INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dual Audio Encoders Based Mandarin Prosodic Boundary Prediction by Using Multi-Granularity Prosodic Representations

Ruishan Li[1], *Yingming Gao*[2], *Yanlu Xie*[1], *Dengfeng Ke*[1], *Jinsong Zhang*[1]

[1]Beijing Language and Culture University, China
[2]Beijing University of Posts and Telecommunications, China

l3r1s7@163.com, yingming.gao@bupt.edu.cn, xieyanlu@blcu.edu.cn

Abstract

Prosodic boundary prediction plays an important role in speech synthesis, phonetic understanding, etc. In previous studies, supra-segmental features such as pitch, energy, and duration have been widely used to explicitly model Mandarin prosodic boundaries. In this paper, we propose to refine implicit prosodic representations with fine-grained information from complex acoustic features including mel-spectrogram and context vectors obtained from a pre-trained model. Pitch and energy are encoded as explicit prosodic representations. These two representations extracted by dual audio encoders are fused by the decoder mainly composed of cross-attention layers. Then the fused representations are used to predict Mandarin prosodic boundaries. The results indicate that our proposed method outperforms the baselines in the Mandarin prosodic boundary prediction task, particularly for the minor prosodic phrases (#2).

Index Terms: prosodic boundary prediction, dual audio encoders, prosodic representation, multi-granularity decoder

1. Introduction

Prosody essentially refers to the parts of speech which involve stress, rhythm, and intonation. It is related to supra-segmental elements like pitch, duration, and intensity. Speakers can express the meaning of the words more effectively and rationally with an effective and logical prosody structure of speech, and the listeners can also follow the intention of speakers more clearly. Prosodic segmentation produces prosodic boundaries, which are crucial for speech communication, syntactic disambiguation, and enhancing the naturalness and comprehensibility of Mandarin speech synthesis. Prosodic boundaries are ranked based on the level of dispersion between the prosodic units, such as prosodic word, prosodic phrase and intonational phrase [1][2]. The Mandarin prosodic boundary prediction is generally regarded as a sequence-to-sequence based classification task to predict whether there is a prosody break after each character of the utterance.

In previous studies, audio and text are two modalities commonly used to model Mandarin prosodic boundary detection. There are some studies of Mandarin prosodic boundaries using audio modality. Ni et al. proposed a hierarchical prosodic break classification method, which utilized the acoustic, lexical and syntactical features [3]. To improve the detection performance, Lin et al. suggested extracting tone nucleus-based supra-segmental features with DNN [4]. Subsequently, Lin et al. added phonological information to their previous research and used LSTM for the detection of syllable-level prosodic boundary. And they achieved 77.85% accuracy for the Mandarin corpus [5]. Lin et al. applied the joint detection of sentence stress as well as prosody phrase boundary using multi-granularity information of phoneme, syllable, and word. Then they obtained an experimental result of 0.91 F1 score with the Aix-MARSEC corpus [6]. In recent years, several studies have proposed similar approaches for predicting prosodic boundaries from texts, utilizing various techniques such as BLSTM-CRF, Attention, and BERT models [7-11]. Some other studies have found that incorporating linguistic information such as syntactic, lexical, and word embedding features can enhance the performance of the model in this task [12-15].

Although the above methods perform well in predicting Mandarin prosodic boundaries, the prediction of the confusable minor prosodic phrase boundary (#2) is still a challenging task. Prosody refers to variations in speech such as tone, pitch, syllable duration, and intensity, which are manifested explicitly in the audio signal. However, the implicit representations of prosodic information in audio signals have been rarely used to explore the task.

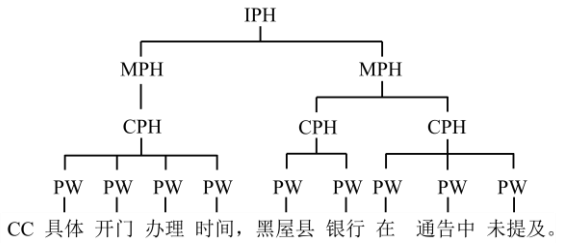

Figure 1: *The structure of prosodic boundary.*

In this paper, explicit prosodic representations, such as supra-segmental features mostly composed of pitch, energy, and duration are employed to predict prosodic boundaries. In the meanwhile, the mel-spectrogram and context vectors are projected to explore fine-grained implicit prosodic representations at the frame level. Fine-grained prosodic representations can help the model deal with the subtle variations of boundary clue. These two representations are fused by the multi-granularity decoder mainly composed of cross-attention layers. After fusion, the multi-granularity prosodic representations comprehensively determine the boundary from multiple prosodic subspaces, thus improving the robustness and generalization performance of the model. Furthermore, the structure of prosodic boundary that we employ is shown in Figure 1, CC represents characters within a prosodic word. PW (prosodic word) usually consists of 1-4 syllables. CPH (minor prosodic phrase) is generally perceived for a shorter period time. MPH (major prosodic phrase) is followed by a relatively long pause and resetting of F0. IPH (intonational phrase group) mainly consists of several major prosodic phrases with decreasing F0 contours [16-18]. In Section 2, with the aid

10.21437/Interspeech.2023-2242

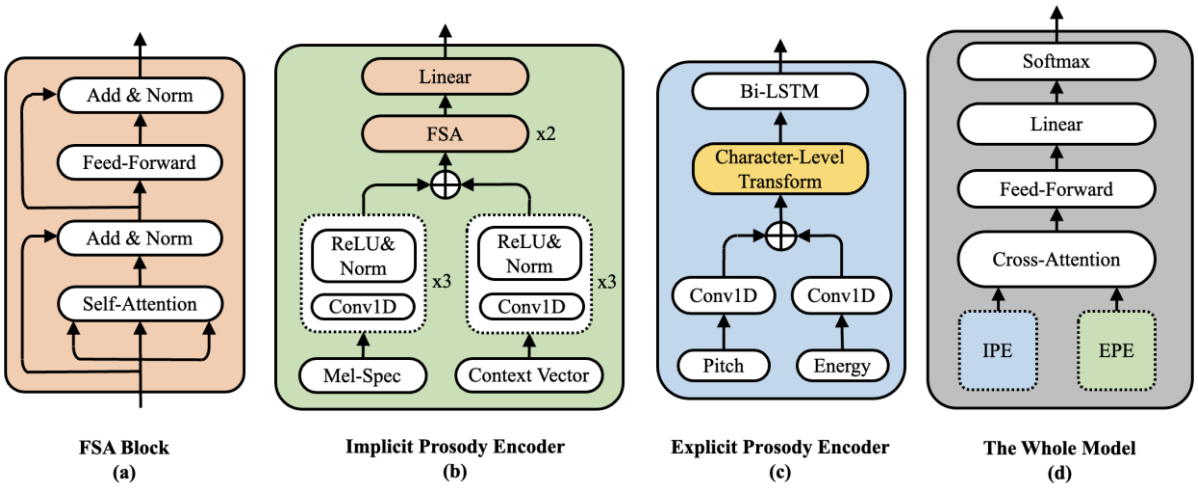

Figure 2: *The architecture of Mandarin prosodic boundary prediction model.*

of pertinent equations and figures, the architecture of Mandarin prosodic boundary prediction model we proposed will be explained. In Section 3, we introduce the corpus used in our work, contrastive experiments, ablation study, and other experimental configurations. The results are presented in Section 4 with detailed analyses. In Section 5, we draw conclusions and provide suggestions for the future work.

2. Proposed method

The architecture shown in Figure 2 mainly consists of FSA block, dual audio encoders and a decoder. The essence of prosody is the semantic information conveyed by the temporal variations of acoustic features in audio signals. Powerful audio encoders that can accurately extract prosodic information from the audio waveform are crucial to improve the accuracy of prediction. In this paper, EPE and IPE denote the explicit and implicit prosody encoder, respectively.

2.1. Explicit prosody encoder

As shown in Figure 2 (c), root-mean-square energy (RMS-Energy) and three-dimension pitch are fed into EPE as suprasegmental features, where pitch is extracted by the Kaldi toolkit and RMS-Energy is calculated by Equation 1.

$$\text{RMS-Energy} = \sqrt{\frac{1}{N}\sum_{n}|x(n)|^2} \tag{1}$$

The sequence of the audio signal is defined as x(n), where N represents the total number of samples in the signal. Firstly, the pitch and RMS-Energy features are processed by 1D-convolution layers, each followed by ReLU activation and layer normalization. The results of the above operation are concatenated to produce a transitional hidden matrix $\boldsymbol{T}$, with a size of (batch_size, out_channel, out_length). Next, to get the prosodic boundary labels for each character of the utterance, we use a trainable parameter matrix $\boldsymbol{W}$ to transform character-level conversions of the concatenated results. The size of matrix $\boldsymbol{W}$ is (batch_size, out_length, character_length). Then matrix $\boldsymbol{W}$ performs a multiplicative operation on matrix $\boldsymbol{T}$ through the matmul function. Then, the character-level hidden states are generated, with a size of (batch_size, out_channel, character_length). Finally, we obtain explicit prosodic representations after feeding the character-level hidden states into the Bidirectional Long Short-Term Memory (Bi-LSTM) module.

2.2. Implicit prosody encoder

As shown in Figure 2 (b), the encoder serves as the central component of the model, and tasked with generating an implicit representation that captures fine-grained prosodic information from the complex acoustic features. The mel-spectrogram is encoded through three 1D-convolution layers, each followed by the ReLU activation and layer normalization. Context vectors extracted by Wav2Vec 2.0 follow the same steps as the mel-spectrogram [19]. These two hidden states are concatenated to produce the transitional hidden sequences. As depicted in Figure 2 (a), the FSA block implements the multi-head self-attention and a feed-forward layer, each followed by a residual connection and layer normalization [20]. The transitional hidden sequences are fed into two FSA blocks to produce frame-level implicit representations with fine-grained prosody. Finally, a linear layer is utilized to ensure that implicit representations and explicit representations have matching hidden states dimensions.

2.3. Multi-granularity fusion decoder

A multi-granularity fusion decoder is needed to fuse the frame-level and character-level representations and then predict the prosodic boundary. The difficulty of fusing these two representations is that the frame-level representations are usually much longer than the character-level representations. In this paper, a cross-attention-based multi-granularity fusion decoder is implemented as a solution to this issue.

As depicted in the bottom half of Figure 2 (d), these two representations from dual audio encoders are fed into the stacked layers, consisting of multi-head cross-attention layers and a feed-forward layer. Since the total number of characters in the utterance sequence matches the length of the prosodic boundary prediction, we use the output of EPE as Query, and the output of IPE as Key and Value in cross-attention layers. The $\mathbf{C} = (C_1, C_2 \ldots C_N) \in R^{N\times D}$ and $\mathbf{F} = (F_1, F_2, \ldots F_M,) \in R^{M\times D}$ respectively denote Query and Key / Value. D is the dimension

Table 1: *Results of contrastive experiments.*

Model	PW (#1)		CPH (#2)		MPH (#3)		IPH (#4)	
	P	F1	P	F1	P	F1	P	F1
BLSTM-CRF	0.72	0.75	0.37	0.35	0.85	0.86	0.91	0.90
Self-Attention	0.82	0.83	0.46	0.47	0.89	0.89	0.95	0.96
Fine-tuned Bert	0.87	0.86	0.36	0.38	0.83	0.82	0.97	0.96
Fine-tuned w2vec	0.81	0.82	0.40	0.41	0.86	0.87	0.98	0.98
Fine-tuned HuBert	0.88	0.85	0.47	0.48	0.87	0.87	0.98	0.99
Proposed	0.90	0.89	0.59	0.60	0.93	0.92	0.98	0.99

of the two representations. The fusion of the explicit prosodic representations and implicit prosodic representations can be described as follows:

$$\mathbf{Q_C},\ \mathbf{K_F}, \mathbf{V_F} = \mathbf{W_Q C},\ \mathbf{W_K F},\ \mathbf{W_V F} \tag{2}$$

$$\mathbf{O} = \text{softmax}\left(\frac{\mathbf{Q_C}\ \mathbf{K_F^T}}{\sqrt{\mathbf{D}}}\right)\mathbf{V_F} \tag{3}$$

$\mathbf{W_Q}$, $\mathbf{W_K}$ and $\mathbf{W_V}$ are trainable matrices. $\mathbf{O} \in R^{N\times D}$ represents fused hidden output of the cross-attention layers. Afterwards, the output is passed through a post-processing module comprising a feed-forward neural network with two linear layers and the softmax function. This module produces a probability distribution of prosodic boundaries. Cross Entropy Loss (CE) is adopted as the train loss function of the proposed architecture.

3. Corpus and experiments

3.1. Joint Mandarin dataset

To conduct our experiments, we used the joint Mandarin dataset by merging the ASCCD (Annotated Speech Corpus of Chinese Discourse) Corpus and the Chinese Standard Girl Voice Corpus. The ASCCD Corpus consists of 4072 utterances recorded at the sampling rate of 16kHz, with a duration of 7 hours. These utterances were recorded by ten different speakers of Beijing Standard Mandarin. And the Chinese Standard Girl Voice Corpus is a public database containing around ten thousand utterances (≈ 12 hours) recorded by a female Mandarin speaker. Then, the corpus is resampled to 16kHz to ensure consistency with the ASCCD Corpus. The joint Mandarin dataset is randomly shuffled and subjected to a uniform data processing to ensure the fairness and robustness in our work. The dataset is split into the training set, development set, and test set at a ratio of 85%, 5%, and 10%, respectively, as presented in Table 2.

3.2. Experimental configurations

The experiments were conducted in the PyTorch framework, with the support of various tools such as Kaldi, Librosa, and Torchaudio. To evaluate the performance of our proposed model, we selected five baselines for comparison, as detailed in Section 3.3. The ablation study verifies the contributions of the IPE, as shown in Section 3.4. Furthermore, the experimental results are evaluated comprehensively by precision (P) and f1 score (F1) .

The IPE utilizes the WAV2VEC2_ASR_LARGE_960H to extract 768-dimensional context vectors as inputs. Due to the significant difference in dimensions of the input acoustic features between dual audio encoders, we use convolutional kernels of different sizes to adapt to each of them. Both multi-head self-attention and multi-head cross-attention all have six heads in our work. It not only improves efficiency but also facilitates the integration of multi-granularity representations across multiple prosodic subspaces. In this paper, every feed-forward layer is composed of two linear layers that are activated by the ReLU activation. The trainable matrix $\boldsymbol{W}$ is randomly initialized at the start of the train procedure in EPE.

During the training stage, the parameters are optimized using Adam optimizer with a learning rate of 0.0001. In addition, L2 regularization is introduced to reduce complexity and prevent overfitting. Four Tesla 3090 GPUs are used to train the model with batch size of 128.

Table 2: *Statistics of the joint Mandarin dataset.*

Annotation	Default CC	#1 PW	#2 CPH	#3 MPH	#4 IPH
Number	149773	56643	22945	17483	14072
Proportion	57.4%	21.7%	8.8%	6.7%	5.4%

3.3. Baselines

Previous studies have typically addressed the prosodic boundary detection task using two modalities: audio and text. LSTM-CRF [7], Self-attention [9], fine-tuned Bert [10], fine-tuned Wav2Vec 2.0 [19][21] and Hubert [22] are designated as the baselines, and these baselines have been replicated using joint Mandarin dataset. Moreover, HuBert gets the same fine-tuned configurations as Wav2Vec 2.0. We have maintained modal and structures of these five baselines to closely resemble the methods used in the relevant references. In Section 4, the results of baselines are contrasted with those of the proposed model.

3.4. Ablation study

We demonstrate the effectiveness of our proposed network with two ablation experiments: (1) The EPE can project implicit representation with fine-grained prosodic information from the complex acoustic features. (2) The IPE plays an essential role and makes valuable contributions to our work. For experiment (1), the prediction task of prosodic boundaries is carried out independently by the IPE and decoder. For experiment (2), it has the same experimental structure as experiment (1), except the IPE. EPE and decoder are used to predict the prosodic boundaries. Specifically, the cross-attention layers in the multi-grain fusion decoder are changed to the self-attention layers due to the absence of the other audio encoder. All relevant analyses are described in Section 4.

Table 3: *Results of ablation study.*

Model	PW(#1)		CPH(#2)		MPH(#3)		IPH(#4)	
	P	F1	P	F1	P	F1	P	F1
Only-IPE	0.81	0.79	0.47	0.48	0.87	0.85	0.95	0.96
Only-EPE	0.80	0.81	0.41	0.39	0.85	0.84	0.93	0.95
Proposed	0.90	0.89	0.59	0.60	0.93	0.92	0.98	0.99

4. Results

In this paper, we evaluate the performance of our work in terms of the accuracy of prosodic boundary prediction on the test set.The results of contrastive experiments are shown in Table 1. Among the four prosodic boundaries, the prediction of #2 is the least satisfactory. The #2 (minor prosodic phrase) is inherently difficult to predict accurately, for #2 and #1(prosodic word) have similar acoustic structure [3]. There is a relatively short perceptual break after #2, which is longer than #1 but shorter than #3 (major prosodic phrase) . Thus #2 may be misclassified as #1 or #3. As compared to baselines, our proposed method predicts #2 better. The output of LPE contains fine-grained prosodic information that helps the explicit prosodic representations fill in the fine-grained prosody spaces and participate in determining the variations of prosodic clues between the #2 boundary and the #1, #3 boundaries. Furthermore, the results of ablation study could also confirm our opinion on this point.

For the prediction of #4 boundary, all models have the relatively good performance. We fine-tune pre-trained models Bert, HuBert and Wav2Vec 2.0, which are representative in audio and text modality. Based on the results, the fine-tuned Hubert performs the best for the downstream task of prosodic boundary prediction. And the reason why HuBert model do the best is that fine-tuned HuBert uses both speech and text features as joint inputs for training, while the Wav2Vec 2.0 and Bert use only audio signals or text as inputs. The multi-modal pre-trained model outperforms the unimodal pre-trained model in our task. Non-pre-trained models such as BLSTM-CRF and Self-Attention, have excellent performances among the training set but weak generalization ability for the test set.

As shown in Table 3, we have found from the results that Only-IPE and Only-EPE have almost equal f1 scores except for the #2 boundary. The Only-EPE performs 9% inferior to the Only-IPE for the prediction of #2 boundary. The implicit representations support the EPE to alleviate confusions to some extent between #2 and #1, #3 boundaries. The results largely confirms that the EPE can project fine-grained prosodic representations.

The results of Only-EPE and proposed model also reveal that the addition of IPE improves the predictions for all prosodic boundaries. For the #2 boundary, the performance of our model upgrades about 20% in f1 score. The proposed model demonstrates a robust capability in tracking all prosodic clues by incorporating IPE, which actually has strong contributions to the proposed model.

5. Conclusion

In this paper, we propose to utilize two acoustic encoders and a multi-granularity decoder to predict Mandarin prosodic boundaries. The dual acoustic encoders capture the variations of prosodic clues and project them from different acoustic features. The outputs of dual acoustic encoders are fused at multiple granularities by the decoder. Our proposed method obtains the best results in the experiments and confirms the contribution of implicit prosodic representations for the model. In the future, we will try to explore Mandarin prosodic boundaries from a multi-modal approach, with a particular focus on #2 (minor prosodic phrase). The task should be beneficial for prosody bias detection and speech synthesis.

6. Acknowledgements

This research project is supported by Key Research Project on International Chinese Language Education (22YH49B) and Science Foundation of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (22YJ080006). Also, this research project is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (23YCX122) , and the Fundamental Research Funds for the Central Universities (Grant Number: 2023RC13).

The corresponding author of the paper is Yanlu Xie.

7. References

[1] F. Liu and M. Chen, “A Study of Chinese Prosodic Boundary Characteristics of Chinese as Second Language Learners,”. *TCSOL Studies*, no. 3, pp. 1–16+86, Sep. 2016.

[2] D. Deng, *Experimental Study of Chinese Prosodic Word*, Peking University Press, 2010.

[3] C. Ni, et al. “Classification of Mandarin prosodic break based on hierarchical structure of prosodic break,” *Application Research of Computers*, vol. 28, no. 7, pp. 2452–2454+2511, Jul. 2011

[4] J. Lin, et al, “Automatic Mandarin prosody boundary detecting based on tone nucleus and DNN model,” *Journal of Chinese Information Procession*, vol. 30, no. 6, pp. 35-39, Nov. 2016.

[5] J. Lin, et al, “Improving Mandarin prosody boundary detection by using phonetic information and deep lstm model,” in *2019 International Conference on Asian Language Processing (IALP)*, Shanghai, China, Nov. 2019, pp. 504–508.

[6] B. Lin, et al. “Joint Detection of Sentence Stress and Phrase Boundary for Prosody,” in *INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 4392–4396.

[7] Y. Zheng, et al. “BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End,” in *INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 47–51.

[8] Y. Du, et al. “Automatic Prosodic Structure Labeling using DNN-BGRU-CRF Hybrid Neural Network,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 1234–1238.

[9] C. Liu, P. Zhang and Y. Hong, “Self-attention Based Prosodic Boundary Prediction for Chinese Speech Synthesis,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May. 2019, pp. 7035–7039.

[10] Q. Wang, et al. “Predicting the Chinese poetry prosodic based on a developed BERT model,” in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Nanchang, China, Mar. 2021, pp. 583–586.

[11] Z. Bai and B. Hu, "A Universal Bert-Based Front-End Model for Mandarin Text-To-Speech Synthesis," in *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6074–6078.

[12] Y. Zheng, et al. "Improving Prosodic Boundaries Prediction for Mandarin Speech Synthesis by Using Enhanced Embedding Feature and Model Fusion Approach," in *INTERSPEECH*, San Francisco, USA, Sep. 2016, pp. 3201–3205.

[13] Z. Chen, G. Hu and W. Jiang, "Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction," in *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, Sep. 2010, pp.1421–1424.

[14] C. Hao, J. Tao and Y. Li, "Improving mandarin prosodic boundary prediction with rich syntactic features," in *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, Sep. 2014, pp. 46–50.

[15] Z. Zhang, et al. "Mandarin Prosodic Phrase Prediction based on Syntactic Trees," in *9th ISCA Speech Synthesis Workshop*, Sunnyvale, USA, Sep. 2016, pp. 160–165.

[16] W. Hu, B. Xu and T. Huang, "An Experimental Study on Prosodic Boundary in Chinese Mandarin," *Journal of Chinese information processing*, vol. 16, no. 1, pp. 43–48, Jan. 2001.

[17] Z. Xiong, "An Acoustic Study of the Boundary Features of Prosodic Units," *Applied Linguistics*, vol. 2, no. 2, pp. 116–121, Feb. 2003.

[18] Z. Yin, "Revisiting the Acoustic Characteristics and the Generation Mechanism of Prosodic Boundary," *Chinese Journal of Phonetics*, vol. 13, no. 1, pp. 38-50, Jun. 2020.

[19] A. Baevski, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, no. 1044, pp. 12449–12460, Dec. 2020.

[20] A. Vaswani, et al. "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[21] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using Wav2Vec 2.0," in *Text, Speech, and Dialogue: 25th International Conference*, Brno, Czech Republic, Sep. 2022, pp. 377–388.

[22] W. Hsu, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.29, pp. 3451–3460, Oct. 2021.