developed for clinical and education purposes [42].

5. Acknowledgements

The study presented in this paper was financially supported by the Children's Hearing Foundation, the National Science Council of Taiwan, under grant NSC 99-2410-H-001-097, and the Ministry of Science and Technology of Taiwan, under grant MOST 106-2410-H-001 -045 -MY2 and 108-2410-H-001-041.

6. References

[1] C. M. Connine, J. Mullennix, E. Shernoff, and J. Yelen, "Word familiarity and frequency in visual and auditory word recognition," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6), 1084-1096, 1990.

[2] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: The Neighborhood Activation Model," *Ear and Hearing*, 19 (1), 1-36, 1998.

[3] S. D. Goldinger, P. A. Luce, D. B. Pisoni, and J. K. Marcario, "Form-based priming in spoken word recognition: The roles of competition and bias," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(6), 1211-1238, 1992.

[4] J. Bybee, *Phonology and Language Use*. Cambridge University Press, Cambridge, 2001.

[5] J. B. Pierrehumbert, "Exemplar dynamics: word frequency, lenition, and contrast," In Bybee and Hopper (Eds.). *Frequency effects and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, 137-157, 2001.

[6] S.-C. Tseng, A. Soemer, T.-L. and Lee, "Tones of reduced T1-T4 Mandarin disyllables," *International Journal of Computational Linguistics and Chinese Language Processing*, 18, 81-106, 2013.

[7] D. Jurafsky, A. Bell, M. Gregory, and D. Raymond, "Probabilistic relations between words: evidence from reduction in lexical production," In Bybee and Hopper (Eds.) *Frequency and the emergence of linguistic structure*, pp. 229-54. Amsterdam: John Benjamins, 2001.

[8] C. Meunier and R. Espesser, "Vowel reduction in conversational speech in French: The role of lexical factors," *Journal of Phonetics*, 39(3), 271-278, 2011.

[9] S.-C. Tseng, "ILAS Chinese spoken language resources," *Proceedings of LPSS 2019*, pp. 13-20. Taipei, 2019.

[10] Y.-F. Liu, S.-C. Tseng, R. Jang, "Deriving disyllabic word variants from a Chinese conversational speech corpus," *Journal of the Acoustical Society of America*, 140(1), 308-321, 2016.

[11] M. Ernestus, H. Baayen, R. Schreuder, "The recognition of reduced word forms," *Brain and Language*, 81, 162-173, 2002.

[12] C. N. Li and S. A. Thompson, "The acquisition of tone in Mandarin-speaking children," *Journal of Child Language*, 4, 185-199, 1977.

[13] Z. Hua and B. Dodd, "The phonological acquisition of Putonghua (Modern Standard Chinese)," *Journal of Child Language*, 27, 3-42, 2000.

[14] S. Peng, A. L. Weiss, H. Cheung, and Y. Lin, "Consonant production and language skills in Mandarin-speaking children with cochlear implants," *Arch Otolaryngol Head Neck Surg*, 130, 92-97, 2004a.

[15] L. M. Chen, and R. Kent, "Development of prosodic patterns in Mandarin-speaking infants," *Journal of Child Language*, 36, 73-84, 2009.

[16] C. K. To, P. S. Cheung, and S. McLeod, "A population study of children's acquisition of Hong Kong Cantonese consonants, vowels, and tones," *Journal of Speech, Language, and Hearing Research*, 56, 103-122, 2013.

[17] Y.-S. Lin, and S.-C. Tseng, "Classifying speech intelligibility levels of children in two continuous speech styles," *ICASSP 2021,* 7748-7752.

[18] S.-C. Tseng, K. Kuei, and P.-C. Tsou, "Acoustic characteristics of vowels and plosives/affricates of Mandarin-speaking children with hearing impairment," *Clinical Linguistics and Phonetic*s, 25, 784-803, 2011.

[19] L. Xu, Y. Tsai, and B. E. Pfingst, "Spectral and temporal features of stimulation affecting tonal speech perception: Implication for cochlear prostheses," *Journal of Acoustic Society of America*, 112, 247–258, 2002.

[20] J. Barry, P. Blamey, L. Martin, K. Lee, T. Tang, Y. Ming, and C. Van Hasselt, "Tone discrimination in Cantonese-speaking children using a cochlear implant," *Clinical Linguistics and Phonetics*, 15, 79-99, 2002.

[21] S.-C. Tseng, "Chinese Demonstratives and their Spoken Forms in a Conversational Corpus," *Journal of the Phonetic Society of Japan*, 21(3), 41-52, 2017.

[22] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," *Proceedings of SSPR-2003* (Tokyo, Japan), pp. 7-12, 2003.

[23] B. Schuppler, M. Ernestus, O. Scharenborg, and L. Boves, "Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions," *Journal of Phonetics*, 39, 96-109, 2011.

[24] S. McLeod, and E. Baker, *Children's Speech: An evidence-based approach to assessment and intervention*. Pearson. Boston. 2017

[25] T. A. Serry and P. J. Blamey, "A 4-year investigation into phonetic inventory development in young cochlear implant users," *Journal of Speech and Hearing Researc*h, 42, 141-54, 1999.

[26] V. Mildner, B. Šindija, and K. V. Zrinski, "Speech perception of children with cochlear implants and children with traditional hearing aids," *Clinical Linguistics and Phonetics*, 20, 219-229, 2006.

[27] S. Nittrouer and L. T. Burton, "The role of early language experience in the development of speech perception and language processing abilities in children with hearing loss," *The Volta Review*, 103(1), 5-37, 2003.

[28] A. M. Robbins, B. K. Koch, M. J. Osberger, S. Zimmerman-Phillips, and L. Kishon-Rabin, "Effect of age at cochlear implantation on auditory skill development in infants and toddlers," *Arch Otolaryngoloy Head Neck Surgery*, 130, 570-574, 2004.

[29] Y.-H. Lin, *The Sounds of Chinese*. Cambridge: Cambridge University Press, 2007.

[30] S. Duanmu, *The Phonology of Standard Chinese*. Oxford University Press, 2000.

[31] P. Boersma and D. Weenink, Praat: doing phonetics by computer. http://www.fon.hum.uva/praat 5.4, 2014.

[32] P. Wong, R. G. Schwartz, and J. J. Jenkins, "Perception and Production of Lexical Tones by 3-Year-Old, Mandarin-Speaking Children," *Journal of Speech, Language, and Hearing Research*, 48, 1065-1079, 2005.

[33] P. Wong, "Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions," *Journal of Phonetics*, 40, 141-151, 2012.

[34] S. Peng, J. B. Tomblin, H. Cheung, Y. Lin, and L. Wang, "Perception and production of Mandarin tones in prelingually deaf children with cochlear implants," *Ear and Hearing*, 25, 251-64, 2004b.

[35] D. Han, N. Zhou, Y. Li, X. Chen, X. Zhao, and L. Xu, "Tone production of Mandarin Chinese speaking children with cochlear implants," *International Journal of Pediatric Otohinolaryngology*, 71, 875-80, 2007.

[36] S.-C. Tseng, "Speech production of Mandarin-speaking children with hearing impairment and normal hearing," *ICPhS*, pp. 2030-2033, 2011.

[37] J. C. Bezdek, *Patten Recognition with Fuzzy Objective Function Algorithms*. Plenum Press: New York, 1981.

[38] J. C. Bezdek, *Partition structures: A tutorial*. In: Bezdek, J.C. (Ed.), The analysis of Fuzzy Information. CRC Press, Boca Raton, FL., 1987.

[39] M. R. Rezaee, B. P. F. Lelieveldt, and J. H. C. Reiber, "A new cluster validity index for the fuzzy *c*-mean," *Pattern Recognition Letters*, 19, 237-246, 1998.

[40] K. M. Yu, "The role of time in phonetic spaces: Temporal resolution in Cantonese tone perception," *Journal of Phonetics*, 65, 126-144, 2017.

[41] H.-F. Chang, H.-Y. Gu, and J. H. Wu, "A pilot study on human listener evaluation and computerized tone-contour analysis of Mandarin disyllable utterances by hearing-impaired students," *Bulletin of Special Education*, 26, 221-245, 2004.

[42] H.-F. Chang, "Visual feedback training to promote Mandarin disyllabic tone perception and production in hearing impaired children," *Bulletin of Special Education*, 32(4), 47-64, 2007.