DNN based phrase boundary detection using knowledge-based features and feature representations from CNN

1st Pavan Kumar J
Electrical Engineering Department
Indian Institute of Science
Bangalore, India
pavanj@iisc.ac.in

2nd Chiranjeevi Yarra
Language Technologies Research Center (LTRC)
International Institute of Information Technology
Hyderabad, India
chiranjeevi.yarra@iiit.ac.in

3rd Prasanta Kumar Ghosh
Electrical Engineering Department
Indian Institute of Science
Bangalore, India
prasantg@iisc.ac.in

***Abstract*—Automatic phrase boundary detection could be useful in applications, including computer-assisted pronunciation tutoring, spoken language understanding, and automatic speech recognition. In this work, we consider the problem of phrase boundary detection on English utterances spoken by native American speakers. Most of the existing works on boundary detection use either knowledge-based features or representations learnt from a convolutional neural network (CNN) based architecture, considering word segments. However, we hypothesize that combining knowledge-based features and learned representations could improve the boundary detection task's performance. For this, we consider a fusion-based model considering deep neural network (DNN) and CNN, where CNNs are used for learning representations and DNN is used to combine knowledge-based features and learned representations. Further, unlike existing data-driven methods, we consider two CNNs for learning representation, one for word segments and another for word-final syllable segments. Experiments on Boston University radio news and Switchboard corpora show the benefit of the proposed fusion-based approach compared to a baseline using knowledge-based features only and another baseline using feature representations from CNN only.**

***Index Terms*—Boundary detection, human-computer interaction, computer-assisted pronunciation tutoring, CNN based representation learning**

I. Introduction

Automatic phrasal boundary detection involves identifying two types of prosodic boundaries that constitute the two highest perceived strength of disjuncture between the words [1], [2]. These two types of boundaries have been known as intermediate, and intonation phrase boundaries [1]. Typically, in a spoken utterance, a rating representing perceived strength of disjuncture between a pair of words is represented using break-index [1], [2]. According to the tone break-index (ToBI) [3] system, the values of the break-index range from 0 to 4, where 0 and 4 represent the least and the highest disjuncture, respectively. The break indices 3 and 4 correspond to intermediate and intonation phrase boundaries. These two types of boundaries majorly describe the quality of fluency in native English speakers, often used in computer-assisted pronunciation tutoring (CAPT) [4], [5]. Also, the automatic phrase boundary detection could be useful in the applications of spoken language understanding [6], [7], [8] and automatic speech recognition [9], [10], [11].

In almost all the existing works, the automatic phrase boundary detection problem has been formulated as a binary classification task by assuming the phrase boundary occurs only at the end of word segments in a given test utterance [12], [13], [14], [15], [16]. Considering this, features have been derived for each word in the utterance, and the classifier has been trained considering word associated boundary labels indicating the occurrence of phrase boundary at the word's end [15], [16], [14]. Classifiers in most of the existing works have considered that word ends associated with intermediate and intonation phrase boundaries, i.e., break indices 3 and 4 are labelled as class-1. The word ends correspond to the remaining break indices as class-0 [17], [18], [16], [19]. Following these works, we, in this work, use the same labelling procedure for formulating phrase boundary detection.

Many existing works on phrase boundary detection have used acoustic features derived from prosodic variations embedded in frame-level frequency and energy values [12], [20], [14], [13]. These features are heuristically computed applying statistics on frequency and energy values within the segments, including syllables and words, referred to as segment level knowledge-based features. Considering these features, boundary detection has been addressed using different modeling techniques include decision trees [12], [13], hidden Markov model (HMM) [20], [14], Gaussian mixture model (GMM) [20], neural network [15] conditional random field [16] and support vector machine (SVM) [21], [19]. In addition to the acoustic features, few works have considered syntactic and lexical based information while modelling these features [14], [21], [17]. However, both the information in those works are obtained from canonical pronunciation, which the learners might not follow. Thus, the models constructed using both the information may not be applicable for all learners in computer-assisted pronunciation training. In this work, we perform boundary detection using only frame-level energy and

978-1-6654-0522-5/21/$31.00 ©2021 IEEE

2021 National Conference on Communications (NCC) | 978-1-6654-4177-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/NCC52529.2021.9530147

978-1-6654-4177-3/21/$31.00 ©2021 IEEE

frequency values, referred to as frame-level prosody features, without considering both the information. Further, we consider the features derived from first-order (Δ) and second-order ($\Delta\Delta$) differences on frame-level prosody features, which are found to capture boundary specific prosodic cues much better.

In contrast to the boundary detection methods involving heuristically computed features at the segment level, few other works have directly used frame-level features within the segments [15], [18]. These works have considered convolutional neural networks (CNNs) for learning high-level representations from frame-level features. After this, boundaries are detected using these representations. In these works, frame-level features include frame-level prosody features and Mel frequency cepstral coefficients (MFCCs). These methods based on representations learning have been shown to be effective compared to the knowledge-based segment level features. However, we believe that these representations might not capture all segmental specific information containing boundary detection cues. One such example could be syllable segment duration, which varies due to pre-boundary lengthening property in the segments (syllable or words). As per this property, the segments at the boundary have a more considerable duration than the respective segments at the other locations in an utterance [22].

Unlike the existing methods, we propose a model to facilitate both learned representations and segment level features obtained from both types of segments – syllable and words. For this, a combined deep learning architecture comprising CNN and DNN is considered. In this, CNNs are used for representation learning which takes frame-level prosody features and its Δ and $\Delta\Delta$ variations within both syllable and word segments. DNN is considered for fusing the representations learnt from CNNs and segmental level features from both syllable and word segments. Experiments performed on Boston University radio news (BURN) [17] and Switchboard corpus [23] containing intonation and intermediate phrase boundaries show that the proposed method is more effective compared to the top two best performing existing methods one [17] involve only segment level features and the other [18] uses only CNN based model.

II. Database

In our work, we use BURN [24] and Switchboard [23] corpora. Both the corpora have been used for most of the existing works on boundary detection. The BURN corpus contains recordings with a sampling frequency of 16kHz of broadcast news from American native English speakers (three female and four male speakers). The data also contains the word and phoneme transcriptions along with their respective time-aligned boundaries obtained from a forced-alignment process. From the phoneme transcriptions and the respective time-alignments, we obtain syllable transcriptions and their time-aligned boundaries using P2TK syllabifier [25]. In addition to these, the corpus contains ToBI style boundary (break index) annotations for a subset containing three male and three female speakers. The recordings' total duration has the ToBI annotations is 2hr 48min, and the total number of words in those recordings is 28862.

Switchboard corpus contains spontaneous speech recordings collected from telephonic conversations recorded with a sampling frequency of 8kHz. This data was collected from American native English speakers. In this data, syllable transcriptions and their time-aligned boundaries are available, along with word transcriptions and their boundaries. A subset was provided with ToBI style break index annotations in the entire data, resulting in 8hr and 10 min of recordings.

III. Proposed model

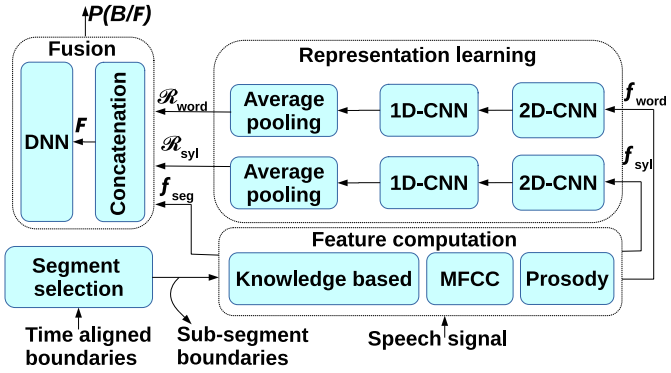

Fig. 1. Block diagram of the proposed model for phrase boundary detection. In the figure f_{word}, f_{syl} represent combined frame-level prosody and MFCC feature of the word and word-final syllable segment, f_{seg} represent Segment level knowledge of prosody based feature and $\mathcal{R}_{word}, \mathcal{R}_{syl}$ represent high-level feature representation of the word and word-final syllable segment

Block diagram in Figure 1 shows four steps involved in the proposed approach. Those are segment selection, feature computation, representation learning and fusion. In the segmentation step, we obtain time aligned boundaries of sub-segments corresponding to words, word-final syllables and word-final syllable nuclei using time-aligned boundaries of phonemes, syllables and words. In the computation step, we obtain frame-level MFCC features and frame and segment-level prosody based features for word and word-final syllable sub-segments. The frame-level prosody based features are the low-level representations of prosodic variations. On the other hand, the segment level features are computed based on the knowledge of prosodic properties at the boundary, in which word-final syllable nuclei sub-segments are used. In the representation learning step, we learn high-level representations for word and word-final syllable segments separately from the combined frame-level prosodic and MFCC features corresponding to those segments in a data-driven manner using CNNs. In the fusion step, the high-level representations of the word and word-final syllable segment belonging to each word are concatenated with segment level prosodic features to obtain a 1-d feature vector. Following this, the 1-d feature vector is considered to detect the phrase boundary using a DNN.

A. Segment selection

For detecting the phrase boundary, many prior works have considered the features computed within the three types of segments – 1) word, 2) syllable, and 3) syllable nucleus by assuming boundary labels are available with the respective segments. However, few of the works have considered only the word segments [13], [15], [18], [16], [12], [14]. On the other hand, the remaining works have considered only syllable and its nucleus segments [17], [20], [19], [21]. Further, it is generally assumed that the phrase boundaries occur only at the last syllable of a word, referred as word-final syllables [12], [17], [20], [19], [21]. Considering these, most of these works have forced the predicted boundary labels associated with the non word-final syllables as '0' [17], [20]. Unlike the previous works, we explore the benefit of the features computed from all the three types of segments by assuming boundary labels are associated with the words only. For this, without loss of generality, we consider only the word-final syllable and the word-final syllable nucleus of a word for feature computation instead of all the syllables and the syllable nuclei in the word.

B. Feature computation

1) Frame-level features:: It has been observed that the phrase boundaries are identified based on acoustic cues indicating prosodic properties such as tonal & stress variations and pre-boundary lengthening [1], [3]. In order to capture these cues for the task of boundary detection, existing works have computed a set of segment-level features based on statistical functions applied on frame-level pitch and energy values [17], [20], [19], [21], which have been assumed to capture prosodic properties. However, these statistical functions are derived heuristically based on knowledge. Unlike these knowledge-driven approaches, we, in this work, propose to model the variations in frame-level features representing prosodic properties, referred to as frame-level prosody based features, in a data-driven manner. Further, considering these features, we allow the models to learn the feature representations at the segment-level. Following the work by Stehwien et al. [18], in this work, we consider the following frame-level prosody based features – smoothed frequency, smoothed root mean square(RMS) energy, harmonic-to-noise ratio, Pulse-code modulation(PCM) loudness and voicing probability. In addition to these features, they have also shown the benefit of MFCC features in the boundary detection task. Thus, we also consider MFCC features for the boundary detection task.

2) Segment-level features:: On the other hand, we hypothesize that the data-driven modelling of frame-level prosody and MFCC features might not capture the pre-boundary lengthening property. This is because it has been shown that the pre-boundary lengthening could be captured using the duration of the segments such as syllable and syllable-nuclei [22], [17], [19]. Thus, in this work, following the work by Ananthakrishanan et al. [17], we also include the duration based features – syllable and syllable nuclei duration [17], referred to as segment-level features.

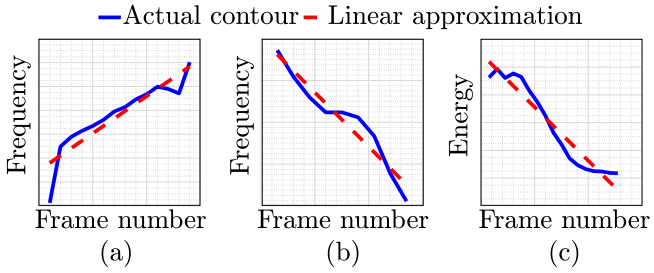

Fig. 2. Approximate linear curve plot for the (a), (b) Frequency variation and (c) Energy variation, within the word-final syllable segment of word 'APPOINTED', 'CALLS' and 'MEMBER' respectively.

3) Δ and $\Delta\Delta$ frame-level features:: In order to represent the dynamic variations in the frame-level features, most of the works have included Δ and $\Delta\Delta$ variations to the frame-level features [26]. In general, pitch and energy variations are related to the following prosodic properties – tone and stress. As described in ToBI, most of the boundary tones have either a linear rising, a linear falling or constant trend in the pitch values [1]. Similarly, Suni et al. [27] have emphasized a linear falling trend in the energy values at the boundary. Thus, we hypothesize that including Δ and $\Delta\Delta$ variations of frame-level prosody features could capture boundary specific prosodic cues much better by capturing such a linear trend in the pitch and energy values at the boundaries.

These are illustrated in Figure 2 with exemplary word-final syllable segments taken from the BURN corpus. Figure 2a and b show pitch values in the word-final syllables of the words 'APPOINTED' and 'CALLS', respectively. The annotations indicate that the pitch values follow rising and falling trends within these syllables' respective segments. In order to verify this, we perform linear regression on the pitch values and plot the obtained results in red colour for each segment separately. From the figures, it is observed that the lines follow their respective ground-truth rising and falling trends as per ToBI [1] as well as closely follows the pitch values in both the segments. Considering these, we believe that the pitch values at the boundary approximately follow a linear trend. Similarly, in Figure 2c, a falling trend in the energy values can be observed for the word-final syllable segment of the word 'MEMBER'.

C. Representation learning

We use CNNs for learning high-level representations from the frame-level prosody and MFCC features for word and word-final syllable segments separately. In the literature, CNNs have been shown to be useful in learning representations [28]. Typically, CNNs learn representations from the input by performing convolution using a kernel followed by pooling. Some of the existing works have used 2-dimensional (2D) kernels for representation learning to capture temporal and spatial dependencies in the input [29], [30]. Similarly, few other works have used 1-dimensional (1D) kernels to obtain

TABLE I
AVERAGE ACCURACY AND F1 SCORE EVALUATED OVER 10-FOLDS WITH DIFFERENT BASELINES AND PROPOSED MODEL FOR BOTH BURN AND SWITCHBOARD CORPORA.

	BURN corpus		Switchboard corpus	
	Accuracy (std)	F1 score (std)	Accuracy (std)	F1 score (std)
Proposed model	85.53 (0.41)	70.72 (3.26)	81.80 (1.58)	61.55 (3.69)
BL-NN	84.61	–	79.17 (1.70)	52.13 (2.75)
BL-CNN	84.08 (0.54)	68.38 (2.11)	80.87 (1.45)	58.04 (3.78)
Results with variants in proposed model				
w/o Δ, $\Delta\Delta$	85.53 (0.81)	70.78 (3.49)	81.92 (1.66)	61.19 (2.95)
w/o f_{seg}	85.48 (0.45)	70.96 (2.35)	81.71 (1.49)	61.28 (2.91)
w/o both	85.04 (0.85)	69.41 (3.48)	81.51 (1.66)	58.22 (4.33)

data-driven filtered output from the input [31]. Recently, the effectiveness of CNN based modelling has been shown in detecting prosodic events and lexical stress [18], [32]. Considering these as well as the effectiveness of 2D and 1D kernels of the CNNs, in this work, we consider a CNN with 2D kernels (2D-CNN) followed by a CNN with 1D kernels (1D-CNN) and then average pooling for learning representations.

In the 2D-CNN, we use 100 number of 2D kernels of size $d\times 6$ with a stride value of 4, where d is the total size of frame-level prosody and MFCC features per frame. In the 1D-CNN, we use and 100 number of 1D kernels of size 4 with a stride value of 2. We perform average pooling on the convoluted output from each kernel obtained after the 1D-CNN.

D. Fusion

We hypothesize that proposed segment-level features and frame-level features are useful for the boundary detection task. However, the segment-level features are 1D feature per segment. Thus, it cannot be directly given as input to CNN. In order to consider the benefit from both the features, we use a DNN, which takes the feature vector of size 202 by concatenating segment-level features (size of 2) with representations learnt from CNNs from both word and word-final syllable segments (size of 100 each). The DNN consists of one hidden layer of 32 units, followed by a softmax layer. We use the relu activation function in each hidden unit.

IV. EXPERIMENTAL RESULTS

A. Experimental setup

1) Feature computation: We obtain frame-level prosody feature using OpenSMILE toolkit [33], and MFCC using Praat toolkit [34] for both the BURN and Switchboard corpora. In the MFCC computation, we use a fixed set of filter cut-off frequencies and band-width as available in the Praat. Thus, MFCC features' size depends on the utterances' sampling frequency, which results in 26 and 19 MFCC features for BURN and Switchboard corpora, respectively. Both the frame-level prosody and MFCC features are computed using a 20ms window with a 10ms overlap. We perform zero-padding to ensure fixed input feature dimensions separately for word and word-final syllable segments.

2) Modeling: We perform experiments using all words from all the utterances that contain ToBI break index annotations separately for both the corpora. A 10-fold cross-validation setup is considered in the experimentation, where eight folds are used for training, one for validation and one for testing. We perform mean and variance normalization only on segment-level features using mean and variance values computed from the training set. Following the work by Stehwein et al. [18], we consider the frame-level MFCC and prosody features without normalization. We implement the model using Keras [35]. The training is performed with 20 epochs with a batch size of 10, including early-stopping and model-checkpoint. In order to know the effectiveness of the proposed approach, we consider two baseline schemes – 1) CNN based work proposed by Stehwein et al. [18], referred to as BL-CNN, 2) NN based work proposed by Ananthakrishnan et al. [17], referred to as BL-NN. We conduct the experiments on Switchboard and BURN corpora.

B. Results & Discussion

Table I shows accuracies and F1-scores (standard deviation in brackets) averaged across all ten folds obtained with the two baselines and proposed models for both the corpora. From the table, it is observed that the proposed model performs better than both the baselines on both the corpora. This indicates the effectiveness of the proposed approach, which benefits from combining the following three contributions – 1) learned representations from word-final syllable segments, 2) Δ and $\Delta\Delta$ frame-level prosody features, and 3) fusion of segment-level features. One baseline (BL-CNN) uses only the learned representations from the word segments, and the other baseline (BL-NN) uses only heuristically derived segment-level features. Further, in order to know the effectiveness of the three contributions, we compute average accuracies and F1-scores under following conditions – 1) learned representations without Δ and $\Delta\Delta$, but with segment-level features (w/o Δ and $\Delta\Delta$), 2) learned representations with Δ and $\Delta\Delta$, but without segment-level features (w/o f_{seg}), and 3) learned representation without both Δ and $\Delta\Delta$ and segment-level features (w/o both).

From the table, it is observed that the accuracies obtained with the proposed model are higher in most of the cases than those obtained with three variants in the proposed model under both the corpora. This indicates the effectiveness of

the combination of all three variants in the proposed model. Comparing the accuracy and F1 score of the proposed model only with Δ, $\Delta\Delta$ and without both the f_{seg} and Δ, $\Delta\Delta$, the higher values with Δ, $\Delta\Delta$ indicates the benefit of the proposed Δ, $\Delta\Delta$ features. Similarly, the higher accuracy and F1 scores values with only f_{seg} comparing to those without both the f_{seg} and Δ, $\Delta\Delta$ indicate the benefit of the f_{seg} features which are obtained in a knowledge-driven manner. Further, comparing the accuracies and F1 scores from the proposed approach without both the f_{seg} and Δ, $\Delta\Delta$ with those from BL-CNN, the higher values indicate the need for using frame-level features within both the word and word-final syllable segments.

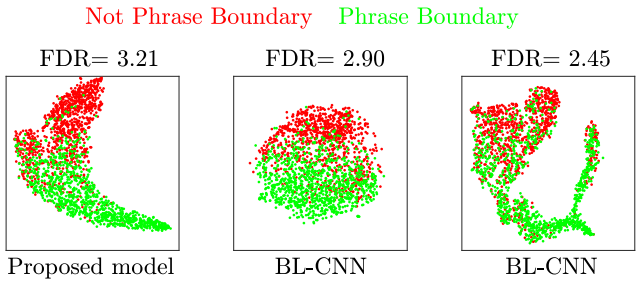

Fig. 3. Scatter plot of the 2D transformed output from t-SNE on the representations obtained with the respective model at the input of soft-max layer for proposed model, BL-CNN and BL-NN. In each case, FDR values are computed and shown above the respective plots.

Further, we analyze the effectiveness of the proposed approach considering two-dimensional (2D) transformed output obtained with t-SNE [36] shown in Figure 3 in comparison to those obtained for BL-CNN and BL-NN. It has been shown that the t-SNE method is useful in visualizing data, in which a non-linear transformation exists between input and output. In order to obtain 2D output, we use the representations at the input of the soft-max layer separately for all the three models on the test set as the input to the t-SNE method. The dimensions of these representations are 32, 100 and 25 in the proposed model, BL-CNN and BL-NN, respectively. We hypothesize that these representations capture the effectiveness of both the input frame/segment level features and the representations learned with the model. We also compute the Fisher discriminant ratio (FDR) on 2D outputs to know the discriminability between the 2D representations between the two classes quantitatively for each model. The 2D representations are more separable from the figure, and the respective FDR is higher in the case of the proposed approach than those with BL-CNN and BL-NN. This also suggests the effectiveness of the learned representations and segment-level features used in the proposed model for the boundary detection task.

V. Conclusions

Unlike the existing works, which consider either knowledge-based features (duration) or data-driven representations from frame-level prosody and MFCC features, we consider both and propose a fusion-based modelling. With this, we could incorporate the duration-based features in the modelling, which the data-driven approaches might not learn. We also consider frame-level features within word and word-final syllable in contrast to only word segments as considered in the existing methods. Experiments on BURN and Switchboard corpora revealed that the proposed method performs better than the best of the existing knowledge-based methods and data-driven methods. Further investigations are required to consider context information for phrase boundary detection. Future works also include estimated lexical and syntactic features and effects on the phrase boundary detection.

References

[1] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," *The Ohio State University Research Foundation*, vol. 3, 1997.

[2] J. B. Pierrehumbert, "The phonology and phonetics of english intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[3] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody." in *Second international conference on spoken language processing*, 1992, pp. 867–870.

[4] C. Yarra and P. K. Ghosh, "voisTUTOR: Virtual Operator for Interactive Spoken English TUTORing," in *Eighth International Speech Communication Association Workshop on Speech and Language Technology in Education*, 2019, pp. 35–36.

[5] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Language learning & technology*, vol. 2, no. 2, pp. 62–76, 1999.

[6] D. Jouvet, "Speech processing and prosody," in *International Conference on Text, Speech, and Dialogue*. Springer, 2019, pp. 3–15.

[7] B. Schuppler and B. Ludusan, "An analysis of prosodic boundary detection in german and austrian german read speech," in *International Conference on Speech Prosody*, 2020.

[8] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, 2004, pp. 105–114.

[9] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition," *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010.

[10] S. Stehwien, A. Schweitzer, and N. T. Vu, "Acoustic and temporal representations in convolutional neural network models of prosodic events," *Speech Communication*, vol. 125, pp. 128–141, 2020.

[11] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech," *Mathematical Foundations of Speech and Language Processing*, vol. 138, p. 105, 2012.

[12] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 469–481, 1994.

[13] X. Sun and T. H. Applebaum, "Intonational phrase break prediction using decision tree and n-gram model," in *Seventh European Conference on Speech Communication and Technology*, vol. 1, 2001, p. 537–540.

[14] V. K. Rangarajan Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 797–811, 2008.

[15] Ken Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. 1–509.

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

[17] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[18] S. Stehwien and N. T. Vu, "Prosodic event recognition using convolutional neural networks with context information," in *Proceedings of Interspeech*, 2017.

[19] H. Jeon, Je and Y. Liu, "Automatic prosodic event detection using a novel labeling and selection method in co-training," *Speech Communication*, vol. 54, no. 3, pp. 445–458, 2012.

[20] S. Ananthakrishnan and S. S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2005, pp. I/269–I/272.

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

[22] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *The Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[23] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1992, p. 517–520.

[24] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[25] J. Tauberer, "P2TK automated syllabifier," 2018.

[26] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Eighth European conference on speech communication and technology*, 2003, pp. 2573–2576.

[27] S. Antti, S. Juraj, and V. Martti, "Boundary detection using continuous wavelet analysis," in *International Conference on Speech Prosody*, 2016, pp. 267–271.

[28] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *The Institution of Engineering and Technology Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[29] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5619–5629, 2018.

[30] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17 913–17 922, 2018.

[31] E. David, M. Ansorge, L. Goras, and V. Grigoras, "On the sensitivity of CNN linear spatial filters: Non-homogeneous template variations," in *Eighth IEEE International Workshop on Cellular Neural Networks and their Applications*, vol. 8, 2004, pp. 40–45.

[32] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic classification of lexical stress in English and Arabic languages using deep learning," in *Proceedings of Interspeech*, 2016, pp. 175–179.

[33] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Association for Computing Machinery international conference on Multimedia*, no. 4, 2013, pp. 835–838.

[34] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[35] F. Chollet *et al.*, "Keras," 2015.

[36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.