Inferring Emphasis for Real Voice Data: An Attentive Multimodal Neural Network Approach

Suping Zhou[1], Jia Jia[1(✉)], Long Zhang[2], Yanfeng Wang[3], Wei Chen[3], Fanbo Meng[3], Fei Yu[2], and Jialie Shen[4]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
1874504489@qq.com, jjia@mail.tsinghua.edu.cn
[2] The Spectrum Division of China Electronic Equipment System Engineering Company, Beijing, China
976890413@qq.com, yfei0210@163.com
[3] Sogou Corporation, Beijing, China
{wangyanfeng,chenweibj8871,mengfanbosi0935}@sogou-inc.com
[4] Queen's University Belfast, Belfast, UK
j.shen@qub.ac.uk

Abstract. To understand speakers' attitudes and intentions in real Voice Dialogue Applications (VDAs), effective emphasis inference from users' queries may play an important role. However, in VDAs, there are tremendous amount of uncertain speakers with a great diversity of users' dialects, expression preferences, which challenge the traditional emphasis detection methods. In this paper, to better infer emphasis for real voice data, we propose an attentive multimodal neural network. Specifically, first, beside the acoustic features, extensive textual features are applied in modelling. Then, considering the feature in-dependency, we model the multi-modal features utilizing a Multi-path convolutional neural network (MCNN). Furthermore, combining high-level multi-modal features, we train an emphasis classifier by attending on the textual features with an attention-based bidirectional long short-term memory network (ABLSTM), to comprehensively learn discriminative features from diverse users. Our experimental study based on a real-world dataset collected from Sogou Voice Assistant (https://yy.sogou.com/) show that our method outperforms (over 1.0–15.5% in terms of F1 measure) alternative baselines.

Keywords: Emphasis detection · Voice dialogue applications · Attention

© Springer Nature Switzerland AG 2020
Y. M. Ro et al. (Eds.): MMM 2020, LNCS 11962, pp. 52–62, 2020.
https://doi.org/10.1007/978-3-030-37734-2_5

1 Introduction

With the rapid development of technology, the Voice Dialogue Applications (Siri[1], Nina[2], Alexa[3], etc.) have gained popularity in recent years. Emphasis plays an important role in conveying speaker's attitudes and intentions in VDAs. Meanwhile, emphasis detection also attracts considerable attention in the field of speech-to-speech translation, emphatic speech synthesis, automatic prosodic event detection, human-computer interaction [2,18].

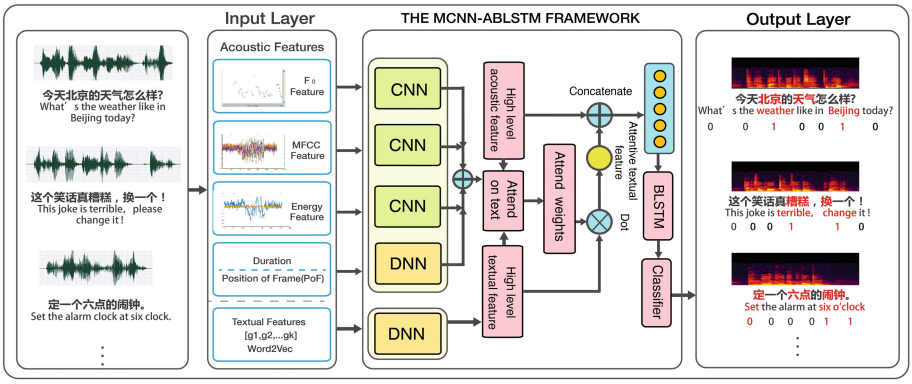

Fig. 1. The workflow of our framework.

Although there have amounts of attempts on emphasis detection, fulfilling the task is still a non-trivial issue. Traditionally, emphasis mainly detected utilizing acoustics information. Kennedy *et al.* [12] propose the use of pitch features as its only acoustic predictor. Ferrer *et al.* [9] uses filtered spectral and segmental features to detect emphasis for each syllable in a word. Although some work also try to utilize both acoustic and textual features [2] for emphatic words detection, they are mainly done on acted corpora data.

Therefore, there remains two challenges unsolved for emphasis detection in the specific situation of real-world VDAs: (1) Except speech information, the speech-to-text information is also provided by VDAs. Can we integrate multiple modalities (speech and text) to help enhance the performance on inferring emphasis? (2) Distinguished from the traditional speech emphasis recognition methods based on acted labeled data, the tremendous amount of uncertain speakers bring in a great diversity of users' dialects and expression preferences. Therefore, how to comprehensively learn user-invariant features by strengthening single modal features to increase the emphasis inferring effectiveness?

[1] https://www.apple.com/cn/ios/siri/.
[2] https://www.nuance.com/index.html.
[3] https://developer.amazon.com/alexa/.

To solve this problem, we introduce a novel approach to detect emphasis in VDAs with extra attention on textual information. Figure 1 illustrates detail architecture. In particular, first we employ a Multi-path convolutional neural network (MCNN) component which considers the independency nature of features [21,22], to extract high-level representation of acoustic features (fundamental frequency (F0), Mel Frequency Cepstral Coefficients (MFCCs), energy, duration and position of frame (POF) [21]) and high-level textual features individually. Then, combining both high-level acoustic features and textual features, we train an emphasis classifier by attending on the textual features with an attention-based bidirectional long short-term memory network (ABLSTM). Our experimental study based on a 1500 real-world dataset collected from Sogou Voice Assistant demonstrate that our method outperforms baseline systems (over 1.0–15.5% in terms of F1 measure). Specifically, we discover that, the textual information enhances the performance for 2.6%, while attention mechanism further enhance the performance for 1.0%. Meanwhile, to demonstrate the adaptability of our method, we also conduct experiments on a 500 real-world English corpus. Our method easily adapts to utterances of other language and outperforms baseline systems (over 0.8–14.4% in terms of F1 measure).

The organization of this paper is as follows: Sect. 2 lists related works. Section 3 presents the methodologies. Section 4 introduces the experiments and results. Section 5 is the conclusion.

2 Related Work

Emphasis Detection Methods. Previous researches on emphasis detection have focused on the features and models perspectives: Ladd *et al.* [15] utilize fundamental frequency to analysis 'normal' and 'emphatic' accent peaks. Heldner *et al.* [11] and Ferrer *et al.* [9] uses spectral features to detect emphasis. In [2], fundamental frequencies, duration, spectral features, lexical features, and identity features are combined together to get a better performance in emphatic words detection. Meanwhile, some previous works have been done on modelling methods. Cernak *et al.* [3] used a probabilistic amplitude demodulation (PAD) method to predict word prominence in speech. Do *et al.* [6] used linear regression HSMMs method (LR-HSMMs) for preserving word-level emphasis. Ning *et al.* [18] propose a multilingual BLSTM model for prosodic event detection. However, these researches mainly focus on inferring emphasis from acted corpora, few have been done to address the problem in real-world VDAs.

Multi-media Modeling. Recently, methods in Multi-media Modeling have shown significant performance improvements. Zhou *et al.* [22] propose a Multi-path Generative Neural Network which consider both acoustic and textual features. Zhang *et al.* [21] propose a MCNN model for emphasis detection. Meanwhile, attention mechanism [1] is gaining its popularity. It have been proved to be effective in learning more attentive features for many areas like sentiment analysis [5]. Therefore, we suppose that these methods may also be helpful for multi-modal emphasis detection in VDAs.

3 Methodology

In this paper, to infer users' emphasis in VDAs. We propose a novel scheme for emphasis detection with extra attention on textual information. Specifically, (1) considering the in-dependency nature in features, we first model the acoustic and textual features utilizing a Multi-path convolutional neural network (MCNN) individually. (2) to comprehensively learn discriminative features from diverse users in VDAs, combining high-level multi-modal features, we train an emphasis classifier by attending on the textual features with an attention-based bidirectional long short-term memory network (ABLSTM).

3.1 Multi-path Convolutional Neural Network Component

As discussed above, traditionally, to model high-level features for emphasis analysis, SVM, CRF [19], HSMMs [7], DNNs, CNNs [14] have been adopted. However, since the different feature has its own characteristics, the traditional methods which utilize the low-level features like F0, MFCCs, energy, et al. as input together may not fully consider the independency nature of different features. These may limit the performance in emphasis detection while combining multi-features [21]. In our solution, considering both textual feature and six kinds of acoustic feature, we employ a multi-path convolutional neural network (MCNN) component to extract high-level representation from multi-modal features respectively to enhance the performance of our proposed approach.

Specifically, for F0, MFCCs and energy, we perform convolution on modelling. We define $\mathbb{F}$, $\mathbb{E}$, $\mathbb{M}$ as the high-level features representation for F0, energy, MFCCs. Let $\boldsymbol{s} \in \mathbb{R}^{L \times d}$ represents a L-frame sentence. For each frame, it has d-dimensional features. The convolution involves a filter $\boldsymbol{m} \in \mathbb{R}^{k \times k}$, which is applied to a window $\boldsymbol{w} \in \mathbb{R}^{k \times k}$ to produce a new feature $\boldsymbol{y} \in \mathbb{R}^{L \times d}$. Each feature $\boldsymbol{y}_i$ is produced as [13]:

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

b respectively denotes bias term and f is nonlinear transformation function. This filter is applied to each possible window in the sentence $\boldsymbol{s}$ to produce a feature map:

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

For duration, POE feature and textual features, we choose the DNNs to model their high-level representation which preforms well for these three features and can make our model more efficient. We define them as $\mathbb{D}$, $\mathbb{P}$ and $\mathbb{W}$. We train all these feature extractors together. Then, all the acoustic features are merged together to generate $\mathbb{A}$ as follows:

$$\mathbb{A} = concat[\mathbb{F}, \mathbb{E}, \mathbb{M}, \mathbb{D}, \mathbb{P}] \tag{3}$$

The output hiddens $\mathbb{A}$ and $\mathbb{W}$ for high-level acoustic and textual features are then fed to the ABLSTM component for further computation directly.

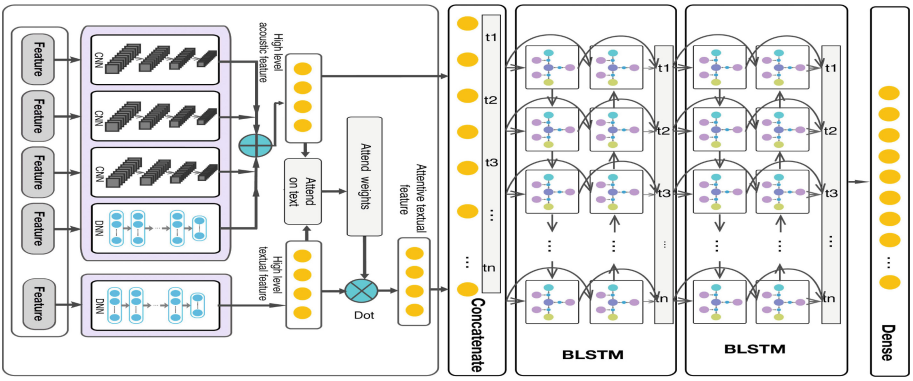

Fig. 2. The multi-path convolutional and attention-based bidirectional long short-term memory neural network (MCNN-ABLSTM)

3.2 Attention-Based BLSTM

Long short-term memory (LSTM) units have been extensively used to learn long span temporal information. In our proposed framework, we apply recurrent neural network architecture with bi-directional long short-term memory (BLSTM) to achieve effective modeling.

Sepecifically, we apply Bidirectional RNN [20] to make full use of speech sequences in the forward and backward directions. Given an input sequence $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, T is the length. $\overrightarrow{h}$ is forward hidden layer, and $\overleftarrow{h}$ is backward hidden layer. The iterative process is as follows [8]:

$$\overrightarrow{h}_t = \mathcal{H}(\boldsymbol{W}_{x\overrightarrow{h}} x_t + \boldsymbol{W}_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t+1} + b_{\overrightarrow{h}}) \tag{4}$$

$$\overleftarrow{h}_t = \mathcal{H}(\boldsymbol{W}_{x\overleftarrow{h}} x_t + \boldsymbol{W}_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{5}$$

$$y_t = \boldsymbol{W}_{\overrightarrow{h}y} \overrightarrow{h}_t + \boldsymbol{W}_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{6}$$

$\boldsymbol{y}$ is the outputs sequence and $\boldsymbol{W}$ is the weight matrix for different layers. b_h is the bias vector for hidden state vector and b_y is the bias vector for output vector. $\mathcal{H}$ is an activation function. For $\mathcal{H}$ in conventional RNN models, it has the limitations of storing past and future information in speech. The Bidirectional long short-term memory (BLSTM) with a memory cell built inside can overcome it. The $\mathcal{H}$ of BLSTM is as follows [10]:

$$i_t = \sigma(\boldsymbol{W}_{xi} x_t + \boldsymbol{W}_{hi} h_{t-1} + \boldsymbol{W}_{ci} c_{t-1} + b_i) \tag{7}$$

$$f_t = \sigma(\boldsymbol{W}_{xf} x_t + \boldsymbol{W}_{hf} x_{t-1} + \boldsymbol{W}_{cf} c_{t-1} + b_f) \tag{8}$$

$$c_t = f_t c_{t-1} + i_t tanh(\boldsymbol{W}_{xc} x_t + \boldsymbol{W}_{hc} h_{t-1} + b_c) \tag{9}$$

$$o_t = \sigma(\boldsymbol{W}_{x0} x_t + \boldsymbol{W}_{ho} h_{t-1} + \boldsymbol{W}_{co} c_t + b_o) \tag{10}$$

$$h_t = o_t tanh(c_t) \tag{11}$$

In order to take advantage of the textual information, we adopt the attention mechanism mentioned in [5] and modify the input layer into the BLSTM units. Let $\boldsymbol{a}_t$ represents the current high-level acoustic feature input and $\boldsymbol{w}_t$ be the current high-level textual feature input learned from MCNN. We first obtain the attentive textual feature $\boldsymbol{v}_t$ as an weighted average of the high-level textual feature $\boldsymbol{w}_t$ based on the self-selected attention mechanism. Let $\boldsymbol{w}_t$, $t \in (1, T)$ represent the t-th frame feature of textual feature $\boldsymbol{w}$ and $f_{att}(\cdot, \boldsymbol{a}_t)$ denote the attention function conditioned on the current high-level acoustic feature $\boldsymbol{a}_t$. The attention weight $\boldsymbol{\alpha}_i$ and attentive textual feature $\boldsymbol{v}_t$ is formulated as follows:

$$u_t = f_{att}(\boldsymbol{w}_t, \boldsymbol{a}_t) \tag{12}$$

$$\alpha_t = \frac{exp(u_t)}{\sum_{t=1}^{T} exp(u_t)} \tag{13}$$

$$\boldsymbol{v}_t = \alpha_t \cdot \boldsymbol{w}_t \tag{14}$$

We choose a fully-connected layer with ELU activation as the attention function, and the attention vector $\boldsymbol{v}_t$ is concatenated with the high-level acoustic feature $\boldsymbol{a}_t$ as the new input of the BLSTM. Thus the input vector $\boldsymbol{x}_t$ becomes $[\boldsymbol{a}_t, \boldsymbol{v}_t]$. The output of the final BLSTM unit is then fed into a fully-connected layer with softmax activation to predict emphasis results. Categorical cross-entropy loss is used as the objection function.

The motivation of this acoustic-guide Attention-based BLSTM (ABLSTM) as shown in Fig. 2 with the textual feature is that we use the acoustic feature to guide the attention weights of the textual feature in order to enforce the model to self-select which frame feature it should attend on. With this mechanism, it can help comprehensively learn discriminative features from diverse users in order to improve the emphasis detection accuracy in VDAs.

4 Experiments

4.1 Corpus and Annotation

Mandarin Corpus. We establish a real-world corpus of voice data from Sogou Voice Assistant containing 1500 Mandarin utterances recorded by 176 users. Every utterance is assigned with its corresponding speech-to-text information provided by Sogou Corporation.

English Corpus. We establish a corpus of voice data containing 500 English utterances. Every utterance is assigned with its corresponding speech-to-text information provided by Sogou Corporation.

Data Annotation. The corpus is labeled by three well-trained annotators. The annotators are asked to label the emphasis by listening to the utterances and reading corresponding words simultaneously. The words are then classified into

labels of 0 and 1 indicating normal and emphasized words. 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, 1500 Mandarin utterances and 500 English utterances are labeled emphasis. 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: 27.03%, normal: 72.97%. An example of the label sentences is shown in Fig. 3.

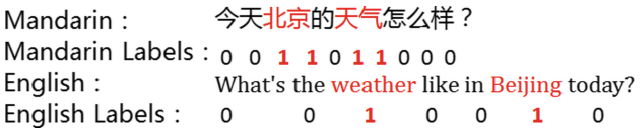

Fig. 3. An example of emphasis labels in Mandarin and English from the VDAs.

4.2 Features

Acoustic Feature. Previous works indicate that emphasis usually has higher F0, longer duration and higher energy [4]. Therefore, we use 19-dimensional acoustic features, including Log F0 (lf0) (1), energy (1), duration (1), Position of Frame (PoF) (4) and Mel Frequency Cepstral Coefficients (MFCCs) (12) prosodic features. The used PoF features include the position of the syllables in the sentence, the position of the frame in syllable and the position of the frame in sentence [21]. The frame length of voice segments is 25 ms and frame shift is 5 ms. Features are normalized to the mean 0 and the variance 1.

Textual Feature. For textual information in Mandarin Corpus, we first use Thulac Tool [17] which is an efficient Chinese word segmentation to get words of an utterance. Then we utilize word2vec to learn word embeddings. Specifically, we use the whole 31.2 million chinese word corpora collected from the 7.5 million utterance from SVAD13 [22] as the training corpora for word2vec. As for the textual information for English databases, we adopt the publicly available 300-dimensional word2vec vectors, which are trained on 100 billion words from Google News to represent word vector.

4.3 Experimental Setup

Comparison Methods. We compared the performance of emphasis detection with some well-known LSTM baseline models for comparison, bi-directional long short-term memory (BLSTM) [18], convolutional bidirectional long short-term memory (CNN-BLSTM) [16], Multi-path convolutional bi-directional long short-term memory neural networks (MCNN-BLSTM) [21]. Our proposed model is

Table 1. The performance on Mandarin corpus and English corpus with different comparison methods.

Method	Mandarin corpus			English corpus		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
BLSTM	0.396	0.453	0.422	0.387	0.639	0.477
CNN-BLSTM	0.493	0.541	0.516	0.524	0.560	0.538
MCNN-BLSTM	**0.542**	0.595	0.567	**0.632**	0.595	0.613
MCNN-ABLSTM	0.523	**0.643**	**0.577**	0.627	**0.616**	**0.621**

Attention-base Multi-path convolutional bi-directional long short-term memory neural networks (MCNN-ABLSTM).

Metrics. In all the experiments, we evaluate the performance in terms of F1-measure, Precision, Recall. The datasets are split by train:val:test = 8:1:1.

4.4 Experimental Results

4.4.1 Performance Comparision

To evaluate the effectiveness of our proposed MCNN-ABLSTM, we compare the performance of emphasis detection with some baseline methods: BLSTM, CNN-BLSTM, MCNN-BLSTM for both Mandarin corpus and English corpus. Table 1 shows the results of emphasis detection with acoustic information and textual information.

For the Mandarin corpus from Sogou Voice Assistant, in terms of F1-measure, the proposed MCNN-ABLSTM outperforms all the baseline methods: +15.5% compared with BLSTM, +6.1% compared with CNN-LSTM, and +1.0% compared with MCNN-BLSTM. Specifically, (1) to demonstrate the Multi-path solution of our proposed method, the MCNN-BLSTM also outperforms the BLSTM (+14.5%) and CNN-BLSTM (+5.1%). This proves the effectiveness of the proposed MCNN component which considers the in-dependency nature of different features, in modeling the multi-modal high-level features. (2) To demonstrate the ABLSTM part of our proposed method, comparing MCNN-BLSTM and MCNN-ABLSTM, although MCNN-BLSTM has a better performance in terms of precision, MCNN-ABLSTM has a more balanced overall performance, +4.8% in terms of recall, +1.0% in terms of F1-measure. Therefore, our proposed MCNN-ABLSTM with acoustic-guide attention on textual feature is a more effective way for emphasis detection in VDAs.

To demonstrate the comparability and the adaptability of our method, we also report experimental results on a real-world English corpus from Sogou Corporation. As shown in Table 1, the F1-measure reaches 0.621, showing +14.4% improvement compared with BLSTM, +8.3% improvement compared with CNN-BLSTM, +0.9% improvement compared with MCNN-BLSTM, indicating that our method still shows advantages on utterances of other language.

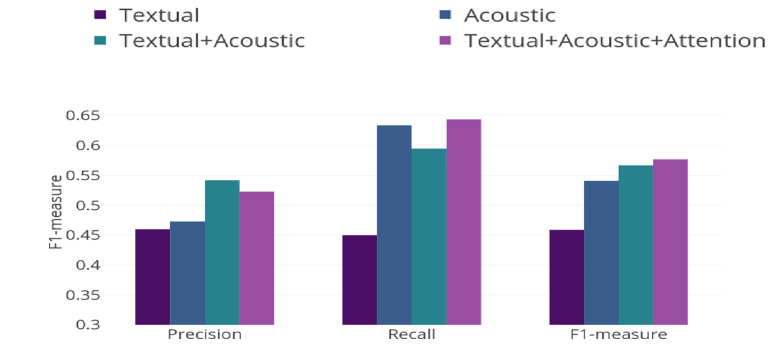

Fig. 4. Feature contribution analysis.

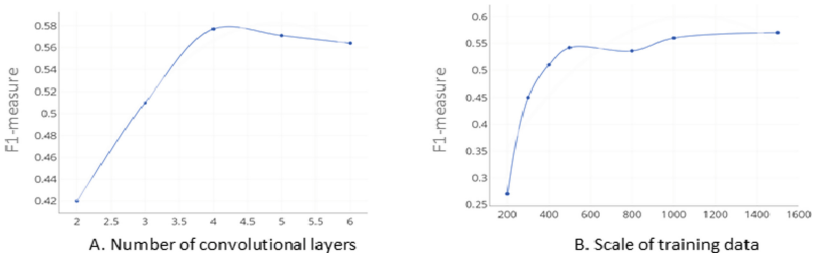

Fig. 5. Parameters analysis.

4.4.2 Feature Contribution Analysis

Then we discuss the contributions of acoustic and textual features. The F1-measure, precision, recall for emphasis detection results for Mandarin Corpus are shown in Fig. 4. Specifically, for 'Textual Only', 'Acoustic Only', 'Textual+Acoustic', we utilize MCNN-BLSTM model, and for 'Textual+Acoustic+Attention', we utilize MCNN-ABLSTM model. As in Fig. 4, the performance of 'Acoustic Only' is better than 'Textual Only', which indicates that the acoustic information can contribute more to the emphasis detection in the real world VDAs. Moreover, 'Textual+Acoustic' which contains both textual information and acoustic information performs better than 'Acoustic Only' +2.6% in terms of F1-measure. The results validate the necessity of taking the textual information into consideration. Moreover, 'Textual+Acoustic+Attention' which consider both acoustic feature and textual feature with our proposed MCNN-ABLSTM has the best performance. Compared with 'T+A', 'T+A+attention' +1.0% in terms of F1-measure and +4.8% in terms of recall. These convince that our proposed attention mechanism can be more effective in modeling multi-modal features.

4.4.3 Parameter Sensitivity Analysis

We show how changes of parameters in MCNN-ABLSTM affect the performance of emphasis detection in Mandarin Corpus.

Multi-path Convolutional Layers Analysis. We first test the parameter sensitivity about Multi-path Convolutional Layers. As shown in Fig. 5(a), the performance reached the highest performance when the layer of Multi-path Convolutional is 4. With the increase of the number of the layers, the performance decreased for over-fitting. So we choose the four convolutional layers as the experimental setup.

Training Data Scalability Analysis. We further test the parameter sensitivity about training data size of Mandarin Corpus. As shown in Fig. 5(b), with the increase of the amount of training data, F1-score performance with rapid ascension, but when the size of training data over 1500, the performance reaches convergence. Considering time efficiency, we choose 1500 as our experiment dataset.

5 Conclusions

In this paper, we propose a novel scheme for emphasis detection with extra attention on textual information. Specifically, we first model the acoustic features and textual features utilizing a MCNN component individually. Then combining high-level multi-modal features, we train an attention-based emphasis classifier ABLSTM, to comprehensively learn discriminative features from diverse users. Experiments based on real-world Mandarin and English corpus show the effectiveness of our methods. Based on our work, VDAs can better understand speakers' attitudes and intentions which contributes to more humanized intelligent service.

Acknowledgements. This work is supported by Tiangong Institute for Intelligent Computing, Tsinghua University and the state key program of the National Natural Science Foundation of China (NSFC) (No. 61831022).

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Brenier, J.M., Cer, D.M., Jurafsky, D.: The detection of emphatic words using acoustic and lexical features. In: Ninth European Conference on Speech Communication and Technology (2005)
3. Cernak, M., Honnet, P.E.: An empirical model of emphatic word detection. In: Interspeech, pp. 573–577 (2015)
4. Chen, J.Y., Lan, W.: Automatic lexical stress detection for Chinese learners' of English. In: International Symposium on Chinese Spoken Language Processing (2011)
5. Chen, Y., Yuan, J., You, Q., Luo, J.: Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. arXiv preprint arXiv:1807.07961 (2018)

6. Do, Q.T., Takamichi, S., Sakti, S., Neubig, G., Toda, T., Nakamura, S.: Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
7. Do, Q.T., Toda, T., Neubig, G., Sakti, S., Nakamura, S.: Preserving word-level emphasis in speech-to-speech translation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 544–556 (2017)
8. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
9. Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., Precoda, K.: Lexical stress classification for language learning using spectral and segmental features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7704–7708 (2014)
10. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. J. Mach. Learn. Res. **3**(Aug), 115–143 (2002)
11. Heldner, M.: Spectral emphasis as an additional source of information in accent detection. In: ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding (2001)
12. Kennedy, L.S., Ellis, D.P.W.: Pitch-based emphasis detection for characterization of meeting recordings. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2003, pp. 243–248 (2003)
13. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Ladd, D.R., Morton, R.: The perception of intonational emphasis: continuous or categorical? J. Phon. **25**(3), 313–342 (1997)
16. Li, L., Wu, Z., Xu, M., Meng, H.M., Cai, L.: Combining CNN and BLSTM to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition. In: Interspeech, pp. 1392–1396 (2016)
17. Li, Z., Sun, M.: Punctuation as implicit annotations for Chinese word segmentation. Comput. Linguist. **35**(4), 505–512 (2009)
18. Ning, Y., et al.: 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, pp. 5615–5619 (2017)
19. Schnall, A., Heckmann, M.: 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)
20. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997)
21. Zhang, L., et al.: Emphasis detection for voice dialogue applications using multi-channel convolutional bidirectional long short-term memory network. In: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 210–214. IEEE (2018)
22. Zhou, S., Jia, J., Wang, Q., Dong, Y., Yin, Y., Lei, K.: Inferring emotion from conversational voice data: a semi-supervised multi-path generative neural network approach. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)