INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Model-assisted Lexical Tone Evaluation of three-year-old Chinese-speaking Children by also Considering Segment Production

Shu-Chuan Tseng[1], *Yi-Fen Liu*[2] *and Xiang-Li Lu*[2]

[1]Institute of Linguistics, Academia Sinica
[2]Department of Information Engineering and Computer Science, Feng-Chia University
tsengsc@gate.sinica.edu.tw, yfliu@fcu.edu.tw, M1105831@o365.fcu.edu.tw

Abstract

This paper presents a hybrid workflow for lexical tone evaluation of 3-year-old Chinese-speaking children. The speech data of 123 children were phonetically transcribed for phoneme accuracy as well as perceptually evaluated for tone accuracy by human judgement. A transformer-based tone model with a BERT input architecture was built using the speech data and tested on twelve children with low speech performance. The accuracy rates between the judged tones and the predicted tones output by our model were high for the overall evaluation. More consistent patterns between judged and predicted tones were observed for high-register Tone 1 and Tone 4 for low-register Tone 2 and Tone 3. We also found that a child's tone production ability is consistently reflected in relation to consonants, vowels, and syllables. Tone accuracy is more related to vowel accuracy than consonant accuracy. In particular, the most diverse differences in tone, consonant, and vowel accuracies were observed for Tone 3.

Index Terms: tone evaluation, F0, contour representations, BERT input architecture, Transformers, speech assessment, screening tool

1. Introduction

Phonological development in children is conventionally evaluated using pronunciation error patterns. Additionally, on the basis of empirical evidence, interdisciplinary studies combining phonetic research and assistive technology for speech assessment of children have recently attracted intensive attention with computational models constructed using annotated data. Ground truth labeling, such as labeling "typically developing" or "at risk" groups, speech disorder severity levels, intelligibility levels or nonnative speech, significantly contributes to the effectiveness of model training and the improvement of diagnostic assessment systems [1, 2, 3, 4, 5]. Acoustics-based models have been proven useful for making contributions to speech assessment. Acoustic features that reflect the phonetic properties of speech are used to provide assistive support in disease evaluation and progression monitoring in dysarthria-related diseases such as Parkinson's disease and multiple sclerosis [6].

In this study, we are concerned with lexical tone evaluation in Mandarin Chinese. Mandarin Chinese is a tonal language spoken in mainland China and Taiwan. A written character represents a syllable with a specific lexical tone. Comprehending a word means being able to interpret the meaning of the phonetic sequences and their associated tone sequences. Therefore, the evaluation of a child's speech performance should also account for the accuracy or acceptability of the production of lexical tones. To examine lexical tones, we propose a hybrid workflow that includes both perceptually judged results and assistive technology that uses acoustic features. In addition, considering tone and segment production in three-year-old children, we will also demonstrate that predictions using the tone model are feasible, as tone patterns are similar to those judged by humans.

A syllable in Mandarin Chinese has a maximum number of four phonemic segments, i.e., an onset, a glide, a vowel, and a coda of /n/ or /ŋ/. The phoneme inventory of Mandarin Chinese consists of two glides /j, w/, 15 vowels /i, ɨ, ɯ, u, y, a, o, ə, e, ɚ, ai, ei, au, ou, ye/, and 22 consonants /p, p^h, t, t^h, k, k^h, f, s, ʂ, ç, x, ʐ, ts, ts^h, tʂ, $tʂ^h$, tç, $tç^h$, m, n, ŋ, l/ [7, 8]. The tone inventory consists of four full lexical tones: Tone 1, Tone 2, Tone 3, Tone 4, and a neutral tone that represents high-level, low-rising, low-dipping, high-falling, and short mid-level tonal contours. Tones are conventionally transcribed by diacritics added to the main vowel, e.g., *mā* (mother), *má* (numb), *mǎ* (horse), *mà* (scold), and *ma* (particle) in the order of Tone 1 to Tone 4, respectively, and the neutral tone.

Previous developmental studies on articulation and phonology of Chinese-speaking children mostly focused on the acquisition of speech sounds and tones separately [9, 10, 11, 12, 13, 14, 15]. Some have also reported error patterns related to syllable structure, assimilation, and substitution processes. Earlier research results on Taiwanese Mandarin-speaking children have pointed out that syllable-initial consonants /p, t, k, k^h, m, n, l, x, tɛ/ and all lexical tones are acquired at the age of three [13], [14]. Tone 1 and Tone 4 are acquired earlier than Tone 2 and Tone 3 based on evidence obtained from the results of perceptual judgment and acoustic analysis [13, 16]. Moreover, tones produced in monosyllables and in continuous speech are likely to have different contour patterns due to contextual tonal coarticulation [14, 17].

In verbal communication, tone production is essential in the process of meaning decoding, so the assessment of phonological development should also account for connections between tones and syllables. However, this type of assessment has not yet been explicitly conducted. In this study, we will show that acoustics-based tone model predictions may actually be able to assist the practice of tone evaluation together with the traditional assessment practice of phonological transcription and judgment-based assessment.

2. A hybrid lexical tone evaluation workflow

The phonetic form of a tonal syllable is mapped to the designated word meaning by simultaneously processing the syllable and the produced tone contour. While the syllable and tone contour may differ in many ways, they are connected to

10.21437/Interspeech.2023-161

each other. Lexical tones are perceived via pitch information referring to the contour of the fundamental frequency (F0), and phone sequences may be processed at different linguistic levels, e.g., the syllable-initial consonant, the vocalic part, the entire syllable or the entire word.

We propose a hybrid technology-assisted workflow by making use of phonologically transcribed phonemes, assessed tones, and model-predicted tone categories, as shown in Figure 1. It is our intention to show that the proposed workflow is a starting point for developing a technology-assisted diagnostic screening tool that helps detect young children's speech problems at an early stage, on the one hand. On the other hand, the output of the tone model will deepen our understanding about the nature of lexical tones in speech acquisition research.

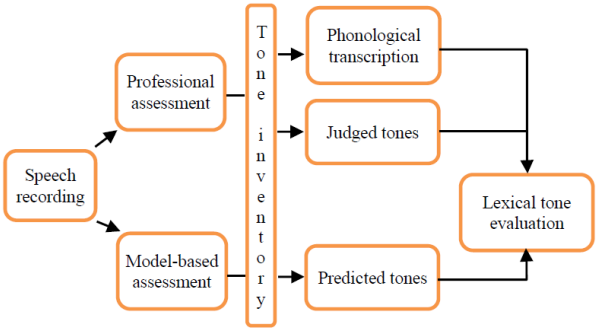

Figure 1: *Workflow of lexical tone evaluation.*

There are prerequisites for this workflow. Professional judgment and tone model prediction are based on the same tone inventory. The tone inventory is regarded as an innate system shared by human and model processing modules. In terms of judgment and prediction results, the notion of a "correct tone" is defined differently. To judge the accuracy of lexical tones, we refer to the lexical access to word meaning. If tone production clearly inhibits the comprehension of the word meaning, it is considered "incorrect"; otherwise, it is referred to as "properly produced". We take tone category into account in an implicit manner in the perceptual experiment.

A tone model is trained and tested using only the acoustic information of F0, without any other spectral information from the signal. For our tone model, if the predicted tone corresponds to the underlying tone category, it is labeled "correct"; otherwise, it is labeled "incorrect". Tone category is explicitly considered in tone model construction. To assess the relationship between segment and tone production, statistics of judged tones and predicted tones are analyzed with segment/syllable productions derived from our phonological transcription results.

3. Speech recording, data annotation and tone model

3.1. Data annotation

A total of 123 three-year-old children from Taipei City and New Taipei City in Taiwan participated in a speech recording project that was approved by the Institutional Review Board on Humanities and Social Science Research at Academia Sinica AS-IRB-HS07-107079. None of the children had known diagnosed diseases related to language, hearing, or cognitive development. All of the children passed a pure-tone audiometric hearing test with a GSI 18 Screening Audiometer at 1, 2, and 4 kHz at 20 dB on both ears. CapiAssess is an online system that facilitates speech recording with illustrative pictures. It also facilitates phonological transcription as well as automatically analyzing phonological development patterns by comparing standard pronunciation with the transcribed results. For speech collection, CapiAssess was installed on a MacBook Air Pro Retina 13.3 laptop with a Sony ECM MS907 microphone [18]. A picture-naming task was conducted to record the *Sinica Child Balanced Wordlist* [19], which consists of 70 children-friendly multisyllabic words. The words are scattered across different semantic fields that are familiar to children, including animals, food, transportation, body parts, movement, objects, games, locations, and natural phenomena. For each child, 148 syllables were recorded. Speech data were digitized at a sampling rate of 16 kHz and automatically processed by the *ILAS phone aligner* with manual verification of syllable boundaries.

Phoneme transcription was performed on 18,204 syllables by a phonetician assisted with spectrogram visualization in PRAAT [20]. As a result, a total of 2,803 syllables were transcribed with at least one segment pronounced incorrectly. When a syllable is evaluated with any improperly pronounced segment, it is deemed incorrect in the calculation of percentages of correct syllables (PCS). In the present study, three-year-old children acquire syllable-initial consonants /p, pʰ, t, tʰ, tɕ, m, n/, glides /j, w/, and vowels /e, ei, i, a, ə, u, ai, y/, based on a passing threshold of 90%. Neither of the nasal coda passed the threshold. These results are slightly different from previous developmental studies [10, 11, 12, 14]. For later analysis, we also calculated the percentages of correct vowels (PCV) and consonants (PCC) for each child according to the results of phonological transcription

For tone judgment, two annotators evaluated the tones in parallel. The agreement rate was 97% with a Cohen's kappa of 0.42. The low kappa value is due to the extremely imbalanced numbers of correct and incorrect cases. Inconsistent cases in the annotation project were then discussed among the annotators and the first author until consensus was reached. As a result, 190 of the 18,204 tokens of tones were judged "incorrect". We calculated the percentages of correct tones (PCT-judged) for each child. Similar to the results reported in previous research on tone acquisition, all but four children were able to produce 90% of the tones correctly. But only thirteen out of the 123 children were able to achieve the threshold for vowels and consonants.

3.2. ContourNet for tone prediction

For the implementation of our tone model, we used the data from the twelve subjects with the lowest PCV as our testing data. The remaining 90% of the data were used for model training (80%) and validation (10%). For our later analysis, the percentages of correct tones predicted by our tone model, noted as PCT-predicted, were accordingly calculated. For our tone model, we propose the ContourNet model, which is similar to BertSum [21], to abstract the contours of lexical tones. As shown in Figure 2, the model takes a sequence of multitonal f_0 values as input and extends BERT [22] by inserting multiple [CLS] symbols to learn (or summarize) *individual* tonal

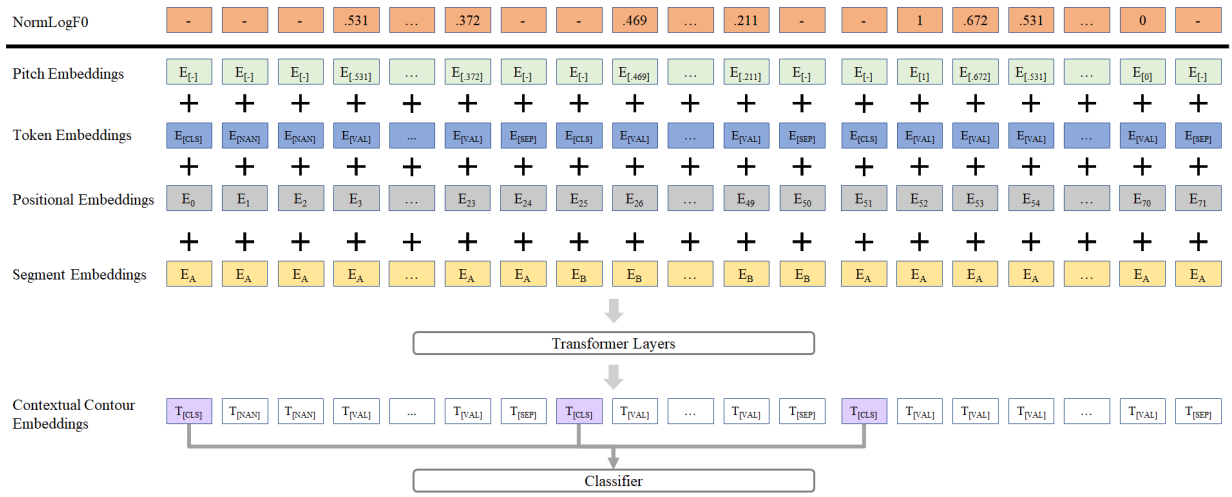

Figure 2: *Transformer-based* CONTOURNET *for tone prediction.* [CLS], [SEP], [VAL] *and* [NAN], *are defined following the architecture of* BERT.

contours and using separator tokens [SEP] to indicate tone boundaries in the Token Embeddings. The input f_0 values were first log-transformed and then normalized to [0, 1] using the speaker-specific ceiling and floor f_0 values determined at 0.1% and 99.9% of the range, respectively, abbreviated as NormLogF0.

The first embedding is the Pitch Embeddings that directly linearly transforms 1-d f_0 values to the hidden states of the dimensionality as $d_{\text{model}} = 256$, where the undefined pitch values [-] and the padded values at the end of every utterance are zeros. Second, for the Token Embeddings, except the embeddings for [CLS] and [SEP], all the estimated, nonzero pitch tokens and all the remaining zeros are specifically indicated as embeddings of [VAL] and [NAN]. Next, as was chosen in [23], we used the sinusoid positional embeddings as the Positional Embeddings to mark the pitch ordering. Finally, we used *interval segment* embeddings (the Segment Embeddings) similar to that in BERTSUM to distinguish the odd tones from the even ones in an utterance with two symbols E_A and E_B. These four embeddings are summed to a single input pitch vector x_i and fed to a bidirectional Transformer with multistacked layers:

$$\tilde{h}^l = \text{LN}(h^{l-1} + \text{MHAtt}(h^{l-1})) \quad (1)$$

$$h^l = \text{LN}(\tilde{h}^l + \text{FFN}(\tilde{h}^l)) \quad (2)$$

where $h^0 = x$ are the input pitch vectors; LN stands for the layer normalization; MHAtt is the multihead attention; the superscript l indicates the depth of the stacked layer (L=4); and FFN is the feed-forward network. Therefore, contextual contour representations are learned hierarchically, where lower Transformer layers focus on adjacent pitches and higher layers in combination with self-attention, which focuses on the multitonal context of tonal coarticulation effects. Finally, additional 2 inter-tone Transformer layers are stacked on top of current BERT outputs to capture utterance-level influences for reshaping contour abstractions. The contextual T vectors are later fed to a fully connected projection layer (with *tanh*) for shuffling the features for tone prediction

We trained the CONTOURNET model on a 52-hour corpus of 255 adults' read speech that was developed at Feng Chia University with the original tone model framework [24]. Then, the experiment was carried out by adapting the contour representations to children's speech in the softmax layer (SL) or from the very beginning stage of the model (FULL). When training from scratch (Scratch), the contour representations are only learned from our small-scale children's speech dataset. As shown in Table 1, we find that training from scratch is robust in terms of abstracting the contour of the four lexical tones: Tone 1 (T1), Tone 2 (T2), Tone 3 (T3), and Tone 4 (T4). The model achieves the best performance despite the lack of pretrained representations using this training strategy. Nearly two-thirds of the neutral tones are retrieved. In the next section, we analyze the data in the test set from the twelve children described above using the predicted tone outputs from the Scratch training strategy.

Table 1: CONTOURNET *model performance.*

	Accuracy (%)	F1-score (%)				
		T1	T2	T3	T4	Neutral
Fine-tuning on SL	62.5	67.9	55.5	50.2	72.4	0
Fine-tuning on FULL	80.6	80.4	80.4	76.9	85.4	58.3
From Scratch	**81**	81.8	82.1	75.8	84.4	61.5

Table 2: *Accuracy rates of tones and segments.*

	Accuracy (%) of				
	judged tone vs			predicted tone vs	
	predicted tone	consonant	vowel	consonant	vowel
All	79.6	70.4	74.9	62.7	65.4
T1	83.6	61.5	77.1	56.2	67.5
T2	78.4	77.4	75	69.4	69.2
T3	71.9	69.7	71.9	57.8	56.1
T4	84.8	72.1	74.1	66.9	66.7

4. Results and discussion

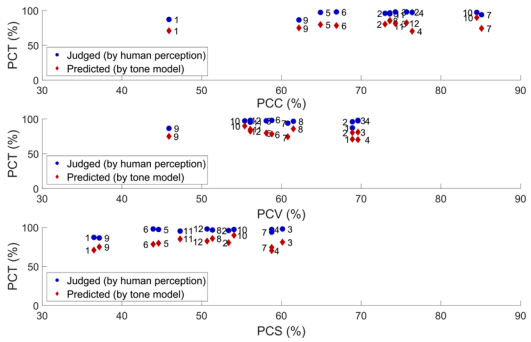

Figure 3: *Comparison of judged and predicted tones in relation to* PCC *(top),* PCV *(middle) and* PCS *(bottom).*

We examine the accuracy rates of judged tones and transcribed consonants and vowels from CONTOURNET model predictions. The results are summarized in Table 2. First, the prediction of our tone model reflects the assessment using human judgment. The overall rate is 79.6%. The accuracy rates descend in the order of Tone 4 > Tone 1 > Tone 2 > Tone 3. Second, the accuracy rates of both judged and predicted tones are more closely correlated with vowel production than consonant production. With respect to individual tones, the largest gap between judged and predicted tones as well as between tone and segment production is observed for Tone 3 data. It seems that the tone judgment results are more related to segment production (both vowels and consonants) than tone model prediction. It may be due to the fact that our tone model only makes use of F0, but for human judgment, spectral features that represent segment information are also available to the annotators. If an additional phoneme model were built to supplement our tone model, performance would likely rise.

As the goal of a clinical screening tool is to evaluate the speech performance of each child, we plotted the PCT-judged and PCT-predicted results of data from the test set, as shown in Figure 3. When comparing PCC, PCV, and PCS, the differences between the PCT-judged and PCT-predicted results are relatively stable, with a gap that does not exceed 27% of the human judgment-based results. As the model prediction is consistently similar to that of human judgment, we further observed the patterns of PCT, PCC, and PCV for individual tones. Figure 4 shows that the PCT-judged and PCT-predicted results show consistent differences in Tone 1 and Tone 4 but display diverse scattering patterns in Tone 2 and Tone 3. The variance of the PCT results among these twelve subjects in relation to the PCV results is more convergent in relation to the PCC results. This pattern is most clearly exhibited in Tone 4.

Previous developmental studies have confirmed that Tone 1 and Tone 4 are acquired earlier than Tone 2 and Tone 3 [13, 14]. Our data provide new support for this result. The acoustic pattern processed by our tone model is the closest to the perceived tone contours for Tone 1 and Tone 4, as the accuracy rates between judged tones and predicted tones are the highest. This suggests that the divergence between the acoustic and auditory inputs may be the lowest in high level Tone 1 and high-falling Tone 4. The rising Tone 2 and the multivariant Tone 3 cause more divergent patterns for the tone model and tone judgment. This may also indicate that the high-register levels of Tone 1 and Tone 4 may be more easily perceived and less misleading than the low-register levels of Tone 2 and Tone 3.

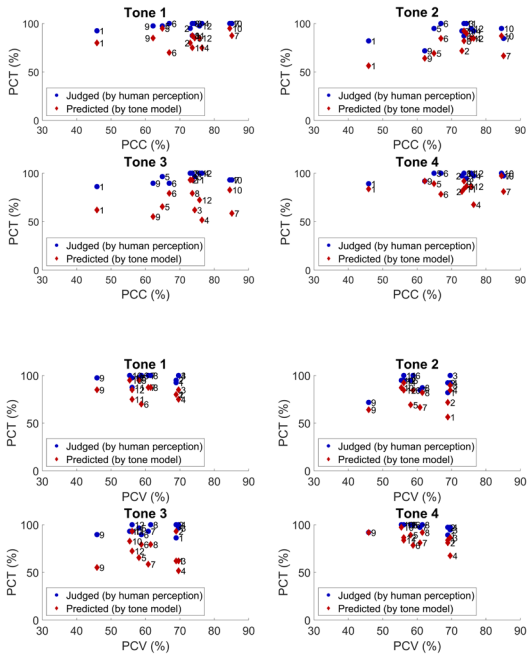

Figure 4: PCT *of the four lexical tones in relation to* PCC *(top) and* PCV *(bottom).*

5. Conclusions

The accuracy rate between the judged and predicted tones is approximately 80%. With more training data, we believe that it is possible to fine tune the model and apply it as a screening tool. Most developmental studies focus on consonants. But in our study tone accuracy is related more to vowels than consonants, suggesting that Mandarin-speaking children may have different developmental processes for consonants, vowels, and tones. The last two linguistic units may be more closely correlated. The accuracy rates between judged and predicted tones and between tones and segments drop more for Tone 3 than for the other tones. Tone 3 has three surface forms in speech production, including canonical Tone 3, Sandhi Tone 3, and half Tone 3. Thus, dealing with acoustic modeling of Tone 3 variations is a challenging task for a screening tool. Our present tone model only uses F0 information, no other spectral information. However, there are clear correlations between tone and segment accuracy. The findings reported in our study suggest that tone model-assisted evaluation may actually serve as an efficient and low-cost means of screening children's speech.

6. Acknowledgements

We would like to express our sincere gratitude to all our team members, the children, parents and kindergarten caregivers who took part in the recording project. This work was supported by the Institute of Linguistics, Academia Sinica and the National Science and Technology Council, Taiwan, with projects [109-2410-H-001-087-MY2, 111-2410-H-001-015-MY3] granted to the first author and [110-2222-E-035-005-MY2] granted to the second author.

7. References

[1] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell, and J. H. L. Hansen, "Fusing text-dependent word-level i-vector models to screen 'at risk' child speech," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 1681-1685.

[2] R. Lileikyte, D. Irvin, and J. H. L. Hansen, "Assessing child communication engagement and statistical speech patterns for American English via speech recognition in naturalistic active learning spaces," *Speech Communication*, vol. 140, pp.98-108, 2022.

[3] Y.-S. Lin, and S.-C. Tseng, "Classifying speech intelligibility levels of children in two continuous speech styles," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7748-7752.

[4] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning," *International Journal of Medical Informatics*, vol. 90, pp.13-21, 2016.

[5] D. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil and A. Morgan, "Improving child speech disorder assessment by incorporating out-of-domain adult speech," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2690-2694.

[6] P. Vizza, G. Tradigo, D. Mirarchi, R. B. Bossio, N. Lombardo, G. Arabia, A. Quattrone, and P. Veltri, "Methodologies of speech analysis for neurodegenerative diseases evaluation," *International Journal of Medical Informatics*, vol. 122, pp.45-54, 2019.

[7] S. Duanmu, *The phonology of standard Chinese*. Oxford University Press, 2007.

[8] Y.-H. Lin, *The Sounds of Chinese with Audio CD*. Cambridge University Press, 2007.

[9] H. Cheung, and B.-H. Hsu, "Chinese children's production and perception of consonants: A developmental study," *Journal of the Taiwan Speech Language Hearing Association*, vol. 15, pp.1-10, 2000.

[10] S.-C. Cho, *The Phonological Development of 3 to 6 year-old preschool children in Taiwan*. National Taipei University of Nursing and Health Sciences, Master Thesis, 2008.

[11] Z. Hua, and B. Dodd, "The phonological acquisition of Putonghua (modern standard Chinese)," *Journal of Child Language*, vol. 27, pp.3-42, 2000.

[12] J.-Y. Jeng, "The speech acquisition of Mandarin-speaking preschool children," *Journal of Child Language*, vol. 14, pp.109-136, 2017.

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

[14] G. B.-G. Lin, and M.-H. Lin, "A Study on the Development of Language Abilities in Preschool Children in Taiwan, R.O.C.," *Bulletin of Special Education*, vol. 10, pp.259-281, 1994.

[15] L.-L. Yeh, B. Wells, J. Stackhouse, and M. Szczerbinski, "The development of phonological representations in Mandarin-speaking children: Evidence from a longitudinal study of phonological awareness," *Clinical Linguistics & Phonetics*, vol. 29, pp.266-275, 2015.

[16] S.-C. Tseng, and Y.-F. Liu, "Segment and Tone Production in Continuous Speech of Hearing and Hearing-impaired Children," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 606-610.

[17] 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*, vol. 48, pp. 1065-1079, 2005.

[18] S.-C. Tseng, and Y.-F. Liu, "Establishment of normative phonological development data for Mandarin-speaking children by applying computational linguistics techniques," in *the Annual meeting of the Speech-Language-Hearing Association of Taiwan*, 2017.

[19] S.-C. Tseng, "ILAS Chinese spoken language resources," in *the third International Symposium on Linguistic Patterns in Spontaneous Speech*, 2019, pp. 13-20.

[20] P. Boersma, and D. Weenink, Praat: doing phonetics by computer. Version 6.2.14. Retrieved 24 May 2022 from http://www.praat.org/.

[21] Y. Liu, and M. Lapata, "Text Summarization with Pretrained Encoders," in *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3730-3740.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171-4186.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Conference on Neural Information Processing Systems (NIPS)*, 2017, pp.1-11.

[24] Y.-F. Liu, "Representation Learning for Discovering Typical Pitch Contours of Spoken Words and Core Communication Speech Rhythm in Utterances," in *Project report: National Science and Technology Council*, 2022, Taiwan.