CNN-BLSTM Based Question Detection from Dialogs Considering Phase and Context Information

Yuke Si[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1,2,*], *Mengfei Wu*[1], *Aijun Li*[3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Institude of Linguistics, Chinese Academy of Social Science, Beijing, China

{siyuke, longbiao_wang}@tju.edu.cn, jdang@jaist.ac.jp

Abstract

Question detection from dialogs is important in human-computer interaction systems. Recent studies on question detection mostly use recurrent neural network (RNN) based methods to process low-level descriptors (LLD) of the utterance. However, there are three main problems in these studies. Firstly, traditional LLD features are defined based on human a priori knowledge, some of which are difficult to be extracted accurately. Secondly, previous studies of question detection only consider features from amplitude information and ignored phase information. Thirdly, previous studies show that the context in an utterance is helpful to detect question, while the context between utterances is not well investigated in this task. To cope with the aforementioned problems, we propose a CNN-BLSTM based framework, where amplitude information is obtained from the combination of spectrogram and LLD, and processed together with the phase information. Our framework also models the context information in the dialog. From the experiments on Mandarin dialog corpus, we revealed the effectiveness of the integrated feature with both amplitude and phase in question detection. The results indicated that the phase feature was helpful to detect the questions with a short duration, and the context between utterances was beneficial to detect questions without special interrogative forms.

Index Terms: question detection, amplitude, phase, convolutional neural network, bidirectional recurrent neural network

1. Introduction

Question detection from conversations is very critical in human-computer interaction systems. The detected questions can be applied to automatic transcription of punctuation marks, meeting summarization with question/answer pairs, thus providing useful information for identifying speakers' roles and understanding users' intentions. However, question detection is still a challenging issue. It is insufficient to define a feature based only on human a priori knowledge for question detection. And some of the artificially defined features cannot be extracted accurately. Besides, it is still difficult to detect questions from the spontaneous speech in Mandarin because of its diverse patterns with varied forms of intonation and expression.

Previous works on question detection from the speech signal consider mostly acoustic and prosodic features. Shriberg et al. investigated the function of different prosodic features in dialog act classification, including statement, question, backchannel etc [1]. The prosodic features they used were F0, energy, duration, pause and so on. Their study revealed the great potential of prosody for identifying English dialog acts. The further work was conducted on F0-based prosodic features in different languages such as French and Vietnamese [2][3]. Researchers to follow used extended acoustic and prosodic features that are called low-level descriptors (LLDs). The LLDs commonly include energy, spectral, and/or voice related features, and other statistical features of speech sound. Tang et al. worked on question detection by using 65 LLDs with an RNN-based model [4]. Ortega and Vu classified dialog acts by combining lexical features and 13-dimensional Mel-frequency cepstrum coefficients (MFCC), and their result explored that the acoustic features are helpful to recognize questions [5]. Arsikere et al. proposed a number of new statistical acoustic features for dialog act classification [6]. Their work demonstrated significant differences of histograms between questions and statements. Wei et al. used prosodic and MFCC features to detect interrogative intonation in Mandarin, which was successful under a speaker-independent condition [7].

Although the previous works above have shown the effectiveness of the common features such as F0, pause and so forth, some problems still remain. Firstly, most of the features used are artificially defined depending on human a priori knowledge. However, it is difficult to extract these features accurately. Furthermore, there may be other under-explored features in speech signal that further contribute to detecting question. As a representation of amplitude information in speech signals, spectrograms can be exploited to extract effective features. Direct applications of spectrograms were found successful for speech recognition [8][9], speaker verification [10] and emotion recognition [11][12]. In the present work we follow this path, and investigate the effectiveness of spectrogram for question detection. Secondly, frequency properties of speech signals consist of two parts: the amplitude and phase information. However, prior research efforts in question detection mainly focus on the amplitude-based features without considering phase-based features. In recent, the role of phase information is investigated in many tasks such as speech enhancement [13], speech recognition [14], emotion recognition [15], anti-spoofing [16], and speaker identification [17][18]. In this study we explores whether phase-based features are meaningful to detect questions. Thirdly, the context information in a dialog contains two types of cues, one is the context between words in an utterance, and the other one is between utterances in a dialog. Previous studies showed that the word-level context is effective on question detection [4], whereas the effect of utterance-level context on question detection performance has not been fully explored.

To address the above problems, we investigate the effec-

[*]Corresponding author

tiveness of three-category information (amplitude, phase and context between utterances) in the speech signals for question detection. Specifically, we extract features of amplitude and phase according to the methods described in Section 2. Then we take advantage of a convolutional neural network (CNN) to fuse different features and extract the high-level representation of each utterance. Besides, we adopt a bidirectional long short term memory (BLSTM) mechanism to model the context information connecting adjacent utterances, thus exploring the significance of context information during detecting questions from dialogs.

2. Feature Description and Extraction

As shown in the formula (1), the spectrum of speech signal can be calculated by using Discrete Fourier Transform (DFT).

$$X(\omega, t) = |X(\omega, t)|\, e^{j\theta(\omega,t)}, \tag{1}$$

where $|X(\omega, t)|$ is the amplitude spectrum, and $\theta(\omega, t)$ is the phase at frequency ω, and time t. To obtain the best available information of the speech signal, the pair of features including amplitude and phase is extracted. The following is the description of the features and the methods for extraction.

2.1. Amplitude Information

2.1.1. Spectrogram with Log Magnitude Spectrum

The log magnitude spectrum is the logarithm of the magnitude part of the signal's Fourier transform $|X(\omega)|$:

$$L(t, \omega) = log(|X(t, \omega))|). \tag{2}$$

The spectrum contains considerable information about the input speech signal, such as formant, pitch, harmonic structure and so on. The logarithmic operation on the spectrum can better describe the differences in the frequency domain, which is also consistent with the characteristics of the human auditory system.

2.1.2. Mel Frequency Cepstral Coefficient (MFCC)

As one of the most popular LLDs in speech processing, the MFCC reflects amplitude-based information of the speech signal. It is defined as the formula (2):

$$Ci = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} Ljcos\left(\tfrac{\pi i}{N}(j - 0.5)\right), \tag{3}$$

where N is the number of Mel-frequency bins of log spectrum L, and i is the number of cepstral coefficients. The MFCC undergoes a cepstral analysis on the log magnitude spectrum along the Mel frequency scale. As we know, Mel-Frequency analysis is based on experiments of human speech perception, thus it pertains characteristics similar to the human ear. This point makes the MFCC suitable to play a role as a feature in speech studies.

2.2. Phase Information

2.2.1. Modified Group Delay Cepstral Coefficients (MGDCC)

The original phase information is difficult to be analyzed directly because all the values are wrapped in $(-\pi, \pi)$, which is called the phase wrapping. To overcome this problem, the group delay is introduced. Then, many studies shows that the modified group delay performs better than the original group delay feature [19][20][21][22]. The modified group delay function can be defined as follows:

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau\,(\omega)|)^{\alpha}, \tag{4}$$

$$\tau\,(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S\,(\omega)|^{2\gamma}}, \tag{5}$$

where $X(\omega)$ and $Y(\omega)$ denote the Fourier transform of the signal $x(n)$, and $nx(n)$, respectively. Subscripts R is the real part of the Fourier transform, while I is the imaginary part. $S(\omega)$ denotes the cepstrally smoothed $X(\omega)$, and the ranges of α and γ are $(0 < \alpha \leq 1)$, and $(0 < \gamma \leq 1)$, respectively.

It can be seen that the amplitude information $X(\omega)$ is required when calculating the MGDCC. Therefore, the MGDCC contains both amplitude and phase information to some extent.

2.2.2. Relative Phase

It is a problem that the original phase information changes depending on the clipping position of the input speech even at the same frequency. To solve it, the relative phase is proposed by Wang, et al. [23]. According to the definition of the relative phase, the phase of a certain base frequency ω is kept constant, and the phases of other frequencies are estimated relative to the base. As an example, by setting the base frequency ω to 0, the following formula is obtained:

$$X^{'}\,(\omega) = |X\,(\omega)| \times e^{j\theta(\omega)} \times e^{j(-\theta(\omega))}, \tag{6}$$

and for the other frequency $\omega^{'} = 2\pi f^{'}$, the spectrum becomes

$$X^{'}(\omega^{'}) = |X^{'}(\omega^{'})| \times e^{j\theta(\omega^{'})} \times e^{j\frac{\omega'}{\omega}(-\theta(\omega))}. \tag{7}$$

In this way, the phase information can be normalized as:

$$\tilde{\theta}(\omega') = \theta(\omega') + \frac{\omega'}{\omega}(-\theta(\omega)). \tag{8}$$

It can be seen that relative phase values are calculated from the pure phase part without using amplitude information. This means that the relative phase contains only phase information.

3. Model Description

Previously, the most commonly used models for question detection are the support vector machine (SVM) [24] and decision tree (DT) [2][3][25], but they cannot model the context information. To solve this problem, the RNN-based method is applied on LLD to model the context in an utterance during detecting question[4][26]. However, the fixed LLDs features may be insufficient or inaccurate for robust question detection.

Inspired by these studies, and considering the great potential of the CNN in spectrogram processing [27][28][29][30], we proposed a CNN-BLSTM model for question detection. The whole framework of our approach is illustrated in Figure 1. Specifically, speech signal at an utterance-level will be processed to generate both amplitude-based features and phase-based features in each time frame according to methods described in Section 2. It can be seen from the formulas that the phase features contain minimal or no amplitude information, and thus the two independent features could be integrated. The final combined feature can be represented as:

$$F^t = [F_a^t, F_p^t], \tag{9}$$

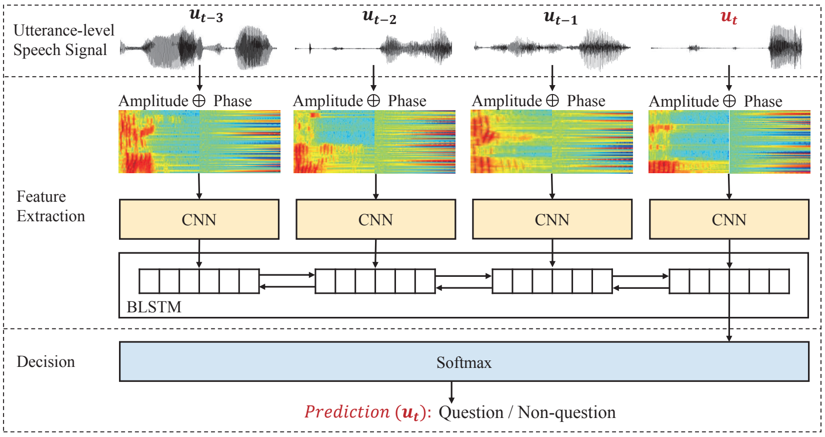

Figure 1: *Framework of CNN-BLSTM method on question detection, ⊕ represents a concatenation.*

where F_a^t and F_p^t are the amplitude and phase of the utterance in time t, respectively.

Here we exploit the CNN to fuse the integrated features of amplitude and phase, and obtain the high-level representation of an utterance. Then a softmax layer is adopted to obtain the prediction of the utterance. Besides, to clarify the significance of the context information in the dialog, we utilize the BLSTM to model the context between utterances. As shown in Figure 1, we use the CNN-output features of a current utterance u_t and previous three adjacent utterances as the input to the BLSTM. This means that totally four utterances will be processed by the BLSTM and the label of u_t will be finally predicted.

4. Experiment

In this section, we conduct a set of experiments on a dialog corpus in Mandarin to evaluate the effectiveness of our proposed method for dialog question detection from speech signal.

4.1. Corpus Description

The CASIA-CASSIL Corpus is a restricted-domain corpus, which is a collection of spontaneous telephone conversations in the real world between two speakers about making restaurant reservation [31]. It is co-established by the Institute of Automation, Chinese Academy of Science, and Institute of Linguistics, Chinese Academy of Social Science. The data we used contains 195 dialogs with 8062 utterances in Mandarin. The durations of all the utterances range from 0.08 to 4.2 seconds approximately, and the number of utterances in each dialog is between 25 and 91. There are totally 2199 (27.3%) question sentences and 5863 (72.2%) non-question sentences. 80% dialogs are selected randomly as the train set, and the other data are used as the test set. Finally, we obtained 156 dialogs of the train set including 1749 question sentences, and 4695 non-question sentences. For the test set, we used 39 dialogs containing 450 question sentences and 1168 non-question sentences.

4.2. Experimental Setup

In this paper, we use only speech data labeled with question or non-question. Each sentence is given in the Microsoft wav format and sampled at 16 kHz. All the features and their sizes are listed in Table 1. The speech signal is divided into frames of a 16-ms window size with a shift window of 8 ms. Since some of the spontaneous utterances are very short, the statistical LLD features cannot be extracted successfully. Considering the aforementioned work, the 32 frame-level LLD features are chosen as the baseline, including 1 energy feature, 12 MFCCs, 1 F0, 2 voicing related features, and 16 first order derivatives of them. The LLD features of each time frame were extracted by using the OpenSMILE toolkit [32] with the default methods. After a padding processing, the final size of LLDs is 530 × 32, which means that each utterance contains 530 time frames and each frame has 32 attributes. Since the spectrograms of the data in this corpus have almost no valid information in the high-frequency domain, 64-dimensional spectrogram (from 0-4000 Hz) is extracted in each time frame. When extracting MGDCC, α is set to 0.1, and γ is set to 0.2. Totally we obtain 36-dimensional MGDCC features in every time frame, including 12 static MGDCCs, 12 ΔMGDCCs, and 12 $\Delta\Delta$MGDCCs. When the relative phase is calculated, the base frequency ω is set to 1000 Hz, and 129-dimensional relative phase features are extracted in each time frame. In addition to clarifying the role of the independent features, we checked the effectiveness of several combinations of them, which can reflect the relationship among features combined for detecting question.

To select the optimal structure and parameters, we tested different numbers of hidden units, layers, etc. For both CNN and BLSTM, we utilized focal loss as the loss function. The structure of the CNN applied to the features from 1 to 7 contains two convolutional layers and two max-pooling layers, while for features 8 and 9, the CNN consisted of three convolutional layers and three max-pooling layers. In order to observe the best performance of each feature, we tried different parameters of the filter size and pooling size. After a flatten layer, we adopt a full connected layer with 128 units. A drop out of 0.5 was applied before output in order to avoid the over-fitting problem. The BLSTM had two hidden layers, and each had 256 units.

4.3. Experiment Results and Discussions

Table 1 gives the experimental results. The overall accuracy (Acc), precision, recall, and F1-measure of the question set were used to compare the performances. It can be seen that the CNN-BLSTM model with the integrated feature of amplitude and phase achieves the best performance for detecting question, which proved the effectiveness of our method. In each experiment, the F1 score of the question set are lower than the overall accuracy. The reason might be that the utterances labeled with

Table 1: *Experiment features and results.*

ID	Feature	Feature Size	Model	Acc (%)	Question Set (%)		
					Precision	**Recall**	**F1**
1	LLD	530 × 32	CNN	76.02	58.12	47.33	52.17
2	Spectrogram (Spec)	530 × 64	CNN	**77.44**	**61.97**	48.89	54.66
3	MGDCC	530 × 36	CNN	76.08	58.92	46.22	51.80
4	Relative Phase (RP)	530 × 129	CNN	63.91	38.98	52.67	44.80
5	Spec + LLD (Amplitude)	530 × 96	CNN	77.07	60.48	50.67	55.14
6	Amplitude + MGDCC	530 × 100	CNN	77.19	60.05	53.78	56.74
7	Amplitude + RP	530 × 193	CNN	**77.44**	59.77	57.78	58.76
8	Amplitude + MGDCC + RP	530 × 229	CNN	77.13	58.06	64.00	60.89
9	Contextual Amplitude + MGDCC + RP	4 × 530 × 229	CNN-BLSTM	77.13	57.84	**65.56**	**61.46**

question hold a proportion lower than 30%. From the results, we can draw conclusions carefully: 1) The spectrogram outperforms the baseline LLD by 5.21% relative error reduction of F1 in the question set. It indicates that the spectrogram can be used as a good feature to detect questions from the speech signal. 2) The combination of the spectrogram and the LLDs performs better than each of them alone, which shows the complementary relationship of the spectrogram and human a prior knowledge based LLDs. 3) The combined feature with both amplitude and phase information outperforms the methods that use spectrogram or phase alone. Comparing the experiments 5 with 8, the relative error reduces by 12.82% in terms of F1 in the question set. The results suggest that there is a complementarity between the amplitude and phase. This conclusion is also proved by the better performance of the MGDCC than the relative phase, because the MGDCC contains amplitude information while the relative phase does not. 4) The CNN-BLSTM based method can model the context information between utterances. Compared with the CNN-based method without contextual information, the result of the contextual method surpassed it by 1.46% relative error reduction in F1 score. It indicates that the context between utterances is helpful for detecting questions.

To clarify the effect of the phase information, we perform a statistical analysis on the predictions of the CNN-based model with only amplitude and the CNN-based model with the integrated feature including the amplitude and phase. The duration of utterances in the test set has the mean of 1.31 s and a variance of 0.83 s. After careful observation and comparison, we divided the test set into two subsets according to the duration of the utterances. One of the subsets consists of 738 (45.61%) utterances with the duration less than 1 second, while the other subset contains the remaining utterances. As shown in Table 2, when the phase information was used, both the absolute error and the relative error of the F1 score in the question set reduced in the whole test set and every sub-test-set. Furthermore, the absolute and the relative error reduced further in the subset containing utterances less than 1 second. The results indicated that the phase information was highly contributing to detect the short questions. The traditional account is that the questions are often associated with a terminal F0 rise, and the F0 variations are accompanied by certain changes of voicing activities. However, when the utterance is very short, the dynamic information of F0 variations is not enough for the model to recognize whether it has a rising trend. In such an ambiguous case, according to our results, the phase information played a complementary role to detect the question. Therefore, it is reasonable to conjecture that the phase information may reflect such alterations of the source sounds signaling the tendency of the voicing activities.

Table 2: *Comparison between using amplitude features (Amp) only and the combined feature of amplitude and phase in different test sets. U_d represents the duration of the utterances. Acc denotes the overall accuracy (%). AER, RER denotes the absolute error reduction, and the relative error reduction of F1 score in Question set when considering phase features, respectively.*

Test Set	Feature	Acc	Question Set (%)		
			F1	**AER**	**RER**
All	Amp Amp+Phase	77.07 77.13	55.14 60.89	5.75	12.82
U_d ≥ 1 s	Amp Amp+Phase	69.77 69.32	57.66 61.43	3.77	8.90
U_d < 1 s	Amp Amp+Phase	85.77 86.45	47.76 59.35	**11.59**	**22.19**

In addition to the phase information, we compared the results when the context information between utterances was added. It was found that questions without special interrogative forms (such as a statement form) could be detected better when the context information between utterances is employed. In this kind of data, the CNN-BLSTM method outperforms CNN method by a 14.3% relative error reduction in terms of accuracy.

5. Conclusion

In this study, we proposed a CNN-BLSTM based framework with the combined feature of amplitude, phase, and context between utterances for detecting question from speech signals in Mandarin dialogs. To the best of our knowledge, our work is the first one to combine the spectrogram and LLDs, and integrate amplitude with phase information in a question detection task. In addition, it is the first work for question detection that combines the CNN and BLSTM to model the context information between utterances. The experimental results demonstrated the effectiveness of our method, and the functions of the phase and context were discussed in detail. In the future, we will concentrate on further investigation on phase and amplitude features as well as improving the imbalanced classification strategy.

6. Acknowledgements

This work is supported by the National Key R&D Program of China (2018YFB1305200, 2017YFE0111900), National Natural Science Foundation of China (61771333), Tianjin Municipal Science and Technology Project (18ZXZNGX00330), National Basic Research Program of China (2013CB329301), and Research Foundation of Beijing Municipal Science & Technology Commission (Z181100008918002).

7. References

[1] E. Shriberg, R. A. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[2] V. M. Quang, E. Castelli, and P. N. Yłn, "A decision tree-based method for speech processing: Question sentence detection," *Lecture Notes in Computer Science*, vol. 4223, pp. 1205–1212, 2006.

[3] M. Vu, L. Besacier, and E. Castelli, "Automatic question detection: prosodic-lexical features and crosslingual experiments," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2007, pp. 221–224.

[4] Y. Tang, Y. Huang, Z. Wu, H. Meng, M. Xu, and L. Cai, "Question detection from acoustic features using recurrent neural network with gated recurrent unit," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6125–6129.

[5] D. Ortega and N. T. Vu, "Lexico-acoustic neural-based models for dialog act classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6194–6198.

[6] H. Arsikere, A. Sen, A. P. Prathosh, and V. Tyagi, "Novel acoustic features for automatic dialog-act tagging," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6105–6109.

[7] B. Wei, Y. Li, M. Gu, J. Tao, L. Chao, and S. Liu, "Combining prosodic and spectral features for mandarin intonation recognition," in *International Symposium on Chinese Spoken Language Processing*, 2014, pp. 497–500.

[8] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, and A. Coates, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[9] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: end-to-end speech recognition in english and mandarin," *International Conference on Machine Learning*, pp. 312–321, 2016.

[10] E. Variani, L. Xin, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4052–4056.

[11] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 1089–1093.

[12] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 2666–2670.

[13] T. Gerkmann, M. Krawczykbecker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[14] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[15] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, and X. Li, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1611–1615.

[16] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017.

[17] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "Dnn-based amplitude and phase feature enhancement for noise robust speaker identification," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2016, pp. 2204–2208.

[18] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[19] R. M. Hegde, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 517–520.

[20] R. Padmanabhan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2009, pp. 2355–2358.

[21] J. M. K. Kua, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Ls regularization of group delay features for speaker recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*.

[22] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[23] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2015, pp. 2092–2096.

[24] A. Li, M. Xu, and L. Cai, "Acoustic features prominence based chinese question detection," *Chinese Science paper*, no. 7, pp. 826–829, 2014.

[25] J. Yuan and D. Jurafsky, "Detection of questions in chinese conversational speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 47–52.

[26] A. Ando, R. Asakawa, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Automatic question detection from acoustic and phonetic features using feature-wise pre-training," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1731–1735.

[27] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5115–5119.

[28] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 1993–1997.

[29] G. Yue, X. Li, S. Chen, J. Zhang, and I. Marsic, "Speech intention classification with multimodal deep learning," in *Canadian Conference on Artificial Intelligence*, pp. 260–271.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1086–1090.

[31] Y. Zhou, Q. Hu, L. Jie, and J. Yuan, "Combining heterogeneous deep neural networks with conditional random fields for chinese dialogue act recognition," *Neurocomputing*, vol. 168, pp. 408–417, 2015.

[32] F. Eyben, F. Weninger, F. Gross, and B. W. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 2013 ACM Multimedia Conference*.

[This paper was published at Interspeech 2019]