DEEP LEARNING FOR PROMINENCE DETECTION IN CHILDREN'S READ SPEECH

Mithilesh Vaidya   *Kamini Sabu*   *Preeti Rao*

Department of Electrical Engineering,
Indian Institute of Technology Bombay, Mumbai, India

ABSTRACT

The detection of perceived prominence in speech has attracted approaches ranging from the design of knowledge-based linguistic and acoustic features to the automatic feature learning from suprasegmental attributes such as pitch and intensity contours. We present here, in contrast, a system that operates directly on segmented speech waveforms to learn features relevant to prominent word detection for children's oral fluency assessment. The chosen CRNN (convolutional recurrent neural network) framework, incorporating both word-level features and sequence information, is found to benefit from the perceptually motivated SincNet filters as the first convolutional layer. We further explore the benefits of the linguistic association between the prosodic events of phrase boundary and prominence with different multi-task architectures. Surpassing the previously reported performance on the same dataset of a random forest ensemble predictor trained on carefully chosen hand-crafted acoustic features, we evaluate further the possibly complementary information from hand-crafted acoustic and pre-trained lexical features.

Index Terms— prosody, word prominence, waveform-based learning, end-to-end training, multi-task learning

1. INTRODUCTION

The prosodic structure of speech carries important information in terms of the syntax and the meaning, both of which are critical to a listener's ease of comprehension of the spoken message [1, 2, 3]. Phrase boundaries embed sentence syntax through word grouping while prominence or emphasis on specific words signals new information or highlights a contrast. Given their importance in applications requiring speech understanding such as scoring of spoken language fluency and text-to-speech synthesis, the automatic detection of perceived prominence and phrase boundary has attracted continuous research efforts. Prominence or word stress, the focus of the current work, has proved more challenging of the two.

Prominence is perceived by a listener when a word stands out of its local context in one or more of the suprasegmental attributes such as duration, F0, intensity and spectral shape [4]. The local context itself refers to the phones and syllables within the word as well as a neighborhood of up to several words. Prosody perception, however, is influenced not only by the low-level acoustic cues but also top-down expectations from lexico-syntactic information [5, 6]. The precise combination and relative importance of the cues depends on the speaker, language and speaking style as also on the listener. Traditionally, various aggregates of the sampled acoustic parameters across the word segment including mean and variance, contour shape descriptors, and differences in these quantities across neighboring words comprise word-level prosodic features [7, 8]. These features are then used to train a conventional supervised classifier for the automatic detection, possibly in combination with lexico-syntactic information [9, 10]. In our own recent work motivated by a children's

oral reading assessment application, we used feature selection on a large set of features, computed across the distinct suprasegmental attributes of speech, in a random forest ensemble predictor to derive a compact set of interpretable features for speaker-independent boundary and prominence detection [11]. It was observed that apart from the expected pitch, duration and intensity based aggregates, the acoustic cues to prominence included a number of spectral shape functionals while the phrase boundary prediction was dominated by pause based features. With the search space for such 'hand-crafted' features being very large, however, the process can miss potentially important features. Further, the pre-selected context windows used in such analyses make it difficult to exploit the long and variable time scale of prosodic relationships across an utterance in any comprehensive manner. The potential for deep learning solutions has therefore been recognized for some time but incorporated successfully in the prominence detection task only more recently, as briefly reviewed next.

Rosenberg et al. [12] used a large number of acoustic-prosodic features and aggregates at word level derived from their previous AuToBI work [7, 13] in a BiRNN classifier where the word sequence context was learned over that explicitly provided in the feature vector. They observed a small improvement ($< 1\%$ absolute) in boundary and pitch accent detection over a baseline conditional random forest classifier. Wu et al. [14] also used similar aggregated acoustic features with an LSTM to find an improvement over the use of an SVM classifier. Lin et al. [15] used a hierarchical BLSTM network to aggregate features across phone, syllable and word to model contextual information at multiple granularities. As opposed to modelling each task separately, joint detection of boundary and prominence was found to improve the F1-score by 2%.

In a departure from the above pre-computed word-level features, Stehwien et al. [16, 17] used CNN on sampled acoustic parameters (energy, F0, loudness, voicing probability, zero crossing rate and harmonic-to-noise ratio) together with a context window of two neighbouring words to optimally learn the word-level aggregated features. The max-pooled CNN feature maps are directly classified with a softmax layer. With word position indicators provided in the input segment, they report an improvement of 1-3% absolute over Rosenberg [7] on lexical stress and phrase boundary detection on the BURNC corpus, with speaker-independent scenarios being more challenging. Both local acoustic features and longer, more global contexts spanning several words and possibly different sentences across the utterance are important in the perception of prominence. Hence, architectures combining low-level feature aggregation with sequence models were realized with the same contour-learned features input to an LSTM classification layer. A multi-channel CNN for low-level feature extraction, combined with a bidirectional-LSTM for capturing longer context, outperformed a feedforward DNN by 22% (F1-score) [18].

Feature learning via end-to-end neural network systems trained

on speech waveforms is being increasingly viewed as the optimal approach to complex classification tasks [19, 20, 21]. Such systems have achieved performances close to, but not always exceeding, those of classifiers with task-specific hand-crafted features. This hints at a need for the introduction of reasonable constraints, or additional information, in the network architectures especially in the widely encountered data-constrained scenarios. In this work, we explore precisely such variations for the prominence detection task from segmented speech waveforms starting from a straightforward CRNN model. This is the first case of prosodic event detection from speech waveforms that we are aware of.

The first variation involves replacing the CNN layer at the input with bandpass constrained, but tunable, filters motivated by the traditionally used Mel filterbanks emulating the low-level auditory processing [20]. SincNet has been applied in frame-level speaker identification where its hyperparameters have been found to be critical, although sometimes counter-intuitive, to achieved task accuracy [22]. Next, we try to exploit the linguistic association between phrase boundaries and prominent words with multi-task learning. The presence of phrase-finality increases the perceived prominence of the word [23, 24] and can potentially contribute to the feature representation for prominence. Recent work on the joint prediction of boundary and prominence is promising, with the boundary predictions computed from the final layer output of a 3-layer BLSTM network while prominence predictions are made at the penultimate layer [15]. In another attempt, prosodic event classification is viewed as a 4-class problem [25]. Developing the above theme further, we explore alternate multi-task architectures for prominence detection that incorporate information about the (typically more reliably predicted) phrase boundary status of the word in distinct ways.

Our performance baseline is previous work on the same dataset and task in which a random forest ensemble predictor uses highly tuned hand-crafted features [11]. We also examine the possibly complementary information of the hand-crafted features. Finally, given the importance of lexical information in prominence detection [26, 27], we report the combination with pre-trained word embeddings.

2. DATASET AND TASK

The children's oral reading dataset used in this work comprises recordings of grade-appropriate text read aloud by selected middle school students with reasonable word decoding ability in English (as second language) but widely varying levels of prosodic skill [11]. The individual utterances are story paragraphs comprising between 50-70 words, each word labeled separately for the presence/absence of prominence and phrase boundary by 7 naive listeners using the RPT methodology [28]. This is reduced to a net rating per word based on number of votes (out of 7), further scaled down to range 0-1, to obtain the 'degree' of prominence (boundary) per word. The dataset contains 41,286 words across 790 utterances by 35 speakers, recorded at 16 kHz sampling rate. The utterances are available segmented at word level by forced alignment with the manual transcript.

The waveform for each segmented word is padded to the maximum size in the dataset, which is approximately 1.79 seconds (or 28,660 samples). A silent pause before the word (with duration limited to 500 ms if higher) is included in the waveform segment. The dataset is split into three equal folds with no speaker overlap for 3-fold cross-validation based testing. The hyperparameters are tuned with 4-fold CV on the train split. The 4 trained models are used for inference and their predictions are averaged to generate results on the corresponding unseen test set. The results for the prominence degree prediction are reported in terms of Pearson correlation between the predicted output and the degree of prominence from the
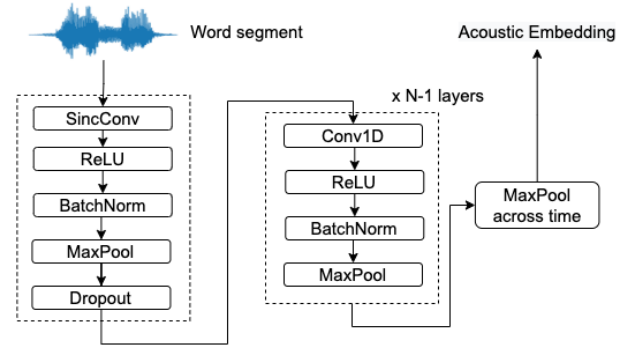

Fig. 1. CNN component of the model, where the first layer is replaced with a Sinc layer.

RPT rater votes which serves as ground truth. We report the mean Pearson correlation and the standard deviation across the three test folds.

3. MODEL

The input to our prominent word detection network is the sequence of word-level waveform segments. We build up our model from a basic CRNN operating on waveform segments to more complex architectures that incorporate additional information.

3.1. Waveform CRNN

As the name suggests, CRNN consists of a convolutional neural network, followed by a recurrent neural network. The CNN layers output a fixed-dimensional feature embedding, while the RNN processes the sequence of embeddings. As in some recent high-level speech classification systems, our CNN is composed of multiple layers, where each layer consists of 1D convolution with batch normalization, ReLU activation and max pooling [19]. After N CNN layers (where N is a hyperparameter), the CNN output is max-pooled across time to get a fixed dimensional embedding for each word.

With a motivation to improve the input representation with meaningful parameter constraints in the first convolutional layer, we experiment with Sinc filters as shown in Figure 1. The constrained band-pass Sinc filters are expected to not only improve performance but also speed up training. The hyperparameters of the Sinc layer (number of filters, window width and stride) are tuned on the data and task.

Prominent words stand out through the acoustic changes in the local context [29]. Prominence detection systems typically include the context explicitly during the feature extraction. In deep learning, the word context that is critical in speech prosody perception is modelled by the sequence classifier GRU. The relevant local context is therefore learned from the sequence of words spanning the entire utterance. At each time step, the GRU takes as input a fixed-dimensional vector corresponding to each word. This vector could be the acoustic embedding extracted from the CNN, word-level features such as lexical features or a concatenation of the two. The output of the, possibly bidirectional, GRU at each time step is fed to a dense network consisting of two fully-connected (FC) layers. The first FC layer, consisting of 128 neurons, is followed by ReLU activation and a dropout layer. The second FC layer has one neuron, followed by Sigmoid activation to get the final score. Mean Squared Error (MSE) between model prediction and ground truth is minimised.
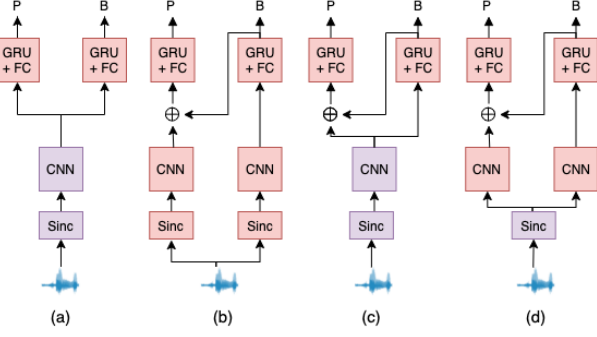
Fig. 2. MTL architectures: (a) Sharing paradigm in which a common CNN feature extractor is followed by separate GRU heads for prediction of prominence (P) and boundary (B). (b) Conditioned MTL with separate CNNs for P and B predictions. (c) Shared CNN combined with conditioning by boundary prediction. (d) Conditioned MTL with only the Sinc layer shared. Note that $\oplus$ denotes concatenation. GRU blocks also include two fully-connected layers at the output for final prediction as discussed in Section 3.1

3.2. Multi-task learning extension

As discussed in Section 1, phrase boundary detection is a closely related task which can bring in complementary information to boost the prominence prediction performance. Given the different ways in which the relevant auxiliary task can be incorporated in the training, we experiment with two distinct multi-task learning (MTL) paradigms: parameter sharing and conditioning [30].

In parameter sharing, the two tasks share the initial feature extraction through shared CNN layers, while the subsequent layers (GRU and Dense) responsible for the final classification are different, as depicted in Figure 2(a). This sharing of parameters at the feature extraction stage prevents overfitting and hence improves generalisation. On the other hand, in conditioned MTL, we have a separate network branch for each task with the final boundary prediction provided as an additional input to the GRU of the prominence prediction task. In this case, the CNN feature extractor can be shared (Figure 2(b)) or separate (Figure 2(c)). Based on an understanding of the role played by the Sinc layer in emulating low level auditory feature extraction, we explore a new architecture in which the Sinc layer is shared across the two tasks but the subsequent CNN, GRU and Dense layers are separate, as in Figure 2(d).

For MTL, the final loss is a convex combination of the prominence MSE loss ($L_{prominence}$) and phrase boundary MSE loss ($L_{boundary}$) i.e.

$$L_{total} = \alpha L_{prominence} + (1 - \alpha)L_{boundary} \tag{1}$$

where α is a hyperparameter which controls the trade-off between performance on the main task and the auxiliary task. For the single-task experiments, $\alpha = 1$.

3.3. Hyperparameters

The hyperparameters for the basic CNN model are tuned for performance on the validation set. These were varied in the following ranges: number of layers, N (2-8), number of filters (16-128), CNN kernel width (7-151), pooling width (2-4) and stride (1 and 2). We found the optimal configuration to be: 4 layers, each consisting of 32 filters of kernel width 51, stride 1 and max pooling with kernel size of 3.

For the GRU, it was found that a 3-layer, 256-dimensional bidirectional GRU with dropout of 0.5 at each hidden layer consistently

Table 1. Performance of different architectures in single-task learning. A34 refers to the 34 hand-crafted features of [11]. Number of filters in all CNN and Sinc layers is 32 while pool size is 3. For layers 2, 3 and 4, the hyperparameters are fixed as in Section 3.3 while Layer 1 variations are reported here in terms of Pearson correlation (s.d. < 0.01).

No.	Input	Acoustic model	Layer 1 (type, width, stride)	Pearson correl.
1.	A34	RFC	-	0.696
2.	A34	GRU	-	0.726
3.	Wav	CRNN	Standard, 51, 1	0.692
4.	Wav	CRNN	Sinc, 51, 1	0.712
5.	Wav	CRNN	Sinc, 31, 2	0.721
6.	A34 + Wav	CRNN	Sinc, 31, 2	0.735

gave the best performance. The hyperparameters of the GRU and the subsequent FC layers are fixed for all experiments.

For training, Adam [31] optimizer is used with a learning rate of 0.001. Batch size is set to 64 and the model is trained on a single NVIDIA GeForce GTX 1080. Early stopping is used on the validation set with patience set to 12 epochs. A single round of training and evaluation on our dataset took about 6 hours.

4. EXPERIMENTS

Prominence prediction is evaluated at the word level via its correlation with the ground-truth degree of prominence. We report here the test data performances for the different models in terms of the mean (all standard deviations were found to be less than 0.01) across the 3 test folds discussed in Section 2.

4.1. Single-task architectures

We start with reporting the performance of the baseline system that uses a set of 34 word-level acoustic-prosodic features computed on the acoustic contours of pitch, intensity and spectral shape versus time as well as various segmental durations including pauses. The final set of 34 features (termed 'A34' here) is eventually obtained by two stages of feature selection in a random forest ensemble predictor [11]. We also test the same A34 set of features with a bidirectional GRU classifier to study whether the implicit 'learning' of word context can bring benefits. The results appear in Table 1. We note that the performance shows a clear improvement with the sequence classifier, indicating the value of learned context over that explicitly represented within the A34 feature computations.

Next, we evaluate the performance of our CRNN model operating directly on the speech waveform. We note, from Table 1, a rise in the Pearson correlation with the Sinc layer replacing the (unconstrained) first convolutional layer but keeping the number of filters and filter widths and stride unchanged. A reduction of the Sinc filter widths to 31 samples (2 ms) gave a further improvement, especially when the stride was concurrently changed to 2 samples from 1 sample. This is consistent with the observations of previous work that smaller filter widths in the Sinc layer are superior in the context of speaker recognition [22]. While this seems counter-intuitive given that auditory filter impulse responses at lower centre frequencies are of duration well over 10 ms, the reduced frequency resolution due to the apparent truncation does not seem to harm the performance. It is also possible that the slightly higher stride (2 samples) helps to counteract the shortened filter widths to some extent. It is encouraging to see that the gap between waveform-learned and the hand-crafted A34 features almost closing with this tuned Sinc version. To check

Table 2. Performance of various multi-task learning architectures and additional features. (Pearson correlation s.d. < 0.01)

No.	MTL variant and additional features	Pearson correl.
1.	Tuned Sinc (without MTL)	0.721
2.	Fig 2(a)	0.726
3.	Fig 2(b)	0.727
4.	Fig 2(c)	0.724
5.	Fig 2(d)	0.740
6.	Fig 2(d) + A34/A27	0.757
7.	Fig 2(d) + A34/A27 + GloVe	0.813

for any complementarity in the two representations, we concatenate the A34 features with the 32-dimensional CNN embedding in the final column of the table to obtain a performance that exceeds that of either. However, ablation studies involving different suprasegmental attributes underlying the A34 features did not reveal any specific contribution as dominant. This indicates the future potential for better waveform-based feature learning, possibly with a larger training dataset.

4.2. Multi-task architectures

The MTL experiments reported in Table 2 were carried out after first tuning α. After a preliminary grid search, we found that the best performance across configurations is obtained when α is set to 0.95 in equation 1 (after scaling the MSE of each to bring them into the same range). From the correlations reported in Table 2, we note that neither of conditioning or shared CNN layers is better than the other. There is no clear improvement over single-task learning seen either. On the other hand, an increase in performance is seen with conditioned MTL when only the Sinc layer is shared and the other CNN layers remain task-dependent. This is consistent with our expectation that the lowest level features extracted from the input waveform correspond to the basic suprasegmental attributes fundamental to all prosodic event detection. Therefore the parameters of the constrained convolutional layer, that is the Sinc layer, get better tuned in the multi-task set-up.

In row 6 of Table 2, we report performance on the concatenation of the hand-crafted acoustic features with the generated CNN embeddings at the GRU input. Similar to the A34 features for prominence, A27 refers to an optimal set of 27 word-level hand-crafted features for the task of boundary prediction obtained in [11]. We expect the boundary task to benefit from the hand-crafted boundary detection features (A27) and therefore influence the prominence prediction performance. The observed improvement in performance confirms the presence of complementary acoustic information not captured by the purely waveform-based architecture.

4.3. Incorporating lexical features

Prominence is linked to the text syntax and semantics with content words such as proper nouns expected to receive prominence most of the time, followed by adjectives, nouns, adverbs, and verbs respectively [6]. Lexical features in the form of GloVe embeddings can implicitly capture such parts-of-speech information and have been found to help in the context of prosodic event detection [27]. To explore the usefulness of lexical information for our task, we test two popular word embeddings: GloVe [32] and BERT [33]. In the present work, we extract 100-dimensional GloVe embeddings pre-trained on Wikipedia using the gensim package. The embedding is passed through a dropout layer followed by a linear layer whose dimension is a hyperparameter [27]. After tuning, we found

that a dropout layer of probability 0.3 and a fully-connected layer of dimension 300 gave the best performance. Although BERT has replaced GloVe embeddings in a wide variety of NLP tasks and demonstrated good performance for prominence detection [26], it did not give any clear improvement over GloVe in our task. This could be attributed to very simple story texts without semantic ambiguities that may benefit from contextualized embeddings. We note a big jump in performance in the final row of Table 2 where the GloVe features are concatenated with the corresponding CNN embedding and hand-crafted features in each branch of the best MTL model, emphasising the importance of lexical features. Although this is at odds with the expectation that beginning readers do not necessarily realize prominence correctly, it supports the important role of the top-down expectations in raters' perceptions.

5. CONCLUSIONS

In this paper, we attempt to replace hand-engineered features for prominent word detection with deep learning models operating on the speech waveform. Our results indicate that it is challenging to surpass the performance of hand-crafted features computed from across the prior extracted suprasegmental contours of speech essential to prosody realization, at least with moderate sized training datasets. However, some human speech and audition motivated constraints such as Sinc based convolution at the lowest feature extraction stage can improve the performance of deep learning models. The optimal hyperparameters for the Sinc filters turned out to correspond to low strides and low widths (high time resolution and low frequency resolution), confirming previous findings in the speaker identification context. Hand-crafted features further enhance the performance, indicating opportunities for further careful constraining and tuning of the deep learning models.

The use of boundary event labeled speech seems to aid the task of prominent word detection, as expected from the linguistic and acoustic association between the two. A multi-task architecture in which prominence prediction is conditioned on the concurrently predicted boundary and where the Sinc convolution filters are shared across the two tasks obtained the best performance. Finally, we investigate the contribution of pre-trained lexical embeddings to prominence prediction to find a large increase in the performance. This indicates that the lexical identity of words guides the top-down expectations of raters, even in the case of not-so-proficient beginning readers.

Although demonstrated on a children's speech dataset, the proposed methods are applicable more generally to any prominent word detection problem. In future work, we plan to explore more sophisticated CNN architectures, attention-based sequence modelling for better context representation and alternate methods to fuse acoustic and lexical information.

6. REFERENCES

[1] J. K. Bock and J. R. Mazzella, "Intonational marking of given and new information: Some consequences for comprehension," *Memory & Cognition*, vol. 11, no. 1, pp. 64–76, 1983.

[2] L. van Maastricht, T. Zee, E. Krahmer, and M. Swerts, "L1 perceptions of L2 prosody: The interplay between intonation, rhythm, and speech rate and their contribution to accentedness and comprehensibility," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 364–368.

[3] J. M. Levis and A. O. Silpachai, "Prominence and information structure in pronunciation teaching materials," in *Proceedings of the Pronunciation in Second Language Learning and Teaching conference*, Ames, IA, USA, 2017, pp. 216–229.

[4] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, no. 7/8/9, pp. 1044–1098, 2010.

[5] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[6] S. Baumann and B. Winter, "What makes a word prominent? Predicting untrained German listeners' perceptual judgments," *Journal of Phonetics*, vol. 70, pp. 20–38, 2018.

[7] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[8] T. Mishra, V. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Proceedings of INTERSPEECH*, Portland, OR, USA, 2012, pp. 1864–1867.

[9] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in French," in *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 116–119.

[10] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. V. Segbroeck, J. Kim, P. G. Georgiou, and S. S. Narayanan, "Automated evaluation of non-native english pronunciation quality: Combining knowledge- and data-driven features at multiple time scales," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015, pp. 493–497.

[11] K. Sabu and P. Rao, "Prosodic event detection in children's read speech," *Computer Speech and Language*, vol. 68, pp. 1–19, 2021.

[12] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015, pp. 3066–3070.

[13] A. Rosenberg, "AuToBI - A tool for automatic ToBI annotation," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010, pp. 146–149.

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

[15] B. Lin, L. Wang, X. Feng, and J. Zhang, "Joint detection of sentence stress and phrase boundary for prosody," in *Proceedings of INTERSPEECH*, Shanghai, China, 2020, pp. 4392–4396.

[16] S. Stehwien and N. T. Vu, "Prosodic event recognition using convolutional neural networks with context information," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2326–2330.

[17] 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.

[18] L. Zhang, F. M. Jia Jia and, S. Zhou, W. Chen, C. Zhang, and R. Li, "Emphasis detection for voice dialogue applications using multi-channel convolutional bidirectional long short-term memory network," in *Proceedings of INTERSPEECH*, Hyderabad, India, 2018.

[19] V. Rajan, A. Brutti, and A. Cavallaro, "Conflictnet: End-to-end learning for speech-based conflict intensity estimation," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1668–1672, 2019.

[20] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[21] J. Fritsch, S. P. Dubagunta, and M. Magimai.-Doss, "Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based CNNS," in *Proceedings of ICASSP*, Barcelona, Spain, 2020, pp. 6534–6538.

[22] D. Oneață, L. Georgescu, H. Cucu, D. Burileanu, and C. Burileanu, "Revisiting SincNet: An evaluation of feature and network hyperparameters for speaker recognition," in *Proceedings of European Signal Processing Conference*, 2021, pp. 1–5.

[23] J. Roy, J. Cole, and T. Mahrt, "Individual differences and patterns of convergence in prosody perception," *Laboratory Phonology*, vol. 8, no. 1, p. 22, 2017.

[24] J. Bishop, GraceKuo, and B. Kim, "Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription," *Laboratory Phonology*, vol. 82, p. 22, 2020.

[25] A. Suni, S. Kakouros, M. Vainio, and J. Šimko, "Prosodic prominence and boundaries in sequence-to-sequence speech synthesis," in *Proceedings of Speech Prosody*, Tokyo, Japan, 2020, pp. 940–944.

[26] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proceedings of Nordic Conference on Computational Linguistics*, Turku, Finland, 2019, pp. 281–290.

[27] S. Stehwien, N. T. Vu, and A. Schweitzer, "Effects of word embeddings on neural network-based pitch accent detection," in *Proceedings of Speech Prosody*, Poznan, Poland, 2018, pp. 314–318.

[28] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," *Computer Speech and Language*, vol. 45, pp. 300–325, 2017.

[29] S. Kakouros and O. Räsänen, "3PRO - An unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[30] R. Kelz, S. Böck, and C. Widnaer, "Multitask learning for polyphonic piano transcription, a case study," in *Proceedings of International Workshop on Multilayer Music Representation and Processing*, 2019, pp. 85–91.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations*, San Diego, CA, 2015.

[32] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019, pp. 4171–4186.