INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improved Contextualized Speech Representations for Tonal Analysis

Jiahong Yuan[1], *Xingyu Cai*[2], *Kenneth Church*[3]

[1]University of Science and Technology of China, China
[2]Google Inc., USA
[3]Northeastern University, USA

jiahongyuan@ustc.edu.cn, xingyu.develop.cai@gmail.com, kenneth.ward.church@gmail.com

Abstract

We propose fine-tuning wav2vec2.0 with a cross-entropy loss to classify tones in an utterance on a frame-by-frame basis. Our study demonstrates that this approach not only improves tone classification accuracy but also generates frame-level representations suitable for tonal analysis. By using these representations, we established that the third-tone-sandhi-rising tone in Mandarin speech differs from the lexical rising tone, and the third tone that doesn't undergo sandhi differs from the third tone that's not in a sandhi context. Our findings suggest that third-tone sandhi in Mandarin Chinese involves a continuous shift from Tone3 to Tone2, rather than a categorical change. [1]

Index Terms: speech analysis, tone, contextualized representations, alignment, classification

1. Introduction

Mandarin Chinese is a tone language with four lexical tones (Tone1 to Tone4) and a neutral tone (Tone5). The phonetic realizations of tones can vary greatly due to factors such as speaker physiology, speaking style and rate, and contextual influences. As a result, recognizing tones in running speech automatically is challenging. In recent years, the use of deep learning models for Mandarin tone recognition has been successful. In particular, the method of fine-tuning wav2vec2.0 with a CTC (Connectionist Temporal Classification) loss has achieved impressive results.

In this paper, we present a method for improving the classification of Mandarin tones by finetuning wav2vec2.0 using a cross-entropy loss for all frames in an utterance. The aim of this approach is to address the issue of peaky output distributions associated with CTC, enabling the use of the model's representations for tonal analysis. Our study demonstrates that the proposed method not only enhances tone classification accuracy but also produces frame-level representations that are suitable for tonal analysis. To illustrate the effectiveness of these representations, we conduct a case study on third tone sandhi in Mandarin Chinese.

2. Related work

2.1. Contextual variation of tones and third tone sandhi

The fundamental frequency, F_0, is the primary acoustic cue for tones in Mandarin Chinese. The four lexical tones, Tone1 to Tone4, are transcribed using a five-point scale of tonal transcription where 1 represents the lowest pitch and 5 represents the highest pitch. Specifically, the four tones are transcribed as 55, 35, 214, and 51, respectively. Besides F_0, other parameters such as duration, amplitude, and voice quality also play a role in the production and perception of tones [1, 2]

Context plays a significant role in the phonetic realization of tones. For instance, a lexical falling tone that appears before a high tone and after a low tone could even be realized as rising, as demonstrated in [3]. Numerous studies have explored contextual variations of tones, considering perspectives such as coarticulation [4], physiological constraints [5], and the interaction between tone and prosodic structure [6]. Several models have been proposed to aid in this endeavor. For instance, in the Stem-ML model [7], tones are treated as flexible templates that can be adjusted due to the interaction between neighboring tones and with other components of prosody. The PENTA model [8] simulates the articulatory realization of underlying pitch targets and provides an operational framework that enables the simultaneous encoding of multiple communicative functions.

Unlike tonal coarticulation, tone sandhi refers to the phonological alteration of lexical tones in a linguistic context. This is a common occurrence in Chinese dialects [9]. In Mandarin Chinese, when a Low tone (Tone3) precedes another Low tone, it is typically pronounced with a rising F0 contour, which is known as third tone sandhi. This phenomenon has been extensively studied in the literature, including the domain of third tone sandhi, which examines how the sandhi is applied across linguistic boundaries [10]. Figure 1 depicts two examples from our research. Both words consist of three third tones, with the first word featuring rising contours on the first two third tones, whereas in the second word, only the second third tone has a rising contour.

An intriguing issue regarding third tone sandhi is whether there is an acoustic distinction between a Sandhi Rising tone and an underlying Rising tone (Tone2). This raises the question of whether third tone sandhi is a complete or incomplete neutralization process. According to [11], who investigated the third tone sandhi in a conversational speech corpus, there are differences between an underlying Rising tone and a Sandhi Rising tone in terms of F_0 rise magnitude and duration of the rise. [12] also showed that the F_0 maximum of a Sandhi Rising tone is lower than that of an underlying Rising tone. Despite the reported acoustic differences, there is no conclusive evidence that listeners can use these subtle acoustic differences to differentiate between the two tones.

2.2. Automatic recognition of tones

In recent years, deep learning models have been successfully employed for Mandarin tone recognition. For example, [13, 14] built a deep neural network using MFCCs to classify tones in Mandarin Chinese, achieving significant improvement com-

[1] This work was done when the authors were at Baidu Research, USA.

10.21437/Interspeech.2023-283

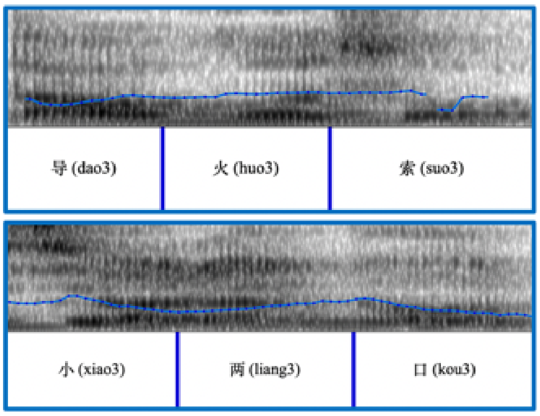

Figure 1: *Third tone sandhi in Mandarin Chinese. The first two tones change to rising in "dao3 huo3 suo3"; whereas only the second tone changes to rising in "xiao3 liang3 kou". F_0s are shown in blue.*

pared to traditional methods that used prosodic features, even though F_0 and other pitch-related features were omitted. In a DNN-HMM framework, [15] investigated the effectiveness of incorporating articulatory information into tone modeling, either by explicitly adding the articulatory features or building phone-dependent tonal models. Their study confirmed that the DNN model can extract useful information from the MFCC parameters for tone recognition and incorporating articulatory information can further improve tone recognition. [16] proposed a method for tone recognition using a convolutional neural network with CTC and achieved a tone error rate of 11.7% on the Aishell-1 dataset [17]. [18] proposed a multi-scale model that gathers information at multiple resolutions to better capture tone variations, achieving competitive results on the Chinese National Hi-Tech Project 863 corpus with a tone error rate of 10.5%. Finally, [19] reported that feeding both the Mel-spectrogram and the short-term context segment features into an end-to-end model could significantly improve automatic tone recognition, with classification accuracy improving from 79.5% to 88.7% on the Aishell-3 database [20].

2.3. Finetuning wav2vec2.0 with CTC loss for tone recognition

Wav2vec2.0 [21] is a self-supervised learning framework based on Transformers that can learn speech representations from raw audio data. The framework processes the speech signal with a multilayer convolutional network to extract latent features every 25 ms. The latent features are then fed into vector quantization and transformer networks. The pre-trained models of wav2vec2.0 can be fine-tuned for speech recognition by using labeled data and optimizing with a Connectionist Temporal Classification (CTC) loss [22].

In a recent study, [23] fine-tuned the wav2vec2.0 framework using a CTC loss to recognize suprasegmentals such as syllables, tones, and pitch accents. The results demonstrated a 50% error reduction in Mandarin tone recognition compared to previous studies. We believe that one of the advantages of using wav2vec2.0 for tone recognition is the self-attention mechanism in its Transformer component [24]. The self-attention mechanism allows the model to effectively learn and model contextual influences on tones.

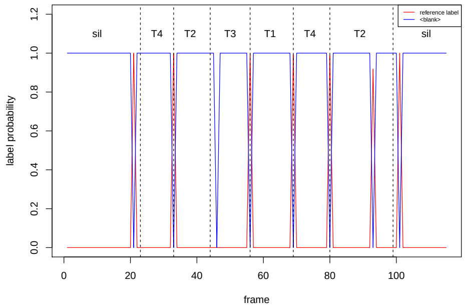

Figure 2: *Output predictions from CTC are spiky. Most frames are mapped to <blank>.*

Despite its success in predicting a label sequence, the output predictions from CTC are often very sparse, with most frames being predicted as <blank>(i.e., no label). Figure 2 shows an example of the output predictions of tone recognition with CTC. As can be seen from the figure, most frames are mapped to <blank>, and therefore, the learned representations of these frames cannot be effectively utilized for tonal analysis.

CTC loss is designed for tasks in which the alignment of input and output sequences is difficult to achieve, for example, the alignment between acoustic frames and characters in ASR. For the task of tone recognition, however, we can first map each acoustic frame to a tone label through forced alignment, and then fine-tune wav2vec2.0 with a cross-entropy loss for frame-wise classification. We expect no spiky predictions from employing a cross-entropy loss, and therefore the learned representations will be useful for tonal analysis.

2.4. Contextualized representations from wav2vec2.0

The pre-trained wav2vec2.0 contextualized representations capture a rich amount of information about speech, as demonstrated by probing experiments that show their effectiveness on a wide range of tasks [25, 26]. In our study, we found that the representation space of a fine-tuned wav2vec2.0 model is very different from that of a pre-trained model. Figure 3 shows a t-SNE plot of contextualized representations from three models on a Chinese dataset: pre-trained wav2vec2.0, wav2vec2.0 fine-tuned for phone recognition, and wav2vec2.0 fine-tuned for tone recognition. It is clear that these representations occupy distinct spaces.

Compared to the pretrained wav2vec2.0 model, models that have been fine-tuned encode task-specific properties in their learned representations. For instance, a wav2vec2.0 model that has been fine-tuned for tone recognition will capture tone-related properties such as pitch patterns and tonal coarticulation. Therefore, contextualized representations from a model fine-tuned for tone recognition are likely to be more effective for analyzing tones in speech than those from the pretrained model.

3. Fine-tuning wav2vec2.0 with cross-entropy loss for tone classification

3.1. Data and alignment

We performed our experiments on the Aishell-1 [17] dataset, which is a widely-used benchmark for Mandarin ASR. The

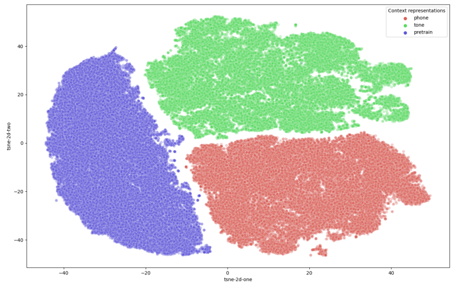

Figure 3: *The space of contextualized representations from three wav2vec2.0 models: pre-trained only, fine-tuned for phone recognition, and fine-tuned for tone recognition.*

dataset includes 165 hours of read speech in Mandarin Chinese from 400 speakers belonging to different dialect regions, with the majority from the northern areas. The dataset is divided into train, dev, and test sets, containing approximately 1.7m, 200k, and 100k tones, respectively.

An HMM-GMM based forced aligner was trained on the dataset using the HTK toolkit. The aligner utilizes a pronouncing dictionary provided in the dataset, which transcribes words into initials and tonal finals in *Pinyin*, a Roman alphabet system of phonetic transcription for Mandarin Chinese.

Using the results of forced alignment, a label sequence is created at every 25ms for each utterance to ensure that the same number of labels and acoustic frames are extracted by wav2vec2.0. The frame-aligned audio and label sequences can then be utilized to finetune wav2vec2.0 with a cross-entropy loss.

Three label creation strategies were implemented: 1. Treating initials as part of a tone; 2. Mapping initials to a separate label ”C”; 3. Mapping only the center frame of a tonal final or silence to a tone or ”sil”, while mapping all other frames to ”O”. For example:

- *Label1*: sil sil sil sil sil T2 T2 T2 T2 T2 T2 T2 . . .
- *Label2*: sil sil sil sil sil C C T2 T2 T2 T2 T2 . . .
- *Label3*: O O sil O O O O O O T2 O O . . .

3.2. Tone classification

The process of finetuning wav2vec2.0 with a cross-entropy loss involves the following steps: First, a randomly initialized linear projection is added on top of the contextual representations of wav2vec2.0 to map the representations into label tokens. Then, the entire network is optimized by minimizing a cross-entropy loss through finetuning. Unlike building a classifier on all frames as independent data points, the finetuning process can learn associations among frames in an utterance.

In our experiments, we utilized the wav2vec2.0 large model pre-trained on 960 hours of Librispeech audio (*libri960_big.pt*) for finetuning. Initially, only the output classifier is trained for the first 10,000 updates, after which the Transformer is updated as well. We set the max tokens to 1 million, which is equivalent to 62.5 seconds of audio with a sampling rate of 16 kHz, and the learning rate to 1e-5. The unit error rate on the dev set was the metric used to determine the total number of updates.

The test set consisted of a total of 1.8 million frames. The frame classification accuracy for *Label1* and *Label2* was 94.0% and 93.5%, respectively. For *Label3*, the majority of frames (1.68 million out of 1.8 million) were ”O”. Among the ”O” frames, 99.7% were classified as either a tone or silence. Excluding the ”O” frames, the frame classification accuracy for *Label3* was 96.4%.

Table 1: *Results of fine-tuning wav2vec2.0 with CE loss for tone classification using different label strategies, compared to fine-tuning with CTC loss.*

Loss function	Labels	Classification accuracy	Recognition errors
cross entropy	T1-T5,sil	95.5%	-
	T1-T5,C,sil	95.0%	-
	T1-T5,O,sil	95.7%	-
CTC	T1-T5,sil	2.5%	Sub: 5.7% Ins: 0.4% Del: 0.4%

The test set comprises 100k tones, and we used the center frame to calculate tone classification accuracy. Table 1 displays the results. Additionally, we performed tone classification on the center frames using a model fine-tuned with a CTC loss on the same dataset. With this model, as demonstrated in section 2.3, the majority of frames were classified as <blank>, leading to an extremely low accuracy in classification. The tone recognition error rate from the CTC model was 6.6%. A rough comparison between the cross-entropy and CTC models can be made by examining their classification and substitution error rates, respectively. The cross-entropy models exhibit classification error rates ranging from 4.3% to 5.0%, which is superior to the CTC model’s substitution error rate of 5.7%.

4. Using contextualized representations for tonal analysis

4.1. Clustering of tones

By fine-tuning wav2vec2.0 for frame-wise tone classification, we can extract contextualized representations from the model that are useful for tonal analysis. These representations were learned in context and contain information about contextual associations and influences. Consequently, there is no need for normalization when utilizing contextualized representations for tonal analysis, unlike with F_0 and other acoustic features.

In Figure 4, we compare the effectiveness of F_0 features and contextualized representations in clustering tones in the test set. The F_0 mean and slope of tonal contours were computed from normalized F_0s, i.e., semitones above the 5th percentile of F_0s in an utterance. The contextualized representations were extracted from the center frame of each tone and have a dimension of 1024. We trained a PCA model on the representations of the training set, and applied it to the test set. Figure 4 shows the first two dimensions of the PCA projection of the test set, where the ellipses represent 95% of the data points in a given category. The results demonstrate that contextualized representations are much more effective than F_0 features for clustering tones.

To quantify the differences, we performed 5-means clustering on two sets of features: the two F_0 features and the two dimensions of PCA projection. We then mapped the resulting clusters to five tones. The results showed that the clustering accuracy was 44.4% for the F_0 features and 87.7% for the contextualized representations on the test set. When using all 1024

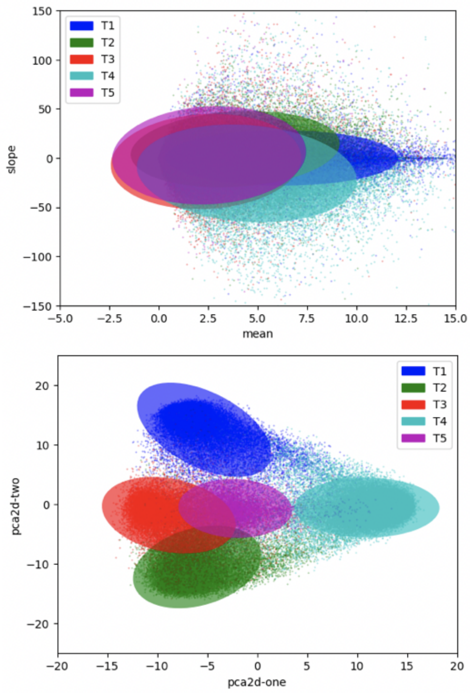

Figure 4: *Tones are better separated in the space of contextualized representations (bottom) than in F_0 features (top).*

dimensions, the clustering accuracy of the contextualized representations increased to 93.7%.

4.2. Third tone sandhi

We cannot predict whether a third tone will undergo sandhi or not based solely on text. Therefore, to investigate third tone sandhi in the dataset, we treated all Tone3s in a sandhi context, i.e., those preceding another Tone3, as a separate category called Tone6. We labeled Tone6 as "O" for fine-tuning wav2vec2.0 with a cross-entropy loss, using the strategy of *Label3* described in section 3.1.

Using the fine-tuned wav2vec2.0 model, we classified Tone6 in the dataset. The training set contained 35,573 Tone6s, of which 19,994 (56.2%) were classified as Tone2, 12,837 (36.1%) as Tone3, and the remainder as other tones or silence. In the test set, 1,187 (56.3%) of the total 2,107 Tone6s were classified as Tone2, and 739 (35.1%) as Tone3. Overall, more than 90% of Tone3s in a sandhi context were classified as either Tone2/sandhi or Tone3/non-sandhi, and the ratio between the two was about 1.6:1. Figure 5 shows the representations (the first two dimensions of PCA projection) of Tone2, Tone3, and Tone6, demonstrating that some Tone6s undergo sandhi while others do not.

To measure the similarity or distance between two tones, we can use the label probabilities from the classification model. Specifically, we define a metric *d* = *logprob(Tone2)-logprob(Tone3)*, where *logprob(X)* is the log probability of a frame being *X*. The metric *d* measures how much more likely the tone is Tone2 than Tone3. Figure 6 shows *d* values for four types: 1. *T2->T2*: Tone2; 2. *T6->T2*: Tone3 in a sandhi context and classified as Tone2 (sandhi); 3. *T6->T3*: Tone3 in a sandhi context and classified as Tone3 (non-sandhi); and 4. *T3->T3*: Tone3 not in a sandhi context.

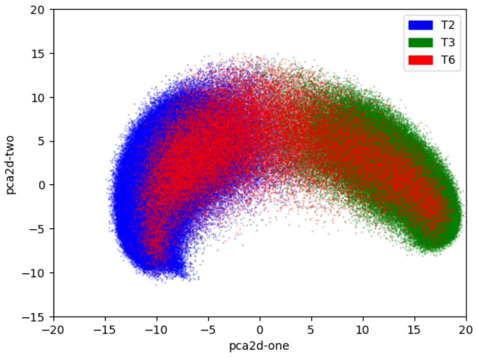

Figure 5: *Representations of Tone6 (i.e., Tone3 in a sandhi context), Tone2, and Tone3. Some Tone6s undergo sandhi while others do not.*

The boxplots show that *T6->T2* has lower *d* values than *T2->T2*, which confirms previous findings that the sandhi rising tone is different from Tone2. Interestingly, *T6->T3* has higher *d* values than *T3->T3*. Two hypotheses may explain this result. First, the difference may be due to contextual influences on tone. *T6->T3* only appears before a Tone3, whereas *T3->T3* does not appear in that context. Secondly, it is possible that third tone sandhi is not a categorical change, but a continuous shift from Tone3 to Tone2. Further research is needed to test these hypotheses.

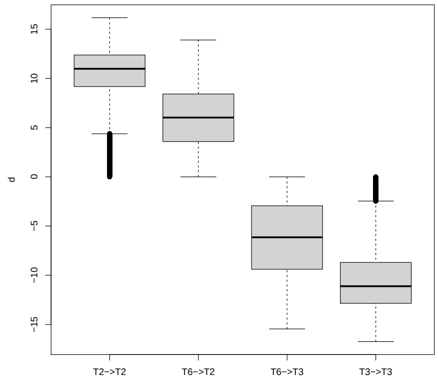

Figure 6: *Boxplots of d values. Sandhi Tone3 (T6->T2) is lower than Tone2; Non-sandhi Tone3 (T6->T3) is higher than Tone3.*

5. Conclusions

The paper presents a new method to classify Mandarin tones using the wav2vec2.0 model, which involves fine-tuning it with a cross-entropy loss for all frames in an utterance. Our study shows that this approach not only improves tone classification accuracy but also enables tonal analysis through frame-level representations generated by the model. Based on these representations, we demonstrate that the third-tone-sandhi-rising tone in Mandarin speech differs from the lexical rising tone, and the third tone not undergoing sandhi also differs from that not in a sandhi context. Our findings suggest that third-tone sandhi in Mandarin Chinese involves a continuous shift from Tone3 to Tone2, rather than a categorical change.

6. References

[1] J. Yuan and K. Church, "Speaking rate and tonal realization in mandarin chinese: What can we learn from large speech corpora?" *Proceedings of ICASSP 2021*, pp. 6463–6467, 2021.

[2] Y.-Y. Kong and F.-G. Zeng, "Temporal and spectral cues in mandarin tone recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2830–2840, 2006.

[3] C. Shih and G. Kochanski, "Chinese tone modeling with stem-ml," *Proceedings of Interspeech 2000*, 2000.

[4] Y. Xu, "Contextual tonal variations in mandarin," *Journal of Phonetics*, vol. 25, pp. 61–83, 1997.

[5] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *Journal of the Acoustical Society of America*, vol. 111, pp. 1399–1423, 2002.

[6] J. Yuan, *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modeling*. Cornell University: PhD thesis, 2004.

[7] G. Kochanski and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, pp. 311–352, 2002.

[8] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[9] M. Chen, *Tone Sandhi*. Cambridge University Press, 2000.

[10] C. Shih, "Mandarin third tone sandhi and prosodic structure," in *Studies in Chinese Phonology*, J. Wang and N. Smith, Eds. De Gruyter Mouton, 1997, pp. 81–124.

[11] J. Yuan and Y. Chen, "3rd tone sandhi in standard chinese: A corpus approach," *Journal of Chinese Linguistics*, vol. 42, 2014.

[12] S. Peng, "Lexical versus 'phonological' representations of mandarin sandhi tones," in *Language acquisition and the lexicon: Papers in laboratory phonology V*, M. Broe and J. Pierrehumbert, Eds. Cambridge University Press, 2000, pp. 152–167.

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

[14] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," *Proceedings of Speech Prosody*, vol. 7, pp. 673–677, 2014.

[15] 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 Systems*, vol. 90, pp. 1077–1087, 2018.

[16] L. Lugosch and V. Tomar, "Tone recognition using lifters and ctc," *Proceedings of Interspeech*, pp. 2305–2309, 2018.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," *Proceedings of O-COCOSDA*, 2017.

[18] L. Peng, W. Dai, D. Ke, and J. Zhang, "Multi-scale model for mandarin tone recognition," *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.

[19] J. Tang and M. Li, "End-to-end mandarin tone classification with short term context information," *ArXiv:2104.05657*, 2021.

[20] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *ArXiv:2010.11567*, 2020.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented data with recurrent neural networks," *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

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

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, K. Lukasz, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[25] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," *Proceedings of ICASSP 2021*, 2021.

[26] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *ArXiv:2101.00387*, 2021.