INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

J-ToneNet: A Transformer-based Encoding Network for Improving Tone Classification in Continuous Speech via F0 Sequences

Yi-Fen Liu and Xiang-Li Lu

Department of Information Engineering and Computer Science, Feng-Chia University

yfliu@fcu.edu.tw, M1105831@o365.fcu.edu.tw

Abstract

Currently, tone classification studies mainly focus on training classifiers by using intrinsic features of isolated segments, i.e. often the syllables. Mostly, the works are not merely in use of fundamental frequency (f_0) but utilizing more information on the spectrograms, MFCCs, or energy to improve model accuracy. However, as we know, more challenges on tone classification lie on modeling the complex f_0 variations from the tonal coarticulations and the interactive effects among tonality in continuous speech. To tackle down this issue, we first aim at in using the sequence of f_0 samples in speech utterance only. In addition, we propose a transformer based network with an extendable BERT input architecture and a joint learning technique to consolidate the contour representations of consecutive tones. Leveraging or fusing more information affected from speech rhythm in utterance, the experiments show that the proposed J-ToneNet is very robust for read speech.

Index Terms: pitch contour, tonal coarticulation, speech rhythm, jointly learning, encoder, BERT, Transformer layers

1. Introduction

Mandarin Chinese has phonemic tones on full syllables, whereby four tones exploited via the changes in the fundamental frequency (f_0) contour can give lexical contrasts. Conventionally, they are transcribed by diacritics added to the main vowel, e.g., *mā* (mother), *má* (numb), *mǎ* (horse) and *mà* (scold). Commonly, the four lexical tones are transliterated into numerals in which the tones are perceptually distinguished by pitch register and direction, such as high tone /55/, rising tone /25/, dipping tone /214/, and falling /51/ in the order of Tone 1 to Tone 4, respectively. In Figure 1, we introduce the four lexical tones and the two tonal variants of Chinese monosyllabic words ending with a phonemic segment /i/. The dotted blue lines are the speaker-specified log-transformed pitch curves and the fitted 2nd-order polynomial lines in the voiced part and full syllable region are visualized as the dotted cyan and dashed red lines. The dotted cyan curves are nearly equivalent to the numeral encoding in perception. As noted in [1, 2], another often realized tonal variant for Tone 3 is a low falling /21/; whereas, for Tone 2, another dipping variant in /323/ would be perceived in spoken Taiwan Mandarin. The variation of tonal contours in continuous speech would worsen model's ability to discriminate one tone from the others, particularly when a model is built in absence of context [3, 4].

Tone classification is important not only for speech evaluation of Mandarin Chinese in computer-aided (CALL) systems of second language (L2) learners [5, 6, 7, 8] but for enhancing the recognition accuracy in most state-of-the-art Mandarin automatic speech recognition (ASR) systems [9, 10, 11, 12]. Often in the two-stage-based tone model, the back-end classifiers trained on a full segment of syllable, or merely on a final, achieve high accuracy with the benefit of deep neural networks (DNNs). Some convolutional neural network (CNN)-based models using spectral features only showed satisfactory results on broadcast news speech [13, 14]. Some other Tone Networks are constructed more complicatedly to account for more context information no matter in using which kind of combinations of the feature frontends on Mel-spectrograms, FFT spectrograms, MFCC, energy, f_0, … etc. For instance, the works in [3, 4, 15], modeling in hybrid with the recurrent neural layers, attention mechanism and the joint training strategy indeed reduce the error rate significantly. However, we notice that there are relatively few documented deep learning models based on transformers or the BERT-based framework to encode the pitch curve information from the f_0 value sequences for tone classification.

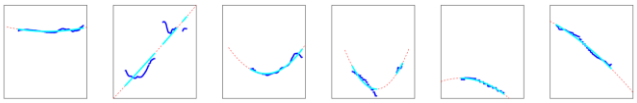

Figure 1: *Six Taiwan Mandarin tonal variants of citation contours in normLogF0 on Tone 1 (di /55/, high level for 'low'), Tone 2 (qi /25/, rising for 'that' vs. ti /323/, dipping for 'to mention'), Tone 3 (di /213/, dipping vs. di /21/, low falling for 'end') and Tone 4 (di /51/, falling for 'place').*

In this paper, we advocate a new, unified deep learning paradigm to enhance tone classification, not only on the isolated syllables, short words but extendable for consecutive syllables in larger speech chunks. The tonal variants in larger speech chunks are diverse and mostly are affected by the contextual contour or rhythmic pattern of different speech rate conditions [16, 17]. Thus, those two factors should be considered and targeted for consolidating tonal representations. When designing the network to learn representations for sequential f_0 contour on lexical tones, we explore an utterance-level encoder based on bidirectional transformers [18, 19], referred as C-Net. Similar to the extendable use of BERT for text summarization in [20], C-Net extends BERT by inserting multiple [CLS] symbols to learn contour representations of Chinese syllables. The exploited technique is to use interval segmentation embeddings for distinguishing multiple syllabic tones in utterance as BERTSUM adopted for learning sentence representation in discourse. Later, the contour representation of a Chinese syllable is fused with corresponding rhythmic representation learned by another encoding network (R-Net) for the task of tone classification. The main research contributions of this work are summarized as follow:

- To the best of our knowledge, we are the first to explore only the sequential pitch information and speech rhythm in utterance for consolidating tonal contours of Chinese syllables in deep neural network.

10.21437/Interspeech.2023-695

- We proposed a fully-transformer-based model, J-ToneNet with a joint learning task on word tone prediction and it is proven effective to exploit more information from utterance context in terms of word inventory.
- We conduct experiments on two datasets of different speaking styles. The results show that our model produces significant improvements on continuous speech.

2. Speech recordings and datasets

Two speech datasets in different speaking styles were used in the conducted experiments. One is the prepared read speech and the other is the spontaneous conversational speech. All the speech materials were recorded in quiet rooms, sampled at 16 kHz and processed by the *ILAS phone aligner* [21] with different degree of manual postediting or human verification.

2.1. Corpus

2.1.1. FCU-VOICE-100

A total of 400 students of Feng Chia University were recruited. Every 40 speakers (20 males, 20 females) received the same set of 250 reading prompts to read. 10 distinct sets of reading prompts are designed and chosen from the *Sinica Core Vocabulary Inventory* [22], half of which are frequently used words and half of which are sentences indicating the word use. Each speaker was asked to read with clarity and naturalness like he/she originally intended to say so as the speaker-specific variations of the speech rhythm and tone changes in coarticulation could be elicited for tone modelling. In the conducted experiments, we use a subset of speech materials recorded by 100 speakers.

2.1.2. MCDC-8

Eight 1-hour spontaneous conversations with talked topics freely decided by the 16 paired speakers were used in this work. This is a released spoken Chinese resource from the Mandarin Conversation Dialog Corpus (MCDC) [23], where the speech materials were truncated in 6,060 speaking turns.

2.2. Datasets, annotation and distribution in length

In use of the *ILAS phone aligner* and orthographic transcriptions with punctuations, e.g., comma, semicolon in a reading prompt, we obtained the force-aligned clausal chunks, words and syllables for every speech recording in FCU-VOICE-100 dataset. Only the "correctly produced and properly aligned" ones are included for the conducted tone experiments. The notion of a "correctly produced and properly aligned" speech recordings is defined as every force-aligned sound on word clearly inhibits the comprehension of its lexical meaning and tone production. Coming upon any hesitation on the perceptual judgement, the annotator discussed with the first author until consensus was reached. After manual verifying, 4,641 out of 25,020 speech recordings were excluded. The released MCDC-8 was thoroughly verified on the annotation of inter-pausing units (IPUs), words, syllables and the tone transcription as it was described in [24].

The adopted processing units for read speech and conversational speech are the utterances on clausal chunks and IPUs. In Table 1, we summarize the characteristics of these two datasets in terms of the number of utterances, syllables, and tone verification. The coverage and distribution of utterances in difference syllable length are shown in Figure 2. Half of utterances, i.e., nearly 11,000 (46.3%) in FCU-VOICE-100 are mono-, di-syllabic words; whereas in MCDC-8, the number is only 1,890 (14.1%). In conversational speech, a coverage of 10% of utterances are long ones over the length of 17 syllables.

Table 1: *Comparison of the two datasets.*

Corpus	#Spkrs	#Utt	#Syls	Tone Verifying
FCU-VOICE-100	100	23,601	125,178	Yes
MCDC-8	16	13,407	131,003	Yes

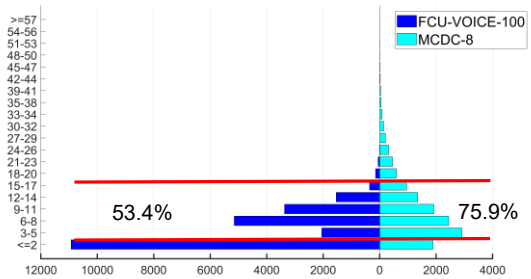

Figure 2: *Coverage and Distribution of utterances in different syllable length.*

3. Classifier on isolated segments

The work here on tone classification mainly explores the contour representations of four lexical tones by using pitch information only. The f_0 values estimated from PRAAT pitch tracking [25] were log-transformed first 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, here abbreviated as *normLogF0*.

3.1. Estimated features

For segment based models stated below, we use a fixed length of 20 points resulting from interpolation of sequential f_0 observations. This *normLogF0*(20) is the first set of estimated features for a Chinese syllable. The survey on tonal features for a Chinese syllable listed in [26] is extended and provided in Table 3. We use this *ToneFea*(17) as a second expanding set.

Table 3: *Estimated features for Tones.*

Fea.	Descriptions
1-3	Coefficients of second order polynomial function fitted on the estimated pitch contour (*normLogF0*).
4-5	Relative positions of minima and maxima of *normLogF0.*
6-11	$(a_{max} - b_{min})$, $(c_{max} - b_{min})$, $(c_{max} - a_{min})$, $(a_{max} - c_{max})$, $(a_{max} - c_{min})$, $(a_{min} - c_{min})$, where a and c is respectively the first and the fourth quartile; b is the union of the second and third quartiles; and the *min*-subscript (*max*-subscript) denotes the minimum (maximum) value in that quartile.
12-17	Corresponding slopes on previously defined regions.

3.2. Segment based models

3.2.1. Random forest

The random forest [27] with 16,384 tree predictors is the baseline model we use for comparison with the state-of-the-art

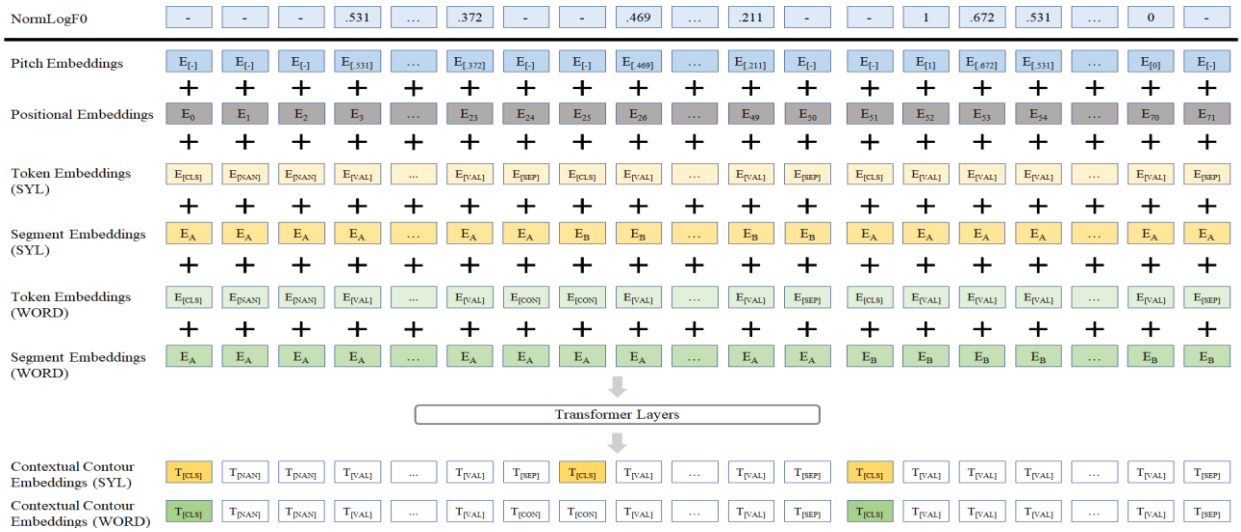

Figure 3: *The proposed Transformer based encoding network for tonal contour* (C-Net).

neural networks or the more complicated deep neural networks. We choose random forest as it is a meta estimator that fits a number of weak learners on various subsamples of the dataset and is often robust with respect to noise, such as the abnormal f_0 estimation incurred from current pitch tracking algorithms.

3.2.2. *Feed-Forward Network (FNN)*

A simple network for segment-level (i.e., on aligned isolated syllables) tone classification is built with two fully connected layers with Gaussian error linear units (GeLUs) [28], which were also adopted in BERT's model for sequence-to-sequence tasks. Later, the last hidden layer is used to generate an output layer of 512 neurons, which is passed through a softmax layer to generate a probability distribution over the four lexical tones.

3.2.3. *1D-CNN*

The core learning networks of the proposed 1D-CNN are the convolution, pooling and the fully connected layers, where the activation functions are also the GeLUs. Six convolutional layers with a convolutional kernel (filter) size of 4 and a stride length of 1 are adopted, and the numbers of filters for every 2 layers are 128, 256 and 512. Only the second and fourth convolutional layers are rightly followed by an average pooling layer (kernel size 2). The output of the sixth convolution is then passed through two fully connected layers with 512 units, and the softmax layer yields 4 outputs, which correspond to 4 bearing lexical tones.

4. Encoders and the tone network

4.1. C-Net: tonal contour abstraction

BERTSUM [20] is a BERT architecture for text summarization. It takes a sequence of multi-sentential inputs and extends BERT by inserting multiple [CLS] symbols to learn sentence representations. We propose an encoding network, which is similar to BERTSUM, to abstract the contours of lexical tones from sequential f_0 values. As shown in Figure 3, the proposed C-Net takes a sequence of multitonal f_0 values as input and is in use of BERT input architecture to define the functions with learnable parameters. These functions are the building blocks from which transformers are made [29]. Full architectures featuring these building blocks in C-Net are stated below.

The first block is the Pitch Embedding which directly linearly transform 1-*d* f_0 values to the hidden states of the dimensionality as $d = 512$ where the untracked pitch values of unvoiced parts, and the padded values at the end of every utterance are zeros. Secondly, as it was chosen in [18], we use the sinusoid positional embedding as Positional Embedding to mark the pitch ordering. Next, in the Token Embedding, we use [VAL], and [NAN] respectively to indicate the estimated and the remaining untracked or padding pitch tokens. The inserted [CLS] token in front of every syllable or word token is to aggregate contour information from the input f_0 values until a syllable or word boundary separator token ([SEP]) is met. Lastly, similar to the BERTSUM, we also use *interval segment* block (the Segment Embedding) both for syllable and word to distinguish the odd tones from the even ones in an utterance with two symbols E_A and E_B. These four embeddings at *i*-th pitch value are summed to a single input pitch vector x_i and fed to a bidirectional Transformer with multi-stacked layers:

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

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

where $h^0 = x$ are the concatenated pitch vectors of matrix $\in \mathbb{R}^{n \times d}$; LN stands for the layer normalization; MHAtt is the multi-head attention; and the superscript *l* indicates the depth of the stacked layer (L=4). This way, contextual contour representations are learned hierarchically where lower Transformer layers focus on adjacent pitches; while higher layers, in combination with self-attention, focus more on coarticulation effects of tones.

Of the C-Net encoding network, the final hidden vector t_i in T is the vector of the *i*-th [CLS] symbol from the top layer h^L. This embedding of tonal contour for *i*-th syllable is t_i^c, and would be fused with its corresponding rhythmic embedding t_i^r.

4.2. R-Net: auxiliary information from speech rhythm

We design a similar encoding network to learn the auxiliary information from speech rhythm. This proposed R-Net takes two sequences as inputs and both are relevant to duration. One is the original durations of syllables in utterance; the other one is the differences of syllable durations and the mean duration. The first used building block in R-Net is the Duration

Table 3: *Tone accuracy (%) of models.*

Models	FCU-VOICE-100		MCDC-8	
	F0	*+ToneFea*	*F0*	*+ToneFea*
Random Forest	52.5	57.2	44.8	48.9
FFN	50.5	55.9	43.6	44.9
1D-CNN	52.6	56.7	45.4	45.4
J-ToneNet	**91.0**		**61.7**	

Component removing	FCU-VOICE-100				MCDC-8			
	Overall	Utt. Length			Overall	Utt. Length		
		<=2	3-20	>=21		<=2	3-20	>=21
J-ToneNet	**91.0**	80.0	93.0	94.1	**61.7**	67.1	61.9	60.7
– joint learning	87.7	78.2	89.4	91.2	58.6	63.4	59.7	55.8
– joint learn. & R-Net	86.0	78.9	87.4	73.5	58.9	63.8	60.2	55.6
only C-Net on SYL	59.0	-			50.6	-		

Embedding which linearly transform 2-*dimensional* durational inputs to the hidden states of the same dimensionality as C-net, i.e., d = 512, by using a fully connected (FC) layer. Another building block, the Positional embedding is then added to the hidden state before feeding into a two-layered Transformer. The architecture of the Transformer is the same as the one depicted in (1) and (2). Here, we are in use of t_i^r to stand for the rhythm embedding of *i*-th syllable in utterance.

4.3. Joint tone classification network (J-ToneNet)

In fully constructed Joint tone classification network (J-ToneNet), we fuse the two embeddgins, t_i^c and t_i^r as it is stated in (3). These two embeddings are first projected into the same hyper space and then concatenated before layer normalization being applied.

$$t_i = \mathrm{LN}((W_c t_i^c) \oplus (W_r t_i^r)) \quad (3)$$

where, $t_i \in \mathbb{R}^{2d}, W_c \in \mathbb{R}^{d\times d}, W_r \in \mathbb{R}^{d\times d}$.

Next, in order to exploit channel dependencies in output feature t_i as the works in [30], we parameterize the gating mechanism by forming a reverse bottleneck [31, 32] with two FC layers around the non-linearity, i.e. a dimensionality-expansion layer with parameters $W_1 \in \mathbb{R}^{2d\times 4d}$, a GeLU (i.e. δ) and a dimensionality-reducing layer with parameters $W_2 \in \mathbb{R}^{4d\times 2d}$.

$$s = \sigma(W_2(\delta(W_1 t_i))) \quad (4)$$

In (5), the attention weight s is then element-wise multiplied to the hidden state t_i. Rightly, the output is passed through a linear layer with parameters $W_3 \in \mathbb{R}^{2d\times d}$ to transform the dimensionality back to the original one, d = 512. Before feeding it into the classifier as the equation (6) states, an additional non-linear (*tanh*) transformation is applied to make output representation $\tilde{t}_i$ more interpretable.

$$\tilde{t}_i = \tanh(W_3(t_i \otimes s)) \quad (5)$$

$$\hat{y}_i = softmax(W_o \tilde{t}_i + b_o) \quad (6)$$

A forming of 'compatible context' or a 'conflicting context' as it is described in [16] says that the neighboring tone preceded or followed has f_0 value similar to, or different from the register/offset of target tone, like the tone pairs of /55/-/55/ or /51/-/55/. With this notion, the loss of the J-ToneNet in (7) defining on prediction tone $\hat{y}_i$ against gold tone label y_i considers the cross-entropy both on tones of syllables and tones of monosyllabic words and certain disyllabic word tone pairs. Much alike in equations of (5) and (6), a word tone classifier is built in use of the *T* vectors indicated in last block of Figure 3.

$$loss = CE(\hat{y}_{syl}, y_{syl}) + CE(\hat{y}_{word}, y_{word}) \quad (7)$$

5. Results and concluding remarks

In the conducted experiments, each dataset was split into training, development and test sets in 80%, 10% and 10% of the utterances, respectively. The top three segment based models on left part of Table 3 reports the tone classification results on the isolated syllables where no contextual information out of the fully aligned syllable range is used. For read speech, appending further with the *ToneFea*(17) instead of simply using the *normLogF0*(20) always achieves the best result. However, when we test spontaneous speech, slightly improvement or nearly no gain is obtained in neural models with the appended set of *ToneFea*(17); on the contrary, it is increased with a rate of 4.1% in random forest. The overall performance and the tone accuracies of consecutive syllables in larger speech chunks achieve the best in the proposed J-ToneNet.

Next, each at a time only one component of the J-ToneNet is removed in a reverse constructing sequence. The model performance is clearly degraded while making the model not jointly train to decide if both monosyllabic words and certain word tone pairs are predicted correctly. Further, taking off the R-Net, the tone accuracy on long speech chunks over 21 syllables is worse, especially for read speech. The last line in the right part of Table 3 shows that when using only the contour encoding network (i.e., C-Net) on isolated syllables, such encoding technique seems effective, but largely falls behind of the achievement seen in utterance-level context. Generally, the results suggest that extending transformers over pitch sequences in utterance works quite well and provide evidence on that utterance context is crucial for model enhancement.

Here are the listed concluding remarks: a) Adopting pitch information only and leverage it with the rhythmic sequence explicitly in a unified deep architecture has proved effective in classifying tones of consecutive syllables in utterance. b) Further, with a joint learning approach to incorporating more information on the use of word tones in spoken utterance context, it shows that the model discriminates tones more robustly in read speech. c) Currently, the model is still in development, and the results are quite preliminary. Enriching rhythmic representations of the timing patterns between syllables in association with the different degrees of vowel merging is our future goal to improve the unsatisfactory results for conversational speech. d) As the work in [33], this proposed model could be valuable for clinical applications of screening children's speech on tone production.

6. Acknowledgements

The work was financially supported by the National Science and Technology Council, Taiwan, with project [110-2222-E-035 -005 -MY2] granted to the first author.

7. References

[1] R. Sandres, "Tonetic Sound Change in Taiwan Mandarin: The Case of Tone 2 and Tone 3 Citation Contours," in *the 20th North American Conference on Chinese Linguistics (NACCL-20)*, 2008, pp. 87-107.

[2] K. Huang, "Phonological Identity of the Neutral-tone Syllables in Taiwan Mandarin: An Acoustic Study," *Acta Linguistica Asiatica*, vol. 8, pp. 9 -50, 2018.

[3] L. Yang, Y. Xie, and J. Zhang, "Improving Mandarin Tone Recognition using Convolutional Bidirectional Long Short-Term Memory with Attention," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 352-356.

[4] J. Tang, and M. Li, "End-to-End Mandarin Tone Classification with Short Term Context Information," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2021, pp. 878-883.

[5] R. Tong, N. F. Chen, B. Ma, and H. Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 801-805.

[6] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Using Tone-based Extended Recognition Network to Detect Non-native Mandarin Tone Mispronunciations," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1-4.

[7] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mandarin Tone Mispronunciation Detection for Non-native Learners with Soft-target Tone Labels and BLSTM-based Deep Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6249-6253.

[8] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mispronunciation Detection of Mandarin Tones for Non-native Learners with Soft-target Tone Labels and BLSTM-based Deep Tone Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 2012-2024, 2019.

[9] G. Peng, and William S.-Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines," *Speech Communication*, vol. 45, pp.49-62, 2005.

[10] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2006.

[11] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly Accurate Mandarin Tone Classification in the Absence of Pitch Information," in *International Conference on Speech Prosody*, 2014, pp. 673-677.

[12] N. Ryant, J. Yuan, and M. Liberman, "Mandarin Tone Classification without Pitch Tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4868-4872.

[13] C. Chen, R. Bunescu, L. Xu, and C. Liu, "Tone Classification in Mandarin Chinese Using Convolutional Neural Networks," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 2150-2154.

[14] Q. Gao, S. Sun, and Y. Yang, "ToneNet: A CNN Model of Tone Classification of Mandarin Chinese," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3367-3371.

[15] H. Huang, K. Wang, Y. Hu, and S. Li, "Encoder-decoder based pitch tracking and joint model training for Mandarin tone classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6943-6947.

[16] Ricky K. W. Chan, "Speaker discrimination: Citation tones vs. coarticulated tones," *Speech Communication*, vol. 117, pp.38-50, 2020.

[17] A. Lee, S. Prom-on, Y. Xu, "Pre-low raising in Cantonese and Thai: Effects of speech rate and vowel quantity," *The Journal of the Acoustical Society of America*, vol. 149, pp. 179-190, 2021.

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

[19] 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 Technology (NAACL-HLT)*, 2019, pp. 4171-4186.

[20] 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.

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

[22] S.-C. Tseng, "Lexical coverage in Taiwan Mandarin conversation," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 18, pp. 1-18, 2013.

[23] S.-C. Tseng, "Spoken Corpora and Analysis of Natural Speech," *Taiwan Journal of Linguistics*, vol. 6, pp. 1-26, 2008.

[24] Y.-F. Liu, S.-C. Tseng, and Roger J.-S. Jang, "Deriving disyllabic word variants from a Chinese conversational speech corpus," *The Journal of the Acoustical Society of America*, vol. 140, pp, 308-321, 2016.

[25] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*. vol. 5, pp. 341-345, 2001.

[26] S. Garg, G. Hamarneh, A. Jongman, J. Sereno, and Y. Wang, "Joint Gender-, Tone-, Vowel- Classification via Novel Hierarchical Classification for Annotation of Monosyllabic Mandarin Word Tokens," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5744-5748.

[27] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[28] D. Hendrycks, and K. Gimpel, "Gaussian error linear units (gelus)," in *arXiv preprint arXiv:1606.08415v4*, 2016.

[29] M. Phuong, and M. Hutter, "Formal Algorithms for Transformers," in *arXiv preprint arXiv:2207.09238v1*, 2022.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," in *arXiv preprint arXiv:1704.04861*, 2017.

[32] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 11966-11976.

[33] B. Lin, and L. Wang, "Attention-based Multi-encoder Automatic Pronunciation Assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7743-7747.