INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Joint detection of sentence stress and phrase boundary for prosody

Binghuai Lin[1], *Liyuan Wang*[1], *Xiaoli Feng*[2], *Jinsong Zhang*[2]

[1]Smart Platform Product Department, Tencent Technology Co., Ltd, China
[2]College of Information Science, Beijing Language and Culture University, China

{binghuailin, sumerlywang}@tencent.com,
fengxiaoli314@163.com, jinsong.zhang@blcu.edu.cn

Abstract

Prosodic event detection plays an important role in spoken language processing tasks and Computer-Assisted Pronunciation Training (CAPT) systems [1]. Traditional methods for the detection of sentence stress and phrase boundaries rely on machine learning methods that model limited contextual information and account little for interaction between these two prosodic events. In this paper, we propose a hierarchical network modeling the contextual factors at the granularity of phoneme, syllable and word based on bidirectional Long Short-Term Memory (BLSTM). Moreover, to account for the inherent connection between sentence stress and phrase boundaries, we perform a joint modeling of these two important prosodic events with a multitask learning framework (MTL) which shares common prosodic features. We evaluate the network performance based on Aix-Machine Readable Spoken English Corpus (Aix-MARSEC). Experimental results show our proposed method obtains the F1-measure of 90% for sentence stress detection and 91% for phrase boundary detection, which outperforms the baseline utilizing conditional random field (CRF) by about 4% and 9% respectively.

Index Terms: Prosodic event detection, contextual information, hierarchical network, MTL, BLSTM

1. Introduction

Prosody plays an important role in many spoken language processing tasks such as automatic speech recognition (ASR), speech synthesis and dialect identification [2]. It is also helpful in determining the pronunciation proficiency of the language learners [3]. Corpora annotated with prosodic information will be beneficial for these spoken language applications. An automatic prosodic labeling system can greatly simplify the task of annotation.

The main prosodic events we are concerned about in this paper are sentence stress and phrase boundary. Generally, phrasing can combine words into prosodic units by intonational patterns [4]. Sentence stress can form a certain natural stress pattern characteristic for a given language and give emphasis on particular words based on their relative importance. Sentence stress is different from pitch accent that carries pitch prominence caused by an intonation event as well as rhythmic prominence caused by sentence stress [5].

Due to the suprasegmental nature of prosody, it has impact on the acoustics of speech over a much longer context such as the word level or phrasal level than the frame level or phoneme level [6, 7]. Many approaches have demonstrated the benefits of adding contextual information in prosody detection. Previous work combined features of its left neighbors and right neighbors to form contextual windows [2, 8, 9]. These methods can only model fixed neighboring context, and thus later many machine learning models are implemented to deliver context information dynamically, such as hidden Markov model (HMM) or CRF [2, 10, 11, 12]. Although models such as CRF can capture relations between sequential prosodic labels, they still need to find useful contextual features by manual feature engineering. To capture contextual information automatically, many neural network architectures have been explored. Convolutional neural network (CNN) was employed in the prosodic event recognition to learn contextual influence in a fixed context window [13]. To encode prosodic features in a larger context, recurrent neural network (RNN) have been applied. A special structure of RNN called LSTM have been applied in pitch accent detection [14]. A particular method based on BLSTM was proposed in the detection of prosody to further capture bidirectional contextual information and achieve better performance than the baseline of the previous work. However, these approaches mainly focus on the broad range contextual influence (i.e., the word and the phrasal level) and account little for the local contextual influence (i.e., the phoneme and the syllable level). The local syllabic effects and broad phrasal contextual influence have been investigated in the task of pitch accent detection and the results showed the improved performance by combining both of them [15]. To investigate the contextual importance over the broad and local ranges, the work [16] examined the performance of pitch accent detection at the phoneme, syllable and word level, respectively, and showed that incorporating information from surrounding context can improve performance at all levels. Inspired by these studies, we combine both the local and broad range of the contextual information at the granularity of phoneme, syllable and word by the BLSTM for the detection of sentence stress and phrase boundaries.

As there are many acoustic cues in prosody, the inherent connection of prosodic events has been investigated in recent studies. The work [17] explored the effects of phrase boundary on sentence pitch accent. The interaction of boundary tone and prominence has also been surveyed [18]. With the development of machine learning, a learning paradigm called MTL is considered to leverage useful information contained in multiple related tasks [19]. Following these studies, we propose a hierarchical network based on the BLSTM which models the contextual information at multi-granularity of phoneme, syllable and word. To further explore the connection of prosodic events, we combine the sentence stress and phrase boundary detection based on an MTL learning framework which shares common acoustic features between them. We will introduce the proposed method in section 2. In section 3, the corpus used in the proposed model is introduced and the hyperparameters of the proposed network are explained. The results are shown and some discussion is made in section 4. We will draw the conclusions and future suggestions in section 5.

http://dx.doi.org/10.21437/Interspeech.2020-1284

2. Proposed method

2.1. Features for prosodic event detection

We investigate the acoustic features extracted from speech signal for prosodic event detection. To model contextual information from different granularity, the phonemes and words in a sentence are first forced-aligned by a Kaldi-based automatic speech recognition (ASR) system [20]. We can obtain the beginning and ending times of each phoneme or word, and then compute the time interval of one particular syllable based on the belonging phoneme sequence of the syllable. Acoustic cues such as fundamental frequency (F0), intensity and duration are regarded as the important measurable parameters for prosody [21]. We also use Kaldi to extract these features. The features of F0 and intensity are computed for each 10ms frame with a 10ms shift. The phoneme and word durations are used to aggregate features of all frames for each phoneme or word into one input feature matrix.

Using aforementioned features, we calculate a number of aggregations to represent the prosodic features for each phoneme, syllable and word. These features include the maximum, minimum, mean, standard deviation of the prosodic features. To account for speaker differences, we normalize the maximum, minimum of these features within the context of a particular phoneme, syllable and word using the z-score normalization [22].

In addition to the above aggregation of features within the context at multi-granularity, the perceptual features related to sentence stress and phrasing are also included. Sentence phrasing is frequently indicated by the presence of silence following the word and the reset of intensity and F0 [2]. We first normalize duration by speech rate, that is the number of phonemes per second, and extract features such as silence duration and the ratio of voiced and silence durations at the phoneme, syllable and word level. Then we compute changes of F0 and intensity between adjacent syllables and words. To take into account that English is a stress-timed language [23], we extract time features mainly consisting of normalized durations of phonemes, syllables and words. These features at multi-granularity are treated as input for the following proposed hierarchical network.

2.2. Modeling method

The hierarchical network is shown in Figure 1. It is composed of three layers : phoneme, syllable and word layer. To capture contextual information at multi-granularity, each layer is constructed by one BLSTM. The BLSTM is shown in Figure 2.

To carry acoustic information of independent phonemes, we employ an independent numerical representation for each phoneme which is called phoneme embedding [24]. The input features of phoneme layer are phoneme embeddings combined with other phoneme level features proposed in the previous section. The final forward and backward outputs of the phoneme BLSTM are concatenated as the input of syllable layer. Similarly, the aforementioned syllable level acoustic features are combined as the input of syllable BLSTM layer. The outputs from the syllable layer and other word level acoustic features are fed into the word BLSTM layer. As the duration between stressed syllables is almost equal in English [23], we conduct the detection of sentence stress from syllable level. Two separate fully connected (FC) layers followed by the sigmoid activation function are applied over syllable and word representations to determine the sentence stress and phrase boundary jointly.

As MTL shares feature representations and facilitates the generalization performance of related tasks [19], the network is optimized by an MTL framework which combines detection of sentence stress and phrase boundaries. Specifically,

$$L_{\text{total}} = (1 - w) \times L_{\text{stress}} + w \times L_{\text{phrasing}} \quad (1)$$

where L_{stress} and L_{phrasing} are the classification losses of sentence stress and phrase boundaries. w is a constant value balancing the weight between two tasks. The classification loss can be defined as Eq. (2),

$$L = -y \times \log(p) - (1 - y) \times \log(1 - p) \quad (2)$$

where y is the ground truth of sentence stress or phrase boundaries and p is the probability derived from the syllable FC layer or the word FC layer of the proposed model.

3. Experimental setup

3.1. Corpus description

We mainly focus on two aspects of prosody: sentence stress and phrase boundary. The standard approach for prosody annotation is based on Tone and Break Indices (ToBI) [25]. As ToBI focuses on pitch accent which is not totally equal to sentence stress, we use the Aix-MARSEC (Aix-Machine Readable Spoken English Corpus) database [26] as in [5]. Aix-MARSEC consists of over 5 hours of BBC radio recordings from 53 different speakers in 11 different speech styles from the 1980s. The corpus includes approximately $55,000$ orthographically transcribed words.

We uses the original phrase break annotations for minor and major boundaries which are equivalent with the break indices 3 and 4 in ToBI [27]. We follow the previous work which treated the syllable to be stressed when first appearing in each Jassems narrow rhythm unit (NRU) notation [5]. For practical purposes, we merge minor and major boundaries into break labels. The labeling of our prosody in this study can be summarized in Table 1.

Table 1: *Modified prosody label*

Label	0	1
Stress	Unstressed	Stressed
Phrasing	No break	Break

The statistics of data used in our experiments is shown in Table 2. An utterance consists of multiple sentences. We use 70% of the dataset for training and 30% for testing.

Table 2: *The statistics of data*

Class	Number
Utterances	408
Sentences	3790
Words	51650
Syllables	90163
Phonemes	107540

3.2. Hyperparameters

The proposed network is composed of three BLSTMs representing the phoneme, word and sentence layers. The input of first

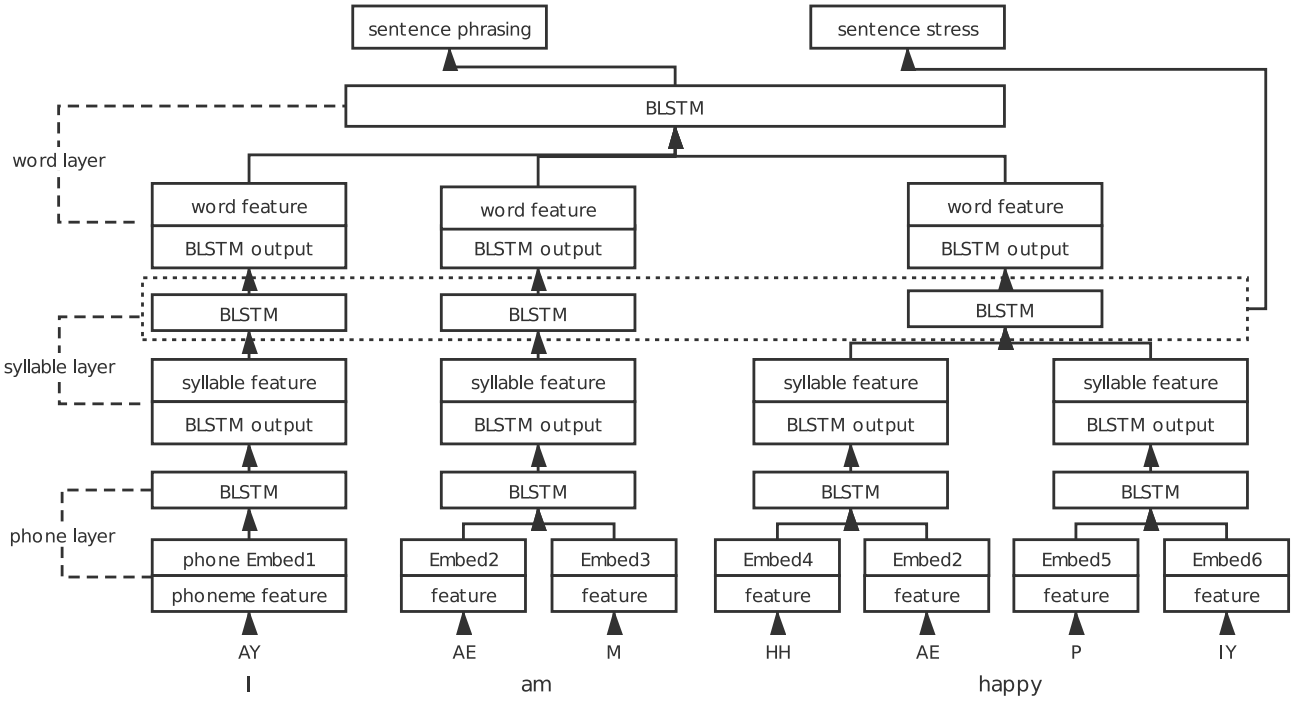

Figure 1: *Sentence stress model structure*

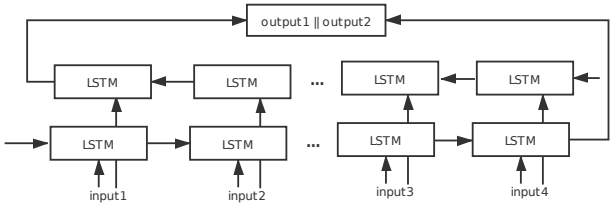

Figure 2: *BLSTM structure*

layer consists of the phoneme embedding and other phoneme features proposed in section 2. The phoneme embedding is composed of 39 different phonemes with feature dimensionality of 15 based on Carnegie Mellon University (CMU) Pronouncing Dictionary [28]. The final shape of the phoneme embedding is 39×15. The hidden sizes of three BLSTMs at phoneme, syllable and word level are 20, 15 and 20 respectively. The input feature dimensionality of the three BLSTMs are 26 (15 for phoneme embeddings and 11 for phoneme acoustic features), 55 (40 from the phoneme layer and 15 from syllable acoustic features) and 45 (30 from the syllable layer and 15 from word syllable features). The dimensionalities of the fully connected layer of the word and syllable layers are 55×2 and 45×2, where 2 indicates a binary classification for sentence stress or phrase boundaries. The model are trained with the adaptive Adam optimizer [29].

4. Results

We evaluate the performance of prosodic event detection by confusion matrix and F1-measure. First, we will show the performance of the proposed method. Then we will illustrate some ablation study to validate the rationality of our proposed methods.

4.1. Performance of prosody detection

The confusion matrix of sentence stress based detection on our proposed network is shown in Table 3 and the confusion matrix of phrase boundary detection is shown in Table 4. It shows the number of syllables or words labeled as stressed (break) or unstressed (no break) in the Aix-MARSEC testing dataset and the number of syllables or words predicted as stressed (break) or unstressed (no break) with our detection models. The accuracy of sentence stress detection is 88% and the accuracy of phrase boundary detection is 95%.

Table 3: *Confusion matrix of sentence stress detection*

labeled \ predicted	Unstressed	Stressed
Unstressed	9700	1916
Stressed	1289	14144

Table 4: *Confusion matrix of phrase boundary detection*

labeled	No break	Break
No break	11240	506
Break	203	3546

We adopt the CRF model, which has been commonly used in prosodic event detection [12, 5, 30] as a fairly strong baseline in our experiment. The CRF model operates on the similar feature as ours. It takes word level features for the detection of sentence phrasing and syllable level features for the detection of sentence stress. As CRF can't merge sequential phonemes features in its model structure, we aggregate features of a phoneme sequence by calculating the maximum, minimum

and mean value of phoneme duration in a word or syllable and the number of phonemes in a word or syllable. To model the contextual information, we explore features from the neighboring words or syllables. Specifically, we include acoustic features of the preceding N words or syllables and the following N words or syllables, where N can be adjusted in the experiment. Two CRF models are created for the detection of sentence stress and phrase boundaries independently. We compare our results with the CRF models using different N. For sentence stress, we also compare results with previous work [5] based on CRF using the same corpus as ours. The results are shown in Table 5.

Table 5: *F1-measure of prosody detection*

	Stress			Phrasing		
	P	R	F1	P	R	F1
CRF(N=1)	0.78	0.86	0.82	0.81	0.80	0.80
CRF(N=2)	0.86	0.87	0.86	0.81	0.83	0.82
CRF(N=4)	0.84	0.86	0.85	0.79	0.83	0.81
CRF [5]	0.85	0.89	0.87	-	-	-
Ours	0.88	0.92	0.90	0.88	0.95	0.91

Our proposed model achieves superior performance by 4% in sentence stress detection and 9% in phrase boundary detection based on F1-measure. From the results, we can see that different window sizes of context influence the performance of CRF. The best performance of CRF is based on window context size of 2 rather than 4. Previous work using CRF achieved the similar results. While the CRF model captures limited static contextual information, our model can capture the contextual influence dynamically.

4.2. Ablation study

We demonstrate the effectiveness of our proposed network from two aspects: (1) MTL combining detection of sentence stress and phrase boundary; (2) the contextual influence at multi-granularity of phoneme, syllable and word.

4.2.1. Effect of MTL

To demonstrate the effect of MTL, the detection tasks of sentence stress and phrasing are carried out independently. Specifically, the sentence stress detection task is based on our proposed model with only the phoneme layer and syllable layer (STL) and the sentence phrasing task is based on our proposed model with the sentence stress detection task removed (STL). The results are shown in Table 6.

Table 6: *Comparison of STL and MTL*

	Stress			Phrasing		
	P	R	F1	P	R	F1
STL	0.87	0.89	0.88	0.86	0.92	0.89
MTL	0.88	0.92	0.90	0.88	0.95	0.91

The results from the table show that the performance of the two single-learning tasks is around 2% inferior to our proposed MTL learning method. It indicates that our proposed method based on the MTL framework can further improve the performance of these two detection tasks by common prosodic features sharing and joint optimizing.

4.2.2. Effect of the multi-granular contextual influence

To demonstrate the effect of contextual influence at multi-granularity, we illustrate the performance of the sentence phrase boundary detection without the phoneme layer (No phoneme) and without both the phoneme and syllable layers (With word), respectively, and the sentence stress detection with the phoneme layer removed (No phoneme). We compare these methods with our proposed model in a single-task learning manner (STL). The results are shown in Table 7.

Table 7: *F1-measure of prosody detection*

	Stress			Phrasing		
	P	R	F1	P	R	F1
No phoneme	0.84	0.87	0.85	0.86	0.90	0.88
With word	-	-	-	0.85	0.90	0.87
STL	0.87	0.89	0.88	0.86	0.92	0.89

The results reveal that the proposed method with only word layer left performs 2% inferior to the full model for the detection of phrase boundaries. The performance of the model with the phoneme layers removed performs 1% inferior to our proposed model. It could indicate that sentence phrase boundary is more correlated with contextual information on the syllable and word level than on the phoneme level. For sentence stress detection, the performance of our network degrades nearly 3% in F1-measure when contextual information of phoneme is removed, which shows the phoneme contextual influence on the sentence stress.

5. Conclusions

In this paper, we propose an automatic prosody detection method for sentence stress and phrase boundaries. A hierarchical network based on BLSTM is developed to dynamically capture contextual information at multi-granularity of phoneme, syllable and word. To further model the connection of sentence stress and phrase boundaries, we implement two detection tasks jointly by an MTL learning framework to share common features at the phoneme and syllable levels. The experiment results on Aix-MARSEC show our prosody model can achieve the F1-measure of 90% in sentence stress detection and 91% in phrase boundary detection which is superior than the baseline CRF. Currently, our proposed method mainly focuses on sentence stress and phrase boundary detection, we will extend this mechanism to other aspects of prosody detection in the future, such as boundary tone. We will also explore detection methods with less feature engineering.

6. References

[1] J. H. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4565–4568.

[2] A. Rosenberg, *Automatic detection and classification of prosodic events*. Columbia University, 2009.

[3] K. Li, S. Zhang, M. Li, W.-K. Lo, and H. Meng, "Detection of intonation in l2 english speech of native mandarin learners," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 69–74.

[4] D. Bolinger and D. L. M. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. Stanford university press, 1989.

[5] G. G. Lee, H.-Y. Lee, J. Song, B. Kim, S. Kang, J. Lee, and H. Hwang, "Automatic sentence stress feedback for non-native English learners," *Computer Speech & Language*, vol. 41, pp. 29–42, 2017.

[6] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] J. Zhao, W.-Q. Zhang, H. Yuan, M. T. Johnson, J. Liu, and S. Xia, "Exploiting contextual information for prosodic event detection using auto-context," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 30, 2013.

[8] A. Schweitzer and B. Möbius, "Experiments on automatic prosodic labeling," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[9] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks," *Speech Communication*, vol. 96, pp. 28–36, 2018.

[10] R. Fernandez and B. Ramabhadran, "Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] M. Fach and W. Wokurek, "Pitch accent classification of fundamental frequency contours by hidden markov models," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[12] M. L. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 677.

[13] S. Stehwien and N. T. Vu, "Prosodic event recognition using convolutional neural networks with context information," *arXiv preprint arXiv:1706.00741*, 2017.

[14] Y. Wu, S. Li, and H. Li, "Automatic pitch accent detection using long short-term memory neural networks," in *Proceedings of the 2019 International Symposium on Signal Processing Systems*, 2019, pp. 41–45.

[15] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] A. Rosenberg and J. B. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," 2009.

[17] Y.-L. Shue, S. Shattuck-Hufnagel, M. Iseli, S.-A. Jun, N. Veilleux, and A. Alwan, "Effects of intonational phrase boundaries on pitch-accented syllables in american english," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[18] A. Katsika, J. Krivokapić, C. Mooshammer, M. Tiede, and L. Goldstein, "The coordination of boundary tones and its interaction with prominence," *Journal of Phonetics*, vol. 44, pp. 62–82, 2014.

[19] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[21] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech communication*, vol. 50, no. 10, pp. 782–796, 2008.

[22] D. Zill, W. S. Wright, and M. R. Cullen, *Advanced engineering mathematics*. Jones & Bartlett Learning, 2011.

[23] T. Mitchell, "David abercrombie, elements of general phonetics. edinburgh: Edinburgh university press, 1966. pp. 203." *Journal of Linguistics*, vol. 5, no. 1, pp. 153–164, 1969.

[24] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, and L. Cai, "Phoneme embedding and its application to speech driven talking avatar synthesis." in *INTERSPEECH*, 2016, pp. 1472–1476.

[25] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," *The OSU Research Foundation*, vol. 3, p. 30, 1997.

[26] C. Auran, C. Auran, C. Bouzon, C. De Looze, D. Hirst *et al.*, "Aix-marsec database," 2008.

[27] C. Brierley and E. Atwell, "An approach for detecting prosodic phrase boundaries in spoken english," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 14, no. 1, pp. 1–11, 2007.

[28] R. L. Weide, "The cmu pronouncing dictionary," *URL: http://www.speech.cs.cmu.edu/cgibin/cmudict*, 1998.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 135–138.