Learning Contextual Representation with Convolution Bank and Multi-head Self-attention for Speech Emphasis Detection

Liangqi Liu*†, Zhiyong Wu*†‡, Runnan Li*†, Jia Jia*†, Helen Meng* ‡
* Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
† Beijing National Research Centre for Information Science and Technology (BNRist),
Department of Computer Science and Technology, Tsinghua University, Beijing, China
‡ Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
E-mail: {llq17, lirn15}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk, jjia@tsinghua.edu.cn

Abstract—**In speech interaction scenarios, speech emphasis plays an important role in conveying the underlying intention of the speaker. For better understanding of user intention and further enhancing user experience, techniques are employed to automatically detect emphasis from the user's input speech in human-computer interaction systems. However, even for state-of-the-art approaches, challenges still exist: 1) the various vocal characteristics and expressions of spoken language; 2) the long-range temporal dependencies in the speech utterance. Inspired by human perception mechanism, in this paper, we propose a novel attention-based emphasis detection architecture to address the above challenges. In the proposed approach, convolution bank is utilized to extract informative patterns of different dependency scope and learn various expressions of emphasis, and multi-head self-attention mechanism is utilized to detect local prominence in speech with the consideration of global contextual dependencies. Experimental results have shown the superior performance of the proposed approach, with 2.62% to 3.54% improvement on F1-measure compared with state-of-the-art approaches.**

I. Introduction

Emphatic speech is widely employed to convey the underlying intention of the speaker in speech communication between humans. In human-computer interaction systems, capturing emphasis can help the understanding of user intention [1], and further enhance user experience. Therefore, automatic detection of emphasis from user's input speech has drawn a broad interest in the human-computer speech interaction research field.

Presented with attention-capturing vocal prominence, emphasis are enhanced speech segments corresponding to one or more words in an utterance, indicating the focus and intention of the speaker. To automatically detect emphasis in speech, techniques are developed from two perspectives: the informative feature extraction and emphasis detection model construction [1] [2]. The former one mainly focuses on extracting informative features related to emphasis from the speech signal. And the latter one aims to construct a robust and efficient framework to recognize emphasis from input speech using the extracted features.

The typical acoustic feature extraction can be divided into three trails with the consideration of different concept level: frame-level features extraction, syllable-level features extraction, and word-level features extraction. [3] proposed a frame-level automatic emphasis detection system using normalized pith for different speakers. [4] conducted a comparative study of emphasis detection at vowel, syllable, and word level, examining the optimal domain for accent analysis. [5] calculated the word-level F0 difference between genuine emphatic speech and synthetic neutral speech to label emphasis for accent words. [6] evaluated a total of nine measures on word-level spectral tilt to detect emphasis in speech. In this work, we employ the word-level feature extraction in the development of emphasis detection system based on the research conclusion of [4], in which the speech prominence is stated to be more significant at the word level. As proposed in [1] and [6], a series of statistic functions are employed to aggregate the emphasis related frame-level features, including F0, energy, and duration, to produce the word-level acoustic features. These features have a clear presentation in emphatic vocalizations, offering robust and discriminative patterns for emphasis detection.

To develop a robust and efficient framework for emphasis detection, various techniques are employed. [7] considered emphasis detection as a classification problem. Motivated by this, [8] employed support vector machines (SVM) and [9] used bayesian network (BN) to predict word prominence in spontaneous speech. However, both SVM and BN cannot incorporate the context information that emphasis detection mainly relies on. [2] and [10] formulated the emphasis detection problem as a sequential learning task and employed bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) to detect emphasis. With enhanced ability in capturing contextual information from both forward and

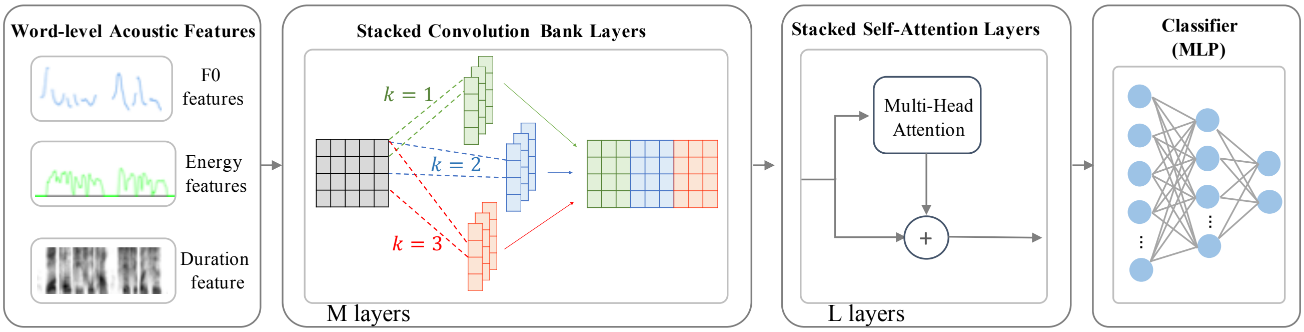


Fig. 1. The architecture of the proposed model (CB-SA) for speech emphasis detection: a stack of M convolution bank layers + a stack of L self-attention layers. k is the kernel size of convolutional filters.

backward directions, the BLSTM based models have shown significant performance improvement on emphasis detection.

Though state-of-the-art approaches have achieved significant performance improvement on emphasis detection, limitations still exist: 1) lack to use the contextual information of different dependency scope; 2) hard to model the long-range dependencies in the speech utterance. Previous research indicated the interaction between the emphasized words and their neighboring words. The words standing out from their environment is perceived to be emphasis [11] [12]. Meanwhile, the emphasized words often affect the acoustic features of their neighboring words [13]. However, the dependencies between the emphasized words and their neighboring words are not modeled in a direct and meaningful way in the above methods. Besides, human usually recognize emphasis based on the understanding of the whole sentence and focus attention selectively on parts of the utterance. But it is hard for current recurrent neural networks (RNNs) to model long-range dependency which is significant for many sequential tasks, including emphasis detection.

In this paper, we focus on addressing the automatic detection of emphasis from a modeling standpoint. Research indicated convolution bank which contains convolutional filters of different kernel size can explicitly model the local contextual information [14] of different dependency scope (the unigram, bigram, up to K-grams of the word). Motivated by this, we use convolution bank to extract the discriminative local features. As to the challenge of modeling long-range dependencies, [15] proposed Transformer network in neural machine translation (NMT), a model architecture relying entirely on self-attention mechanism. By using self-attention mechanism, any two inputs at different times are connected directly, which solves the long-range dependencies problem effectively. Inspired by this, we employ self-attention to model the long-range global dependencies between each word and the whole utterance.

The overall architecture of the proposed model is illustrated in Fig.1. We use a stack of M convolution bank layers and a stack of L self-attention layers to capture local and global dependencies respectively, and a fully connected multi-layer perceptron (MLP) as the classifier. The goal of the stacked convolution bank layers is to extract robust local representations for emphasis. Each convolution bank layer consists of a bank of 1-D convolutional filters of different kernel size (from 1 to K) to explicitly model local context in the unigram, bigram, up to K-grams manner, then the feature maps generated from convolutional filters of different kernel size are concatenated together. By relating different positions of a sequence directly, self-attention can effectively address the problem for modeling long-range dependencies. So we use the stacked self-attention layers to model the global dependencies between each word and the whole utterance. Each self-attention layer is made up of the multi-head attention mechanism and a residual connection [16] which is used to accelerate the training process.

The main contributions of this work can be summarized as:

1) using convolution bank to explicitly model the dependencies between the emphasized words and their neighboring words, extracting local contextual information that is indicative of emphasis from the utterance.
2) using multi-head self-attention to model the global dependencies of the utterance to each word.
3) superior performance of the proposed framework in experiments compared to state-of-the-art approaches.

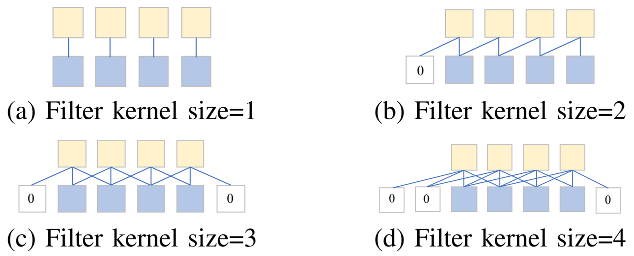
II. Methodology

In this paper, we propose the combination use of convolution bank and multi-head self-attention to recognize emphasis in speech, which can learn informative local contextual features and detect attention-aware prominence with the considering of global contextual dependencies.

A. Convolution bank

To model the local information of different dependency scope, convolution bank is adopted in this work. The convolutional filters in convolution bank have $k = 1, .., K$ kernel size to explicitly model the local contextual information in the unigram, bigram, up to K-grams manner, providing more discriminative feature learning performance for the following components.

As depicted in Fig.2, individual 1-D convolutional layers with different kernel size are employed to process along the temporal axis of input features $n = 1, ..., N$. When kernel size



(a) Filter kernel size=1 (b) Filter kernel size=2

(c) Filter kernel size=3 (d) Filter kernel size=4

Fig. 2. 1-D convolution with filters of different kernel size.

$k = 1$, the n-th hidden output is correlated with the n-th input frame only; when kernel size $k = 2$, the n-th output is related with both the n-th and (n-1)-th input frames; when kernel size $k = 3$, the n-th output is correlated with the (n-1)-th, n-th and (n+1)-th input, and so on. Zero-padding around the border is used to keep the hidden output length being the same with the input sequence, half at the beginning and half at the end.

For input word-level acoustic feature sequence x with N words, K sets of 1-D convolutional layers are employed simultaneously for feature learning, where the k-th convolutional layer is constructed with filters of kernel size k. Specifically, to boost the network training and enhance the model performance, batch normalization [17] is employed for all convolutional layers. The hidden output of convolutional layer is calculated as:

$$h^k = \phi(BN_{\gamma,\beta}(x \oplus w^k)) \tag{1}$$

where $\oplus$ denotes the 1-D convolution operation, w^k is the filter of the k-th convolutional layer, h^k is the feature map generated from given input sequence x. γ and β are learned parameters in batch normalization to scale and shift the normalized value, and ϕ is rectified linear unit (ReLU) activation function.

Feature maps generated from convolutional layers with different kernel size are then concatenated together to produce the learned features, and fed to the following components to recognize the emphasis.

B. Multi-head self-attention

Inspired by human perception mechanism, self-attention is employed in this work to capture the long-range global dependencies between each word and the whole utterance, enhancing the attention ability in detecting prominent vocalizations.

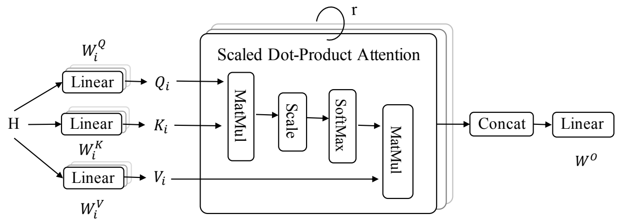


Fig. 3. Architecture of multi-head self-attention, where W_i^Q, W_i^K, W_i^V, and W^O are weight matrices.

For input sentence with N words, the d-dimensional hidden output $H = \{h_1, ..., h_N\}$ is firstly computed in previous convolution bank component, and scaled dot-product attention [15] is then employed to compute the attention value. Note that query sequence Q, key sequence K, value sequence V all come from H, which means $Q = K = V = H \in \mathbb{R}^{N \times d}$. And the attention value is calculated as follow:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

In emphasis detection, employing self-attention mechanism allows the model to capture the paired dependencies between words in all positions of the utterance, reflecting the impact of the query word to all words in the speech.

As depicted in Fig.3, the multi-head attention mechanism is further proposed to exploit the dependencies in different representation subspaces of the input sequence. Enhanced with parallel multi-head computing, multi-head self-attention performs multiple attention function r times to the sub queries, keys, values sequences $Q_i, K_i, V_i (i = 1, ..., r)$ which are computed with different learned linear projection using H as input. This mechanism has reported with higher attention ability in producing the representation in [15]. Modified from Eq.2, the multi-head self-attention is calculated as

$$\begin{gathered} MultiHead(Q, K, V) = Concat(head_1, ..., head_r)W^O \\ where\ head_i = Attention(HW_i^Q, HW_i^K, HW_i^V) \end{gathered} \tag{3}$$

where W_i^Q, W_i^K, W_i^V are the weight matrices in parallel attentions with dimension d/r. W^O is the output weight matrix with dimension d_o.

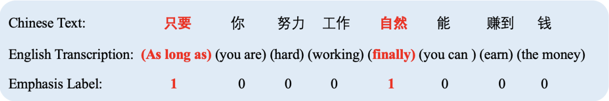
For emphasis detection, considering the various individual characteristics of speakers, the prominence in speech may present with different property in different representation subspaces. Using multi-head attention mechanism can enhance the model ability in capturing emphasis, and improve the robustness of the proposed approach.

III. Experiments

A. Experimental setup

Dataset. A well-organized emphatic speech corpus was employed to assess the performance of the proposed approach. 500 text prompts, each of which contains one or more emphatic words at different positions, were carefully designed to cover all kinds of pronunciation mechanisms and context characteristics of phonemes. Then, a professional native Mandarin female speaker was instructed to record the above emphatic speech data according to the text prompts and the emphasis labels. The word boundary was automatically labeled with an HMM-based forced alignment tool and manually corrected. Fig.4 depicts an example of emphasis labeling.

Features. The raw low-level descriptors (LLDs) extracted from speech signals contain three kinds of features: duration, fundamental frequency (F0), energy. Four statistical functions are then used to aggregate the frame-level F0 and energy:



Chinese Text:	**只要**	你	努力	工作	**自然**	能	赚到	钱
English Transcription:	**(As long as)**	(you are)	(hard)	(working)	(**finally)**	(you can)	(earn)	(the money)
Emphasis Label:	**1**	0	0	0	**1**	0	0	0

Fig. 4. An example of text prompts and corresponding labels.

(i) mean, (ii) maximum, (iii) standard deviation (SD), and (iv) range, resulting in 4-dimensional aggregated F0 related features, 4-dimensional aggregated energy related features, and 1-dimensional duration feature. 9-dimensional aggregated acoustic features are thus used for each word in utterances. Specially, the minimum value is not included following the suggestion in [6]. In this work, the sampling rate of speech signal is set to 16 kHz, and the TUM's open-source openSMILE [18] feature extractor is used for LLDs feature extraction, with 25 msec frame window length and 5 msec frame intervals.

Comparison methods. Besides the proposed CB-SA model, state-of-the-art emphasis detection approaches with reported performance are collected to compare with the proposed approach, including SVM based approach and the temporal memory enhanced recurrent neural network LSTM and BLSTM based approaches. The combination use of convolution bank and BLSTM (CB-BLSTM) is also implemented to analyze the contribution of convolution bank in emphasis detection, which contains stacked convolution bank layers, a BLSTM layer, and the classifier.

Evaluation metrics. In all the experiments, we evaluate the detection performance in terms of Accuracy, Precision, Recall and F1-measure [19]. The emphasis corpus is divided with a proportion of 3:1:1 for training, validation, and testing. All experimental results are based on 5-fold cross-validation.

Hyperparameters. In the proposed framework and the comparisons, M=2 stacked convolution bank layers and L=2 stacked self-attention layers are employed for feature learning. More layers may cause the overfitting problem in our experience. Convolutional layers in convolution bank have $\{k = 1, 2, 3, 4\}$ kernel size respectively, each contains 32 filters. The head number r employed in multi-head self-attention is set to 4. The emphasis classifier is constructed with 3 fully-connected layers, containing 64,16,2 units respectively. Adam [20] optimization algorithm is employed to train all the implementations, with an initial learning rate at 0.001.

B. Experimental results

Table I lists the performance of emphasis detection using different comparison methods. From the results, we can see that the performance of LSTM is better (in terms of F1-measure) than SVM, indicating contextual dependencies are important. And when both past and future contexts are considered (for BLSTM), the performance can be further improved (from 0.8074 to 0.8256 in terms of F1), which means bidirectional contextual dependencies are useful for our task.

Furthermore, comparing CB-BLSTM with BLSTM, both precision and recall of the former model is better indicating the effectiveness of convolution bank in modeling the local context. When self-attention mechanisms are considered, the CB-SA can achieve further performance improvement in emphasis detection (from 0.8473 to 0.8549 in terms of F1) compared with CB-BLSTM.

TABLE I
RESULTS OF USING DIFFERENT COMPARISON METHODS ON EMPHASIS TEST SET.

Method	Accuracy	Precision	Recall	F1
SVM	90.08%	81.41%	77.88%	0.7961
LSTM	90.48%	81.12%	80.37%	0.8074
BLSTM	91.57%	84.78%	80.46%	0.8256
CB-BLSTM	92.45%	85.04%	**84.44%**	0.8473
CB-SA	92.91%	**86.97%**	84.05%	**0.8549**

TABLE II
PERFORMANCE COMPARISON OF USING DIFFERENT FILTER KERNEL SIZE.

kernel size	Accuracy	Precision	Recall	F1
k=1	91.62%	84.43%	81.19%	0.8278
k=1,2	91.96%	84.57%	82.67%	0.8361
k=1,2,3	**92.91%**	**86.97%**	**84.05%**	**0.8549**
k=1,2,3,4	92.37%	85.71%	83.04%	0.8435

C. Discussion

1) Analysis on convolution bank: The convolution bank contains K sets of 1-D convolutional filters, where the k-th convolutional layer is constructed with filters of kernel size k. ($k = 1, 2, ..., K$). The number of sets (K) affects the model performance. Shown in Table II, with the growing of K, the performance of the model is improved gradually at first and then tends to decrease. Probably because the emphasis is closely related to the nearest words and the word itself, the influence of words further apart are relatively small. Besides, the overfitting problem caused by the limited training data may explain the decrease. Hence we use the convolution bank containing K=3 sets of 1-D convolutional filters to compare with the baseline methods.

2) Analysis on self-attention: Self-attention mechanism is analyzed on an example shown in Fig.4, the emphasis is placed on the 1st word which means "as long as" and the 5th word which means "finally". Fig.5 and Fig.6 depict the attention weights of the last self-attention layer when the number of heads r=1 and r=4 respectively.

Compared with BLSTM, attention mechanism can learn to focus on more relevant information. As Fig.5 shows, the 1st column, 5th column, and last column are darker than the others. And the 1st word and 5th word are successfully predicted as emphasized words, while the last word is misjudged as emphasis. As depicted in each row of Fig.5, it is interesting to find that the neutral words tend to attend to the surrounding emphasized words, which means except the word itself (residual connection), the surrounding emphasized words affect the representation of current words greatly. Fig.6 consists of 4 subplots, representing attention weights of different heads. Each subplot attends to different columns and is finally combined to make the final decision: the 1st word and 5th word are perceived as emphasized words. Comparing Fig.6 and Fig.5, multi-head attention (4 heads) helps the model to focus on information from different representation subspaces for improving the robustness of the proposed approach.

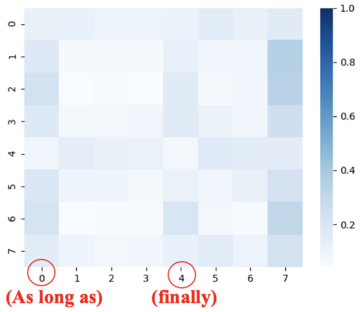


Fig. 5. Attention weights (1 head) of the last self-attention layer. X-coordinate and Y-coordinate represent the time step of keys and queries respectively.

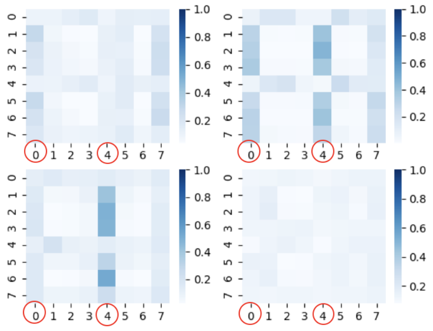


Fig. 6. Attention weights (4 heads) of the last self-attention layer

IV. Conclusions

This paper proposed a CB-SA model for emphasis detection in speech. We used the convolution bank to explicitly model local context from neighboring words, extracting the local prominence that is indicative of emphasis. Self-attention was employed to model the global dependencies between each word and the whole utterance. Analysis showed that the emphasized words will often affect the representation of other words in the utterances greatly. Experimental results demonstrated the effectiveness of our proposed method. Future work will explore the combinition of linguistic features and acoustic features for emphasis detection.

Acknowledgments

This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027), and National Social Science Foundation of China (13&ZD189).

References

[1] T. Mishra, V. R. Sridhar, and A. Conkie, “Word prominence detection using robust yet simple prosodic features,” in Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[2] Y. Ning, Z. Wu, R. Li, J. Jia, M. Xu, H. Meng, and L. Cai, “Learning cross-lingual knowledge with multilingual blstm for emphasis detection with limited training data,” i n Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5615–5619.

[3] L. S. Kennedy and D. P. Ellis, “Pitch-based emphasis detection for characterization of meeting recordings,” in 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721). IEEE, 2003, pp. 243–248.

[4] A. Rosenberg and J. Hirschberg, “Detecting pitch accents at the word, syllable and vowel level,” in Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009, pp. 81–84.

[5] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, “Hmm-based emphatic speech synthesis using unsupervised context labeling,” in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[6] S., Kakouros, O. Räsänen, and P. Alku. “Evaluation of spectral tilt measures for sentence prominence under different noise conditions,” Proc. Interspeech 2017, pp. 3211–3215, 2017.

[7] F. Tamburini, “Prosodic prominence detection in speech,” in Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on, vol. 1. IEEE, 2003, pp. 385–388.

[8] A. Schnall and M. Heckmann, “Integrating sequence information in the audio-visual detection of word prominence in a humanmachine interaction scenario,” in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[9] Y. Ning, Z. Wu, X. Lou, H. Meng, J. Jia, and L. Cai, “Using tilt for automatic emphasis detection with bayesian networks,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[10] Y. Ning, J. Jia, Z. Wu, R. Li, Y. An, Y. Wang, and H. Meng, “Multitask deep learning for user intention understanding in speech interaction systems,” in AAAI, 2017.

[11] J. Terken and D. Hermes, “The perception of prosodic prominence,” in Prosody: Theory and experiment. Springer, 2000, pp. 89–127.

[12] D. B. Pisoni and R. E. Remez, The handbook of speech perception. Wiley Online Library, 2005.

[13] Y. Ning, Z. Wu, J. Jia, F. Meng, H. Meng, and L. Cai, “Hmm-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4934–4938.

[14] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and et al, “Tacotron: A fully end-to-end text-to-speech synthesis model,” CoRR abs/1703.10135, 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 835–838.

[19] D. Powers, “Evaluation: from precision, recall and Fmeasure to ROC, informedness, markedness and correlation”. Journal of Machine Learning Technologies Vol. 2(1), pp. 37-63, 2011.

[20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, pp. 18–25, 2014.