Interspeech 2024
1-5 September 2024, Kos, Greece

Deep Prosodic Features in Tandem with Perceptual Judgments of Word Reduction for Tone Recognition in Conversed Speech

Xiang-Li Lu[1], *Yi-Fen Liu*[2]

[1]Feng-Chia University Artificial Intelligence Research Center, Taichung
[2]Department of Information Engineering and Computer Science, Feng-Chia University, Taichung
coh4ry7z@gmail.com,yfliu@fcu.edu.tw

Abstract

To tackle the tone classification problem in conversational speech, we propose a transformer-based encoding network to classify tones in an utterance on a syllable-by-syllable basis. Using just F0 and rhythmic information, the interaction encoder consolidates contour representations first. By jointly predicting word tones using perceived judgments on reduction degrees, the learning architecture improves automatic recognition of the underlying syllable tones. Leveraging these enhancements, the experiments show that the proposed model is very robust and achieved a 12% increase in tone classification accuracy.

Index Terms: suprasegmentals, contour, tonal coarticulation, word reduction, rhythmic variation, Transformers

1. Introduction

Suprasegmental information, also known as prosody, consists of the rhythm, timing, meter, pitch and stresses of words and utterances that we speak. Mandarin uses pitches as the primary acoustic correlate of tones to signal changes in word meaning, e.g., mā (mother), má (numb), mǎ (horse) and mà (scold). Following the order of Tone 1 to Tone 4, tones are likewise transliterated into numerals to indicate pitch register and direction, for instance, high level tone /55/, rising tone /25/, dipping tone /214/, and falling tone /51/. The four lexical tones for Chinese monosyllabic words ending with phonemic segment /i/ are shown in Figure 1. Blue dotted lines show the speaker's log-transformed pitch curves, while cyan dotted lines and red dashed lines represent the fitted 2nd-order polynomial lines in the voiced part and full syllable region. Almost like encoded numerals, the dotted cyan curves resemble tonal contours.

Just as segments, tones in conversed speech are often coarticulated and varied in contours while words are reduced or spoken in fast speed [1]. It's production and perception of contours would be affected in some way. To judge a reduction degree for any disyllable, whether it is a word or just a concatenated syllabic sequence, follows the perceptual criteria established in [2]. With a clearly identifiable word-internal syllable boundary and none of the non-vocalic segments across the syllable boundary (i.e., the nasal coda N of the first syllable and onset C of the second syllable) are deleted, we classify the words as a category of Canonical Form (CAN). In the opposite extreme case, a Syllable Merger (SYM) occurs when two syllables are thoroughly merged into one eligible Chinese syllable. Merging Continuum between two Extremes (MCE) occurs when a syllable boundary becomes blurred but still recognizable, and some segments (perhaps all) across the boundary are left out. Figure 1 illustrates how tonal contours vary with different degrees of word reduction on a disyllabic word, yīn wèi (because).

In running speech, especially on the broadcast news or narrative reading, automatic recognition of Mandarin tones has gained successes in recent years with the benefit of deep neural networks. It is well accepted that suprasegmentals, such as tones are mainly distinguished by pitch, or affected more or less by duration and energy. As more evidence emerges that spectral balance is also a reliable acoustic correlate of lexical contrast of tones, [3, 4] experimented and found that MFCCs outperform prosodic features for automatic recognition of Mandarin tones. With either the DNN-HMM framework in [5] or a convolutional neural network with CTC (Connectionist Temporal Classification) used in [6], they further confirmed the effectiveness of incorporating articulatory information in tone modeling. Recently, in [7, 8], more robust representations of tonal contours were learned in context well with a self-supervised learning framework based on Transformers. However, researchers as in [9] found using an Encoder-Decoder framework with gating mechanism for better pitch tracking performed quite well on downstream tone recognition tasks, even not worse than using MFCCs, FFT spectrograms, and raw waveforms.

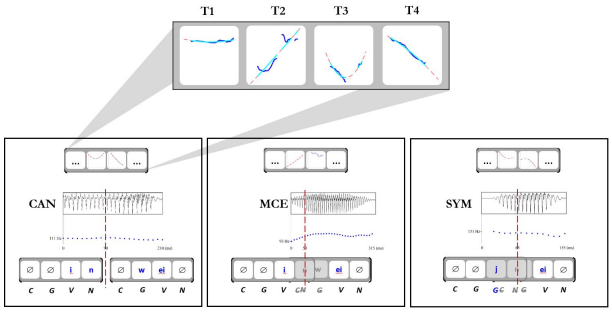


Figure 1: *Four Taiwan Mandarin tonal contours on the top are Tone 1 to Tone 4 in log-transformed f0. The contours of T1 /55/ and T4 /51/ for a disyllabic word, yīn wèi (because) vary with word reduction, from a nearly unchanged CAN to a merging continuum MCE, then toward a thoroughly merged SYM.*

Enlightened by the simplicity of the used input, the novelty of this work lies in two aspects: (1) Attention-mechanism is exploited thoroughly for syllable-wise tone classification while concurrently incorporating rhythmic impact on tonal contours.(2) Jointly trained model on word tone prediction with contextualized contour representations affected by word reductions achieves substantially better results on syllable tone recognition.

2. Conversational speech

The data we use for the task are taken from a released resource, the Mandarin Conversation Dialog Corpus (MCDC) [10] which consists of 16 paired speakers conversing spontaneously about

10.21437/Interspeech.2024-1869

freely selected topics for an hour each. To develop speech resources for investigations in speech science and technology, spontaneously conversed speech needs annotations. To improve tone recognition for conversed speech with estimated f0 requires more information beyond segments. As follows, verified annotations, judgments on word reductions, and rhythmic contrast in conversed speech are clearly evident and provided.

2.1. Annotations in definable, feasible processing units

Using an automatic tool, ILAS aligner [11], a HMM-GMM based forced aligner trained on 39 phones for ordinary syllables and 13 conversational-specific speech sounds, such as filters, particles, word fragments, and paralinguistic sounds, the released MCDC-8 was assured to be thoroughly transcribed and annotated for words, including corrections of tone labels and deviated boundaries [2]. Afterwards, syllable boundaries were automatically aligned under verified word stretch. The natural pauses specified mainly by inhaling and other paralinguistic sounds were also verified, as well as their use as indicators for the beginning and end of inter-pausing utterances (IPUs). This results in 6,060 released speaking turns chunked into 13,407 IPUs. Due to spontaneity, 10% of conversational speech (MCDC-8) are long utterances over 20 syllables encompassing at least two or more grammatical clauses; in comparison, backchannel responses sounding like 'right', 'yeah', 'yes' and 'uh-huh', among others, are short less than three syllables, with a ratio of 14.1%. The remaining IPUs range from 3-20 syllables, covering 75.9%.

2.2. Perceptual judgements on word reductions

An inventory of 1,121 words frequently used in writing and conversation is released from the Sinica Core Vocabulary Inventory [12]. A total of 706 disyllabic words are used to select all spoken tokens in MCDC-8. This ends up with 22,030 disyllabic word tokens of 642 different word types being chosen for advanced perceptual judgements of a word reduction degree. The labeling can be a CAN, a MCE, or a SYM. 400 students were recruited for reduction degree judgments. Each set of nearly 166 word tokens was judged by the same group of three raters.

Prior to a human rater making a judgment, a short training phase was conducted to ensure the perceptual criteria for reduction degree were well established and of the same level of consistency. Nine times consecutively, the rater had to answer the correct type of word reduction degree among a set of 270 randomly selected disyllabic stimuli with the type of reduction degree already determined by a majority vote among three trained labellers. Based on these 270 matched disyllabic stimuli, we compute intraclass correlation coefficients (ICC) using a mean-rating ($k = 3$), consistency, two-way mixed-effects model with three labellers. The resulting ICC value is 0.80, indicating "good" interrater reliability [13]. The 95% confidence interval of ICC ranges between 0.742 and 0.866. Through a majority vote of three raters, 96.3% of 22,030 disyllabic word tokens could eventually be determined with a certain type of word reduction. If none of three types of word reduction won out, we consulted two additional raters' judgements. As a result, the number achieved was 98.3%. Having a sixth rater rated the remaining 1.7% of 368 word tokens, we completed the judging process with a final decision on the reduction degree.

2.3. Rhythmic contrast of different utterance lengths

Acoustic metrics of contrastive speech rhythm, often based on vocalic and intervocalic interval durations, are used for quantifying stable typological differences between languages, or across levels of language proficiency, or even among different speaking styles [14, 15, 16, 17]. To anchor on the average differences between consecutive vocalic intervals, [18] presented a normalized pairwise variability index (nPVI) as a local metric of rhythm on utterance. A variant developed in [19, 20] for calculating speech rate variability in a global sense normalizes the standard deviation of vocalic intervals (ΔV) based on mean duration, named $VarcoV(= \Delta V/meanV)$. These two rhythmic metrics of utterances are extendible to other intervals of interest, such as final parts of Chinese syllables or whole syllabic stretches.

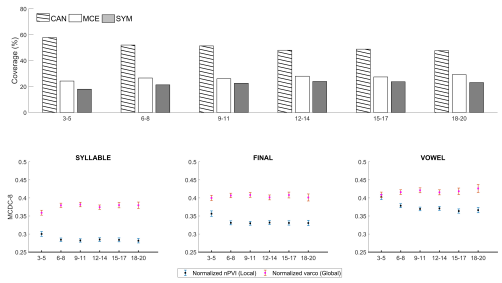


Figure 2: *The reduction distribution and rhythmic contrast*

For MCDC-8, each IPU calculates the nPVI-Inv (local) and varco-Inv (global) based on different types of consecutive intervals (abbreviated as Inv), such as on syllables, or on the finals, which ignore the initial consonants, or on the vocalic parts within syllables. We further normalize the measures of nPVI-Inv and varco-Inv by min-max standardization within stretches of the same speaker since they differ not just from utterance to utterance, but also from speaker to speaker. With speaker effects removed, as shown in Figure 2, the rhythmic contrast in two rhythmic measures widens along with an increasing in utterance length, as more reduced disyllabic words (i.e. with more MCE and SYM) are introduced in response to spontaneity.

3. Joint net for tone recognition

This work is inspired by the model architecture presented in [21], in which tone classification is anchored on contour representations of four lexical tones learned only from suprasegmental information, such as pitch and rhythm. Their J-ToneNet consists of two transformer-based encoding networks (i.e., the contour encoder and rhythm encoder, also abbreviated as C-Net and R-Net, respectively) , summarizing syllabic contour and rhythmic representations from pitch tokens and durational variabilities on syllables. Then these representations are fused together for jointly predicting tones in sequenced syllables and words. Current model is enhanced within three aspects: (1) An transformer-based interaction encoder is further stacked on top after fusing syllable-level contour and rhythmic representations from C-Net and R-Net. (2) To gain local and global rhythmic information about syllables, R-Net is fed three measures of durational variability over different unit intervals. (3) A loss term for word tones, denoted as $\mathcal{L}_{wt}$, is jointly considered under different specified sets of reduction types (RTs).

3.1. More interaction cared via attention-mechanism

The input vector of C-Net is an ℓ-length $f0$ sequence (also known as pitches) in one speech utterance, noted as $\mathcal{P} = \{p_1, p_2, \ldots, p_\ell\} \in \mathbb{R}^{1\times\ell}$. With the aligned boundary information both on syllable and word, C-Net manages to abstract

tone contour representations in response to m-length syllables. The C-Net first introduces a building block $f : \mathcal{P} \mapsto \mathcal{H}$, which takes $\mathcal{P}$ pitches as input and produces latent representations $h_0, h_1, \ldots, h_\ell$. This f contains a projection layer with a learnable weight matrix $W_f \in \mathbb{R}^{d\times 1}$ to project the one-dimensional pitch inputs p onto high-dimensional representations ($d = 64$). Additionally, to account for pitch ordering, token changing of different syllables and of different words within the input sequence, other building blocks featuring formations of embeddings on POSITION, PITCH TOKEN$_{@\text{SYL}}$, PITCH SEGMENT$_{@\text{SYL}}$, PITCH TOKEN$_{@\text{WORD}}$ and PITCH SEGMENT$_{@\text{WORD}}$ are also considered and summed up as $\mathcal{H}$.

Through a Transformer $g : \mathcal{H} \mapsto \mathcal{C}$, contour representations $c_0, c_1, \ldots, c_m$ were generated in aim of capturing the contour shape across entire pitch sequences. As the formal algorithms for transformers indicate in [22], the Transformer g in C-Net comprises stacked transformer layers ($L = 4$) that encode individual latent representations $\mathcal{H}$ with attention. The latent representation $\mathcal{H}$, as shown in Equation 1 and 2, is sequentially processed through the transformer layers to produce the contour representation $\mathcal{C}$.

$$\tilde{\mathcal{H}}^l = LN(\mathcal{H}^{l-1} + MHAtt(\mathcal{H}^{l-1}, \mathcal{H}^{l-1}, \mathcal{H}^{l-1})) \quad (1)$$

$$\mathcal{H}^l = LN(\tilde{\mathcal{H}}^l + FFN(\tilde{\mathcal{H}}^l)) \quad (2)$$

To obtain the rhythmic representations $\mathcal{R}$ in the same dimension of 64, R-Net employs an architecture closely resembling that of C-Net to encode the rhythmic information $\mathcal{V}$ on syllables. The input $\mathcal{V} = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{9\times m}$ is a sequence of m-length duration tokens that have 9-dimensional features measuring durational variations that occur pairwise or globally. Section 3.2 will detail the durational features used, and here R-Nets and C-Nets differ in two key aspects as follows: (1) The number of transformer layers used in R-Net's Transformer is reduced to ($L = 2$). (2) As opposed to pitch tokens, R-Net embedding blocks deal syllables instead.

To explore contextualized effects of syllables in greater detail, the obtained contour and rhythmic representations $\mathcal{C}$ and $\mathcal{R}$ are fused first as they were in [21], and then passed to an interaction encoder for final consolidation of contour representations on tones, resulting in the hidden outputs $\mathcal{Q}$. The fused representations are initially summed with embeddings on POSITION and SYLLABLE SEGMENT$_{@\text{WORD}}$; and after that, a stacked transformer layer with $L = 2$ is applied. The POSITION embeddings follow the ones defined in [23] to use sinusoidal version to mark the ordering of syllables. To signify how cared syllable tokens change within a word stretch, the SYLLABLE SEGMENT$_{@\text{WORD}}$ embeddings use two symbols, E_A and E_B in a similar fashion to PITCH SEGMENT$_{@\text{SYL}}$. The later cares about pitch token changes from one syllable to another. As a result of self-attention, syllables will interact more and tones will be recognized more accurately when spoken in conversation.

3.2. More durational variations considered for rhythm

As evidenced by the rhythmic contrast between local and global rhythmic measures on conversed speech, a syllable's durational variability can be quantitatively measured in three ways over three specified intervals. The three intervals are a whole syllable stretch (SYL), a final part without the initial consonant ($FINAL$) and a vocalic part only (V). Measures include: (1) simply the duration of the defined interval; (2) the current interval's deviation from the utterance's mean duration over a defined interval, $Devi\mu$; (3) the averaged durational difference between consecutive intervals, $\mu Diff$.

$$Devi\mu_i = Inv_i - \mu_{Inv_i} \quad (3)$$

$$\mu Diff_i = \frac{Inv_i - Inv_{i+1}}{(Inv_i - Inv_{i+1})/2} \quad (4)$$

where Inv = interval under observation (either V, $FINAL$, SYL) and Inv_i = duration of the i-th interval. At the m-th interval, $\mu Diff_\ell$ is replaced with a nPVI-Inv, i.e., a normalized pairwise variability index for a certain type of speech intervals aiming of measuring the speech rhythm of an utterance [18]. Experimenting with this expanded version of durational features used in [21], the rhythm encoder is called exp-R-Net.

3.3. Joint word tone loss with account of reductions

During training, the network constantly learns contour representations of tones from two tasks: $\mathcal{L}_{st}$ for identifying individual syllable tones, and $\mathcal{L}_{wt}$ for identifying tones of words within a specified set taken into account particular RTs.

$$\mathcal{L} = \mathcal{L}_{st} + \alpha\mathcal{L}_{wt} \quad (5)$$

where α is a tuned hyperparameter set to 0.4, yielding the highest result within a searched range of [0, 1] on the development set.

Syllable Tone Loss. With the predicted syllable tone probability sequence $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m] \in \mathbb{R}^{m\times K}$, and given the output $\hat{y}_i$ centered on i-th syllable in the speech utterance, the model needs to recognize the labeled tone y_i with the loss defined as:

$$\mathcal{L}_{st} = -\sum_{k}^{K} y_i^k log(\hat{y}_i^k) \quad (6)$$

where the four lexcial tones are targeted with $K = 4$.

Word Tone Loss. To consolidate contextualized contour representations for syllable tone classification, the joint loss $\mathcal{L}_{wt}$ is designed to account for the effects of word reduction and utilize the contour representations learned from word tones. We utilize the word tone sequence $Z = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{n\times K}$ from the same speech utterance for network training, where n represents the number of word tones in the speech utterance, and $n \leq m$. Two word types, monosyllabic words and disyllabic words without RT labels are defaulted, abbreviated WTD, and unioned with the RT set $\{\text{CAN}, \text{MCE}, \text{SYM}\}$ specified on filtered disyllabic words to form a set, designated as swr. With a control gate ψ_j, the word tone loss is as follows:

$$\mathcal{L}_{wt} = -\sum_{k}^{K} \psi_j z_j^k log(\hat{z}_j^k), \psi_j = \begin{cases} 1, \text{if chosen}(w_j) \in swr, \\ 0, \text{otherwise} \end{cases} \quad (7)$$

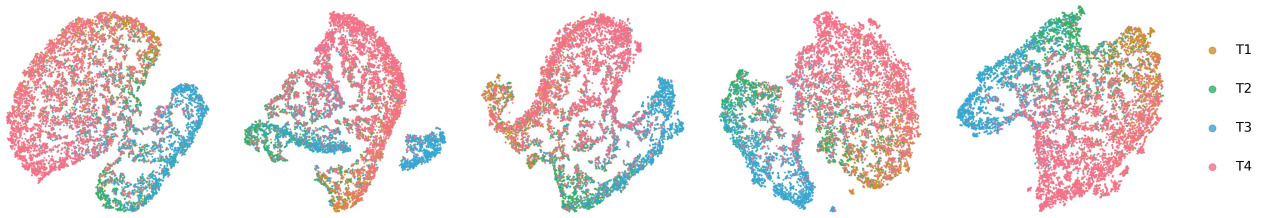
where ψ_j indicates whether the loss of the j-th word w_j should be used concurrently to update the network. Here, any tone combinations with neutral tones are excluded and K stands for the number of word tones to be predicted. As a result, 20 of them are slated for classification.

Classification Network. To predict the tones of syllables or words, we designed the same, two-layer feed-forward classification network to be applied to output representations of q_i or q_j. In Equation 8, syllable tone classification is illustrated by first applying a non-linear transformation δ (tanh) to the output before feeding it to the softmax layer (σ). The parameters $W_1 \in \mathbb{R}^{d\times d}$, $W_2 \in \mathbb{R}^{K\times d}$, b_1, and b_2 are the learnable weights and bias terms, where d stands for the input dimension and K is the number of tones intended to be classified, which are syllables or words.

$$\hat{y}_i = \sigma(W_2(\delta(W_1 q_i + b_1)) + b_2) \quad (8)$$

Table 1: *Tone accuracy and F1 score (%) of models.*

Model	#Param.	Acc. (%)		F1 Score (%)				Inference Time (ms)
		dev	test	T1	T2	T3	T4	
ToneNet (con) [21]	0.20M	59.4	58.9	0.478	0.440	0.603	0.683	1.7
re-ToneNet (con $\rightarrow$ int)	0.30M	61.1	60.0	0.499	0.483	0.615	0.688	2.4
re-com-ToneNet (con $\oplus$ rhy $\rightarrow$ int)	0.48M	**67.0**	**65.8**	**0.571**	**0.537**	**0.691**	**0.732**	37.3
re-com-J-ToneNet ($swr\backslash\{$MCE, SYM$\}$)	0.48M	66.9	65.8	0.547	0.558	0.692	0.732	38.4
re-com-J-ToneNet ($swr\backslash\{$SYM$\}$)	0.48M	68.7	68.1	0.601	0.570	0.708	0.750	38.4
re-com-J-ToneNet (swr)	0.48M	**71.5**	**70.3**	**0.626**	**0.600**	**0.722**	**0.771**	38.4



(a) (b) (c) (d) (e)

Figure 3: *t-SNE on models: (a) re-ToneNet (con $\rightarrow$ int) (b) re-com-ToneNet (con $\oplus$ rhy $\rightarrow$ int) (c) re-com-J-ToneNet ($swr\backslash\{$MCE, SYM$\}$) (d) re-com-J-ToneNet ($swr\backslash\{$SYM$\}$) (e) re-com-J-ToneNet (swr)*

4. Ablations and analyses with t-SNE

13,407 IPUs of conversed speech were divided into training, development, and test sets at a rate of 80%, 10%, and 10% respectively. The baseline ToneNet relies solely on the C-Net specified in [21] to learn contour representations without any rhythmic information and joint learning. As follows, several network components with certain kinds of revisions are stacked onto the baseline ToneNet stage by stage to improve model performance. First, extra transformer layers are stacked on and take in contour representations at the beginning of syllables to capture more coarticulated effects between syllable contours, referred to as the interaction encoder; henceforth, reported as re(vised)-ToneNet. Next, a expanded version R-Net (exp-R-Net) in transformers is used along with the C-Net to account for more rhythmic information, hereafter named re-com(plete)-ToneNet. Finally, we consider word tone classification loss in the re-com-J(oint)-ToneNet, where the words are chosen based on degrees of reduction.

In Table 1, the re-com-ToneNet model achieves an accuracy of 65.8%, improving a rate of around 7% both in the development and test sets than baseline ToneNet. Adding an interaction encoder after rhythmic representations are fused benefits T1, T2 and T3 by around 10% each, while T4 benefits by about 5%. Compared to baseline ToneNet, F1 scores show a trend of T4 > T3 » T1 > T2, where "»" signifies the preceding twos prevailing over the following twos by more than 10%. Further, when using the interaction encoder only on C-Net, reported as re-ToneNet, T2 increases its F0 score by over 4%, suggesting that focusing more on contextualized contours might moderate the confusion associated with contours.

Right beneath the double horizontal line, the models integrate word tone loss $\mathcal{L}_{wt}$. Chosen word tones are from two types of words already specified as WTD and filtered disyllabic words with three reduction degrees, CAN, MCE, and SYM. With the exclusion of disyllabic words that are moderately reduced and solely merged, i.e., the cases on MCE and SYM labels, word tone loss did not improve syllable tone accuracy. The F1 scores for T3 and T4 are almost unchanged, while T1 is slightly worse and T2 is marginally better than found in re-com-ToneNet. With moderately reduced and sorely merged words, syllable tone accuracy improves steadily up to 5%. In other words, by jointly learning word tones, deep features in tandem with more reductions considered in the loss $\mathcal{L}_{wt}$, make automatic tone recognition more robust for conversational speech.

To reduce the dimensionality of the latent space for visualizing contour representations of different modeling stages, we applied t-SNE [24]. Our results are presented in Figure 3. The changes in contour clustering in Figure 3a and 3b suggest that T1 is more or less separated from T4 and is developing into its own cluster. Furthermore, some contour patterns of T3 began grouping and could not be easily confused with other tones, but some were not. In Figure 3c, the contour representations begin to form distinct clusters for different tones with four clear grouping directions. As we move further from Figure 3d to Figure 3e, with more reductions considered, the distances between tone groups would become wider, while the confused area would become narrower, centered on the cross area. This indicates that the revised components in proposed network have certain impact of learning contour representations for tone discrimination.

5. Conclusions

Conversational speech is often a challenge for automatic tone recognition, particularly when utilizing suprasegmental information only. Our work integrates rhythmic variability incurred from word reduction into the network, increasing accuracy by 12%. The approach is effective and validated through experiments with revisions on: (1) incorporating the interaction encoder at syllable-level with the embeddings defined as SYLLABLE SEGMENT$_{@\text{WORD}}$, (2) enriching rhythmic information on syllables with the expanded durational features, and (3) adding a word tone loss $\mathcal{L}_{wt}$ constrained further by specified swr. Our future research will aim to detect word reduction degrees through information learned from spectrum changes across boundaries.

6. Acknowledgements

The work was financially supported by the National Science and Technology Council, Taiwan, with project [112-2221-E-035-067] granted to the second author.

7. References

[1] A. Lee, S. Prom-on, and Y. Xu, "Pre-low raising in Cantonese and Thai: Effects of speech rate and vowel quantitya," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 179–190, Jan. 2021. [Online]. Available: https://doi.org/10.1121/10.0002976

[2] Y.-F. Liu, S.-C. Tseng, and J.-S. R. Jang, "Deriving disyllabic word variants from a Chinese conversational speech corpus," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 308–321, Jul. 2016. [Online]. Available: https://doi.org/10.1121/1.4954745

[3] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4868–4872.

[4] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information," in *Proc. Speech Prosody*, 2014, pp. 673–677.

[5] J. Lin, W. Li, Y. Gao, Y. Xie, N. F. Chen, S. M. Siniscalchi, J. Zhang, and C.-H. Lee, "Improving mandarin tone recognition based on dnn by combining acoustic and articulatory features using extended recognition networks," *Journal of Signal Processing System*, vol. 90, no. 7, pp. 1077–1087, Jul. 2018.

[6] L. Lugosch and V. S. Tomar, "Tone Recognition Using Lifters and CTC," in *Proc. Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 2305–2309.

[7] J. Yuan, N. Ryant, X. Cai, K. Church, and M. Liberman, "Automatic recognition of suprasegmentals in speech," *arXiv preprint arXiv:2108.01122*, 2021.

[8] J. Yuan, X. Cai, and K. Church, "Improved Contextualized Speech Representations for Tonal Analysis," in *Proc. INTERSPEECH 2023 - 24th Annual Conference of the International Speech Communication Association*, 2023, pp. 4513–4517.

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

[10] S.-C. Tseng, "Spoken corpora and analysis of natural speech." *Taiwan Journal of Linguistics*, vol. 6, no. 2, pp. 1–26, 2008.

[11] ——, "Ilas chinese spoken language resources," in *Proc. LPSS 2019 - 3rd International Symposium on Linguistic Patterns in Spontaneous Speech*, 2019, pp. 13–20.

[12] ——, "Lexical coverage in taiwan mandarin conversation," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 18, no. 1, pp. 1–18, Mar. 2013.

[13] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[14] L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. L. Mattys, "How stable are acoustic metrics of contrastive speech rhythm?" *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1559–1569, Mar. 2010. [Online]. Available: https://doi.org/10.1121/1.3293004

[15] V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, Mar. 2015. [Online]. Available: https://doi.org/10.1121/1.4906837

[16] M. Ordin and L. Polyanskaya, "Development of timing patterns in first and second languages," *System*, vol. 42, pp. 244–257, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0346251X13001802

[17] ——, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 533–544, Aug. 2015. [Online]. Available: https://doi.org/10.1121/1.4923359

[18] E. Grabe and E. L. Low, *Durational variability in speech and the Rhythm Class Hypothesis*. Berlin, New York: De Gruyter Mouton, 2002, pp. 515–546. [Online]. Available: https://doi.org/10.1515/9783110197105.2.515

[19] V. Dellwo, *Rhythm and Speech Rate: A Variation Coefficient for deltaC*. Frankfurt am Main, Germany: Peter Lang Publishing Group, 2006, pp. 231–241.

[20] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447007000101

[21] Y.-F. Liu and X.-L. Lu, "J-ToneNet: A Transformer-based Encoding Network for Improving Tone Classification in Continuous Speech via F0 Sequences," in *Proc. INTERSPEECH 2023 - 24th Annual Conference of the International Speech Communication Association*, 2023, pp. 2138–2142.

[22] M. Phuong and M. Hutter, "Formal algorithms for transformers," *arXiv preprint arXiv:2207.09238*, 2022.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html