Emphasis Detection for Voice Dialogue Applications Using Multi-channel Convolutional Bidirectional Long Short-Term Memory Network

Long Zhang[1], *Jia Jia*[1*], *Fanbo Meng*[3], *Suping Zhou*[1], *Wei Chen*[3], *Cunjun Zhang*[1,3], *Runnan Li*[1],

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
[2]Academy of Arts & Design, Tsinghua University, China
[3]Beijing Sogou Technology Co. Ltd., Beijing, China

zhanglon16@mails.tsinghua.edu.cn, jjia@mail.tsinghua.edu.cn

Abstract

Emphasis detection is important for user intention understanding in human-computer interaction scenario. Techniques have been developed to detect the emphatic words in speech, but challenges still exist in Voice Dialogue Applications (VDAs): the tremendous non-specific speakers and their various expressions. In this work, we present a novel approach to automatically detect emphasis in VDAs by using multi-channel convolutional bi-directional long short-term memory neural networks (MC-BLSTM), which can learn various expressions of large amounts of speakers and long span temporal dependencies across speech trajectories. In particular, we first use a multi-channel convolutional component in the proposed approach to extract high-level representation of input acoustic features for emphasis detection. The experimental results on a 3400 real-world dataset collected from Sogou[1] Voice Assistant outperform current state-of-the-art baseline systems (+6.2% in terms of F1-measure on average).

Index Terms: emphasis detection, human-computer interaction, voice dialogue applications, multi-channel convolutional bi-directional long short-term memory neural networks

1. Introduction

With the rapid development of technology, the Voice Dialogue Applications (Siri[2], Alexa[3], Cortana[4], etc.) have gained great interest recently [1]. According to the statistics from Nuance[5], about 73% of people prefer improved communication experiences through the smarter technology, which can help them understand the dialogue intention better. Emphasis is an important factor to convey speaker's attitudes, intentions and emotions [2]. Detecting emphasis can help other interlocutors capture the important paralinguistic information of utterance in dialogue [3]. It has also attracted considerable attention in the past few years due to their potential application in the field of speech-to-speech translation [4, 5], emphatic speech synthesis [6], automatic prosodic event detection [7, 8], human-computer interaction, etc.

Previous researches on emphasis detection have focused on the acoustic features and models perspectives: Ferrer *et al.* [9] used filtered energy features to detect pitch accents. Meng *et al.* [10] considered this task as a classification problem and the emphasis is divided into six classes. Schnall *et al.* [11] used support vector machines (SVM) and conditional random fields (CRF) to predict word prominence in spontaneous speech. Cernak *et al.* [12] used a probalilistic amplitude demodulation (PAD) method for prosodic event detection. Cernak *et al.* [13] devised a sound pattern matching method for automatic prosodic event detection. Do *et al.* [14] used linear regression HSMMs method (LR-HSMMs) for preserving word-level emphasis. Ning *et al.* [15] propose a multilingual BLSTM model to detect emphasis. Although the methods have shown significant performance improvements for emphasis detection, it is still be challenged in real-world voice dialogue applications(VDAs) attributed to three main factors: The large amounts of non-specific speakers, various expression and the long span temporal information including long-distance dependencies across speech trajectories. There are tremendous amounts of users in VDAs, which are non-specific speakers with different gender, age and geographical characteristics, bringing in a great diversity of users dialects and expression preferences. This diversity increases the difficulty of inferring users emphasis.

To solve this problem, we introduce a novel approach to detect emphasis in VDAs by using multi-channel convolutional bi-directional long short-term memory neural networks (MC-BLSTM), which can learn tremendous non-specific speakers' expressions of emphasis and long span temporal dependencies across speech trajectories (shown in Figure 1). In particular, we employ a multi-channel convolutional neural network component in the proposed approach to extract high-level representation of acoustic features (fundamental frequency (F0), mel frequency cepstral coefficients (MFCCs), energy, duration and position of frame (POF)). Bi-directional long short-term memory neural network is recurrent neural network which is capable of storing information over extended time intervals. Our approach results on a 3400 real-world dataset collected from Sogou Voice Assistant outperform current state-of-the-art baseline systems (+6.2% in terms of F1-measure [16] on average).

The organization of this paper is as follows: the multi-channel convolutional bi-directional long short-term memory neural networks (MC-BLSTM) architectures are described in section 2. To evaluate the performance of our approach, experiments were conducted. The results and analysis are presented in Section 3. Finally, Section 4 gives a summary and conclusion of this work.

[1]http://yy.sogou.com
[2]https://www.apple.com/cn/ios/siri/
[3]https://developer.amazon.com/alexa/
[4]http://www.microsoft.com/zh-cn/windows/cortana/
[5]https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/conversational-ivr.html

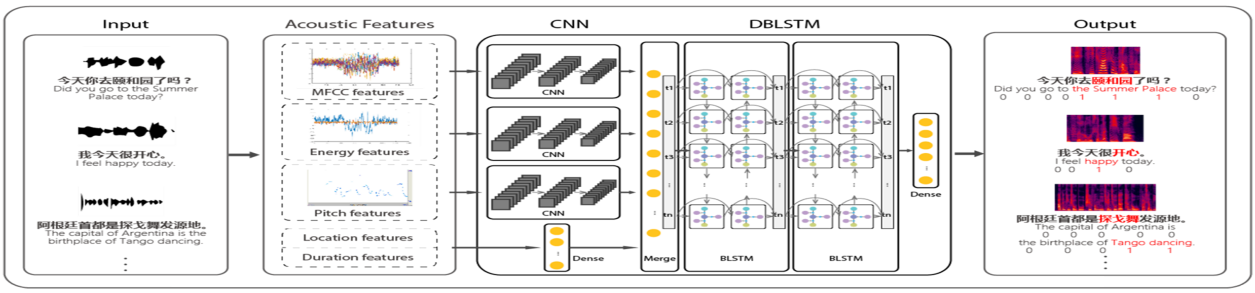

Figure 1: *A graphical overview of our emphasis detection training line: The rectangles in the input image sequence shows the different acoustic features. Here we show the extraction of fundamental frequency (F0), MFCCs and energy. Representation extracted from input features is then used as input for BLSTM network for further prediction.*

2. Methodology

We propose a multi-channel convolutional bi-directional long short-term memory neural networks (MC-BLSTM) model to detect emphasis in VDAs, which can learn various expressions of tremendous non-specific speakers and long span temporal dependencies across speech trajectories[17]. In particular, each individual has his own characteristics, thus we employ a multi-channel convolutional neural network component to extract high-level representation from input features for enhancing the performance of the proposed approach.

2.1. Multi-channel convolutional neural network component

Each of the acoustic features (F0, energy, MFCCs) has its specific representation in VDAs. For example, [4] has compared the difference between the duration and energy features in emphasis prosodic. Therefore, we employ a multi-channel convolutional neural network component to deal with acoustic features (F0, energy, MFCCs) respectively. The overall structure of multi-channel convolutional neural network component [18, 19] is illustrated in Figure 2.

Let $\boldsymbol{s_i}$ be the d-dimensional fundamental frequency (F0) feature for the i-th frame in a sentence, $\boldsymbol{s} \in \mathbb{R}^{L\times d}$ denote the sentence with L frames, k be the lenth of filter vector $\boldsymbol{m} \in \mathbb{R}^{k\times k}$. To extract k-gram features, we design a window vector $\boldsymbol{w} \in \mathbb{R}^{k\times k}$ with k consecutive vectors.

The idea behind convolution is to take the element-wise multiplication of filter vector $\boldsymbol{m}$ with each window vector $\boldsymbol{w}$ in the sentence $\boldsymbol{s}$ to obtain a feature map $\boldsymbol{y} \in \mathbb{R}^{L\times d}$, where each element $\boldsymbol{y}$ is produced as:

$$\boldsymbol{y} = f(\boldsymbol{w} \circ \boldsymbol{m} + b) \tag{1}$$

where $\circ$ is a convolutional operator, b respectively denotes bias term and f is nonlinear transformation function. In our case, we follow the work in [20] to choose ReLU as the nonlinear function. We use multiple filter vectors to generate different feature maps and then concatenate them together to produce new features. Let n be the numbers of filter vectors, we have:

$$\mathbb{Y} = [\boldsymbol{y}_1; \boldsymbol{y}_2; ...; \boldsymbol{y}_n] \tag{2}$$

Semicolons represent column concatenation and $\boldsymbol{y_i}$ is the feature map generated by the i-th filter. Each row of $\mathbb{Y} \in \mathbb{R}^{L\times d}$ is the new higher-level feature representation. Similarly, we perform convolution on energy and MFCCs features. Because the F0 and the energy are one-dimensional feature, while the

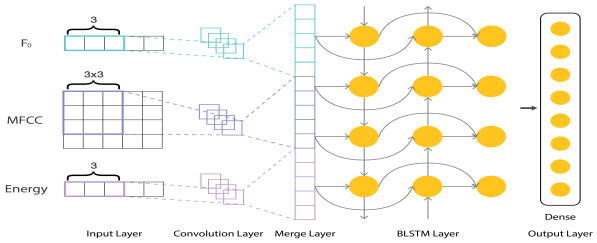

Figure 2: *The structure of multi-channel convolutional bi-directional Long Short-Term Memory Neural Networks.*

MFCCs is a matrix feature, the 1D convolutional processing is used for the F0 and the energy and the 2D convolutional processing is used for the MFCCs respectively. Different from $\mathbb{Y}$, $\mathbb{E}$ and $\mathbb{M}$ denotes the the new features of energy and MFCCs instead of $\mathbb{Y}$. All the features are merged together to generate $\mathbb{C}$ as followes:

$$\mathbb{C} = concat[\mathbb{Y}, \mathbb{E}, \mathbb{M}] \tag{3}$$

The output hiddens $\mathbb{C}$ is then feeded to the BLSTM component [21, 19] for further computation directly.

2.2. Bi-directional long short-term memory neural network

In order to learn long span temporal information, recurrent neural network architecture with bi-directional long short-term memory units is then employed in the proposed framework.

Given an input sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$, Recurrent Neural Network (RNN) computes the hidden state vector sequence $\boldsymbol{h} = (h_1, \ldots, h_T)$ and outputs vector sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$ from $t = 1$ to T by iterating the following equations:

$$h_t = \mathcal{H}(\boldsymbol{W}_{xh}x_t + \boldsymbol{W}_{hh}h_{t-1} + b_h) \tag{4}$$

$$y_t = \boldsymbol{W}_{hy}h_t + b_y \tag{5}$$

where $\boldsymbol{\mathcal{H}}$ is activation function, $\boldsymbol{W}$ is the weight matrix(e.g., $\boldsymbol{W}_{hy}$ is the hidden-output weight vectors), b is the bias vectors, b_h is the bias vector for hidden state vector and b_y is the bias vector for output vector.

Furthermore, in order to make full use of speech sequences in the forward and backward directions, Bidirectional RNN[22] separates the hidden layer into forward frame sequence $\overrightarrow{h}$ and backward frame sequence $\overleftarrow{h}$. The iterative process is as

follows[23]:

$$\overrightarrow{h}_t = \mathcal{H}(\boldsymbol{W}_{x\overrightarrow{h}} x_t + \boldsymbol{W}_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t+1} + b_{\overrightarrow{h}}) \quad (6)$$

$$\overleftarrow{h}_t = \mathcal{H}(\boldsymbol{W}_{x\overleftarrow{h}} x_t + \boldsymbol{W}_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (7)$$

$$y_t = \boldsymbol{W}_{\overrightarrow{h}y} \overrightarrow{h}_t + \boldsymbol{W}_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \quad (8)$$

$\boldsymbol{\mathcal{H}}$ is usually a sigmoid or hyperbolic tangent function in the conventional RNN models[24], which leads to the limitations of storing past and future information in speech. For bidirectional long short-term memory (BLSTM)[25], which build a memory cell inside, can overcome the problems in conventional models. The $\boldsymbol{\mathcal{H}}$ of BLSTM is implemented with the following functions[26]:

$$i_t = \sigma(\boldsymbol{W}_{xi} x_t + \boldsymbol{W}_{hi} h_{t-1} + \boldsymbol{W}_{ci} c_{t-1} + b_i) \quad (9)$$

$$f_t = \sigma(\boldsymbol{W}_{xf} x_t + \boldsymbol{W}_{hf} t-1 + \boldsymbol{W}_{cf} c_{t-1} + b_f) \quad (10)$$

$$c_t = f_t c_{t-1} + i_t tanh(\boldsymbol{W}_{xc} x_t + \boldsymbol{W}_{hc} h_{t-1} + b_c) \quad (11)$$

$$o_t = \sigma(\boldsymbol{W}_{x0} x_t + \boldsymbol{W}_{ho} h_{t-1} + \boldsymbol{W}_{co} c_t + b_o) \quad (12)$$

$$h_t = o_t tanh(c_t) \quad (13)$$

Combining the advantages of Multi-channel Convolutional Neural Network component and BLSTM, which can extract locality and instability features[19, 27] and make the best of long-range context in both forward and backward directions.

3. Experiments

3.1. Corpus

We evaluated our proposed method on the corpus of voice data from Sogou Voice Assistant containing 12,000 utterances recorded by 706 users. We randomly selected 3,400 utterances from the database and labeled the emphasis for each utterance. The estimated word-level emphasis is then classified into labels of 0 and 1 indicating normal and emphasized words. The corpus is labeled by three well-trained annotators. Labels are regarded as emphasis only when three inter-annotator reach an agreement. If they are controversial or ambiguous about labels, utterance will be labeled as ambiguous or discarded. Finally, 3,093 utterances are labeled emphasis as frame level. Each of the utterances contains one or more emphatic words. These emphatic words are located at different positions in sentences. The emphasis distributions of these utterances are: emphasis: 17.03%, normal: 82.97%. An example of the label sentences are shown in Figure 3.

3.2. Features

In this experiment, voice segments of each utterance are sampled with 5ms frame shift and 25ms frame length. Numerical features are normalized to the range of (0, 1]. The acoustic features we used are extracted by librosa [28], including the following features:

English:	Did	you	go	to	the	Summer	Palace	today?	
English labels:	0	0	0	0	0	1	1	0	
Chinese:	今	天	你	去	颐	和	园	了	吗？
	Jin	tian	ni	qv	yi	he	yuan	le	ma?
Chinese labels:	0	0	0	0	1	1	1	0	0?

Figure 3: *An example of labels in English and Chinese sentences from the VDAs.*

- **lf0**: Log F0 (lf0) feature and related features(mean, minimum, maximum, and the range of lf0).
- **Energy**: Energy feature and related features (mean, minimum, maximum, and the range of energy).
- **MFCCs**: MFCCs (12-demensional) features, delta (12-demensional) and acceleration (12-demensional) of MFCCs features. These 36 MFCCs features have been normalization for acoustic models.
- **Duration**: The duration of a word.
- **Position of Frame (PoF)**: The position of the syllables in the sentence, the position of the frame in syllable and the position of the frame in sentence.

3.3. Experimental setup

3.3.1. Comparison methods

We compared the performance of emphasis detection with some well-known machine learning methods, including support vector machine (SVM) [29], bayesian network (BN)[30], deep neural network (DNN) and convolutional neural network (CNN). We also designed some kinds of LSTM models for comparison, bi-directional long short-term memory (BLSTM)[15], convolutional bidirectional long short-term memory (C-BLSTM)[19], the proposed model of the multi-channel convolutional bi-directional long short-term memory neural networks(MC-DBLSTM).

3.3.2. Network setup

In implementation of comparisons, we employ four convolutional layers for each CNN based component: for F0 and energy processing, the filters employed in the 1D convolutional layers are [32, 32, 16, 1] respectively; for MFCC processing, the filters employed in the 2D convolutional layers are [64, 128, 32, 1]. Specially, the kernel size and stride of the proposed 1D convolutional component are 3 and 1, while in 2D convolutional component are 3×3 and 1×1, respectively. ReLU was applied as the activation function. In BLSTM component, one hidden layer contains one forward LSTM layer and one backward LSTM layer [31] with 64 units. Specially, for LSTM based comparisons, the number of LSTM blocks are expanded to match the similar scale. As the convolution layer requires fixed-length input, the $maxlen$ of input sentence is set to be 2000 frames after statistics analysis of the dataset. For sentences with fewer frames, zero padding zero is applied.

3.3.3. Evaluation metrics

In all the experiments, we evaluate the performance in terms of F1-measure[32, 11], Precision, Recall. The data corpora [33] are split by train:val:test = 8:1:1.

3.4. Experimental results

3.4.1. Performance Comparision

We selected six models as baseline to compare the performance of emphasis detection: BN, DNN, CNN, LSTM, BLSTM, C-BLSTM, MC-BLSTM. Table 1 shows the results. The proposed model outperforms all the baseline methods: +23.8% compared with BN, +22.3% compared with DNN, +20.7% compared with CNN, +17.5% compared with LSTM, +12.6% compared with BLSTM, and +6.2% compared with C-BLSTM. By the results of several comparisons, we can see (in terms of F1-measure),

Table 1: *The Precision, Recall and F1-Measure in different models.*

Method	Precision	Recall	F1-measure
BN	0.462	0.272	0.343
DNN	0.391	0.330	0.358
CNN	0.412	0.342	0.374
LSTM	0.463	0.313	0.406
BLSTM	0.457	0.468	0.455
C-BLSTM	0.481	0.562	0.519
MC-BLSTM	0.520	0.658	0.581

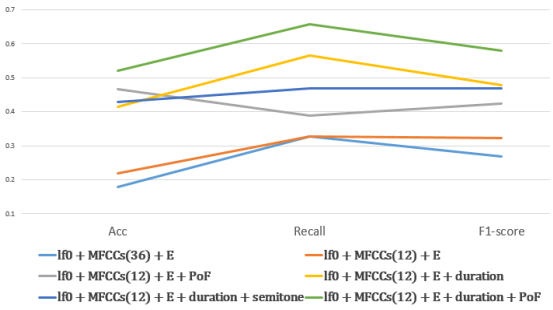

Figure 4: *different acoustics features of emphasis detection.*

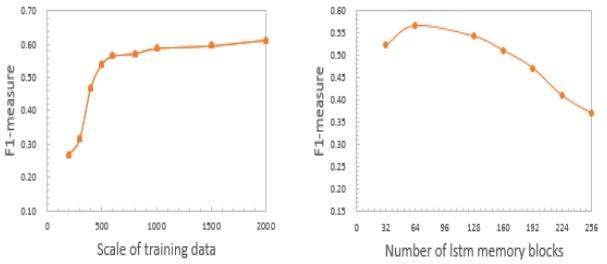

Figure 5: *the number of emphasis detection datasets and the number of lstm memory blocks.*

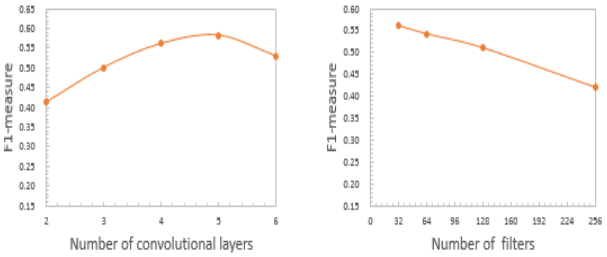

Figure 6: *the number of convolutional layers and filters.*

the MC-BLSTM model is capable of learning the tremendous speaker-independent representation of emphasis and and long span temporal dependencies in VDAs. As shown in Table 1, we find that the novel method has the best performance. These proof that our proposed MC-BLSTM method does have an effective impact on the real-world voice dataset. Moreover, it verifies that the multi-channel convolutional component is more powerful to extract high-level representation of input acoustic features.

To demonstrate the comparability and the adaptability of our method, we also report experimental results on previously examined Database[2]. As shown in Table 1, the accuracy reaches 0.581, showing +12.6% improvement compared with [15], +23.8% improvement compared with[30], indicating that our method still shows advantages on the acted database and utterances of other language.

3.4.2. Feature contribution analysis

We validate that F0, energy(E), MFCCs, duration and PoF are more powerful features to estimate emphasis. As we can see in the Figure 4, the performance of 'lf0 + energy + MFCCs (12)' features is better than 'lf0 + energy + MFCCs (36)'. We conclude that a large amount of redundant information (such as the delta and acceleration MFCCs features) will bring interference to the detection and reduce the recognition for the emphasis. In addition, location (PoF) features also play an important role in emphasis detection.

3.4.3. Scalability

As shown in Figure 5, with the increase of the amount of training data, the performance tends to be stable and the trend of growth is slow. With the increase of the amount of data, F1-score performance with rapid ascension, but when the number of data achieve 500, performance growth slowing. As the data volume increases to 2000, this trend remains the same. So we choose 500 as the main test data quantity of the data.

3.4.4. Parameter sensitivity analysis

We compared the effects of different numbers of LSTM memory blocks, multi-channel convolutional layers and convolutional filters, as shown in Figure 5 and Figure 6. The results of the experiment are as follows: 1) The performance reached the highest performance when the number of LSTM memory blocks is 64. With the increase of the number of the blocks, the performance decreased. Therefore, we choose 64 as the number of LSTM memory blocks in the experiments; 2) The four and five convolutional layers both show good performances in emphasis detection. However, the four layer networks reduces the time consumption compared to the five layer networks, so we choose the four convolutional layers as the experimental setup; 3) For the convolutional filters (first layer of F0), the performance reached the highest when the number is 32. Therefore, we choose 32 as the number of filters in the experiments. Other parameters are shown in section **3.3**.

4. Conclusions

In this paper, we proposed a multi-channel convolutional bi-directional long short-term memory neural networks (MC-BLSTM) which learns various expressions of the tremendous non-specific speakers and long span temporal dependencies across speech trajectories in Voice Dialogue Applications (VDAs). In particular, we employ a multi-channel convolutional component in the proposed approach to extract high-level representation of input features. The experimental results on a 3400 real-world dataset collected from Sogou Voice Assistant outperform current state-of-the-art baseline systems (+6.2% in terms of F1-measure on average).

5. References

[1] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," 2018.

[2] Y. An, Y. Wang, and H. Meng, "Multi-task deep learning for user intention understanding in speech interaction systems," 2017.

[3] Q. T. Do, S. Sakti, S. Nakamura, Q. T. Do, S. Sakti, S. Nakamura, Q. T. Do, S. Sakti, and S. Nakamura, "Toward expressive speech translation: A unified sequence-to-sequence lstms approach for translating words and emphasis," in *INTERSPEECH*, 2017, pp. 2640–2644.

[4] Q. T. Do, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 544–556, 2017.

[5] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8396–8400.

[6] R. Li, Z. Wu, X. Liu, H. Meng, and L. Cai, "Multi-task learning of structured output layer bidirectional lstms for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5510–5514.

[7] 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.

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

[9] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Lexical stress classification for language learning using spectral and segmental features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 7704–7708.

[10] F. Meng, H. Meng, Z. Wu, and L. Cai, "Synthesizing expressive speech to convey focus using a perturbation model for computer-aided pronunciation training," in *Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.

[11] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] M. Cernak and P.-E. Honnet, "An empirical model of emphatic word detection," Idiap, Tech. Rep., 2015.

[13] M. Cernak, A. Asaei, P.-E. Honnet, P. N. Garner, and H. Bourlard, "Sound pattern matching for automatic prosodic event detection," in *Interspeech*, no. EPFL-CONF-218851, 2016.

[14] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression hsmms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] 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," in *ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5615–5619.

[16] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize f1 measure," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 225–239.

[17] S. Sudhakaran and O. Lanz, "Convolutional long short-term memory networks for recognizing first person interactions," *arXiv preprint arXiv:1709.06495*, 2017.

[18] J. Bouvrie, "Notes on convolutional neural networks," 2006.

[19] L. Li, Z. Wu, M. Xu, H. M. Meng, and L. Cai, "Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition." in *INTERSPEECH*, 2016, pp. 1392–1396.

[20] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," *arXiv preprint arXiv:1511.08630*, 2015.

[21] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional cnn-lstm model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2016, pp. 225–230.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[23] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[24] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," in *IEEE International Joint Conference on Neural Networks, 2005. IJCNN '05. Proceedings*, 2005, pp. 2047–2052 vol. 4.

[26] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[27] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *arXiv preprint arXiv:1511.08308*, 2015.

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[29] C.-c. Chang and H. Lin, "A library for support vector machines," 2007.

[30] Y. Ning, Z. Wu, X. Lou, H. Meng, J. Jia, and L. Cai, "Using tilt for automatic emphasis detection with bayesian networks," in , 2015.

[31] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4869–4873.

[32] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," in *Spoken Language Technology Workshop*, 2013, pp. 210–215.

[33] B. Wu, J. Jia, T. He, J. Du, X. Yi, and Y. Ning, "Inferring users' emotions for human-mobile voice dialogue applications," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.