Interspeech 2018
2-6 September 2018, Hyderabad

# Discourse Marker Detection for Hesitation Events on Mandarin Conversation

*Yu-Wun Wang[1], Hen-Hsen Huang[1], Kuan-Yu Chen[2], Hsin-Hsi Chen[1,3]*

[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
b02902033@ntu.edu.tw, hhhuang@nlg.csie.ntu.edu.tw, kychen@mail.ntust.edu.tw, hhchen@ntu.edu.tw

## Abstract

The occurrence of hesitation events in spontaneous conversations can be associated with the difficulties in memory recall. One indicator of hesitation in speech in Taiwanese Mandarin is the usage of discourse markers. This paper introduces an approach to the detection of discourse markers that denote hesitation events. We propose a sequential labeling model to detect discourse markers in conversations by taking information on both acoustic level and word level into account. Experimental results show the integration of word-level acoustic feature extraction network significantly enhances the detection performance. Our approach for further applications is also discussed.

**Index Terms:** discourse markers detection, hesitation event, Mandarin spontaneous conversation

## 1. Introduction

Failing to retrieve an event in human memory frequently occurs in our daily life. Hesitation is a common phenomenon when people have difficulties in memory recall while speaking. Hesitation events reflect people's feeling or attitude, which reveals useful information for subsequent applications. For instance, dialogue systems can provide more explanation to users sounding more uncertain, and a reminding system can detect users' difficulties in memory recall and trigger the assistance. To analyze the "failing to recall" events, we propose an approach to detect discourse markers for hesitation events in speech. Recent study shows that the mild cognitive disorder, a syndrome rated to memory, can be detected from speech [1], which reveals the link between speech and memory recall processes.

Hesitation can affect speech in several ways such as repetitions, false starts, filled pauses, and unfilled pauses [2]. Different from speech with prepared text, people think and talk at the same time in spontaneous speech. Therefore, the hesitation occurs in speech with varying speaking rate, filled pause, and unfilled pause [3]. The association between hesitation and the usages of discourse markers (DMs) are observed in terms of doubt expression and uncertainty in Taiwanese Mandarin [4] [5]. DMs allow speakers to have more time to think or react [6]. In Mandarin speech, discourse markers, which are words with their semantic meaning lost, are used for pragmatic purposes in conversation [7]. In different context, the literally same (LS) words can have semantic meanings and are used as content words. For example, in Table 1, the word 那個, tagged with [NE_GE] in the first utterance, is a DM, which has little semantic meaning and serves as a prolongation for the speaker to recall the following term. By contrast, the LS word in the second utterance serves as a determiner. In this paper, we aim to detect the hesitation in daily conversation by differentiating the discourse markers from LS content words.

Table 1: *Examples of discourse markers (DMs) and literally same (LS) content words.*

| Type | Utterance |
|---|---|
| DM | 搭 捷運 到 **那個 [NE_GE]** 捷運 忠孝復興 站<br>(Take the metro to **[NE_GE]** Zhongxiao Fuxing Metro Station.) |
| LS | 喔 為什麼 **那個** 不是 很 近 嗎<br>(Oh why? Isn't **that** close?) |

Prosodic information is widely used in detecting disfluency in spontaneous speech and improving automatic speech recognition (ASR) [8]. Previous work explores the prosody variation of disfluency events and discourse markers in spontaneous speech [9], and suggests that prosodic features could be useful information for discourse marker detection. In this paper, we learn hesitation-relevant acoustic features from various types of acoustic features related to speech recognition, prosody, and emotion recognition. Furthermore, we propose a weight filtering method to learn word-level hesitation-relevant acoustic features from frame-wise features, which effectively enhance the performance of discourse marker detection. Our model is evaluated on Sinica Mandarin Conversational Dialogue Corpus (MCDC8). In the setting where the transcript is recognized manually, our model achieves a precision of 82.87%. In the fully automatic setting, where the transcript is performed by an ASR service, our model also achieves a high precision of 79.00%. That confirms the feasibility of our approach on practical applications.

## 2. Methods

Our model addresses the task of discourse marker detection from lexical aspect and acoustic aspect. Word-based features and acoustic features are used in two kinds of networks, respectively. We combine features learned from two networks to enhance the robustness of the final model. Figure 1 shows the architecture of our discourse marker detection model.


10.21437/Interspeech.2018-2129

Three major components include hesitation-relevant feature extraction network, word sequence labeling network, and the feature combination network for the two aforementioned networks. The first two components and the overall system will be evaluated individually in Section 4.

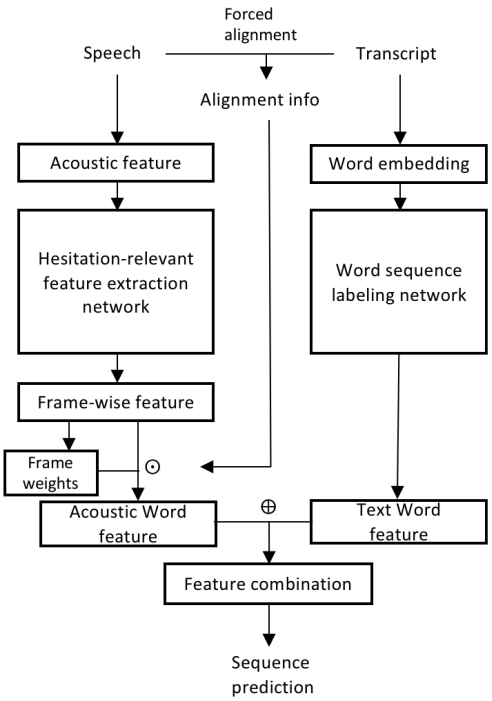


Figure 1: *Architecture of discourse marker detection model.*

### 2.1. Word Sequence Labeling Network

Word sequence labeling network is shown in the right part in Figure 1. We regard discourse marker detection as a sequence labeling task that labels the words served as DM in a transcript. The recurrent neural network (RNN), bidirectional Gated Recurrent Unit (Bi-GRU), and bidirectional Long Short-Term Memory (Bi-LSTM) models are adopted for the task of sequential labeling.

GRU and LSTM show their capacity in modeling long-term dependencies among time steps that is suitable for sequential data like sentences and speech. In contrast to LSTM, GRU is equipped with fewer gates so that it is usually more efficient in training. Bi-GRU and Bi-LSTM learn input features in both directions. At each time step, the prediction is based on the information from both preceding and successive states. This is especially useful for our task since we usually need to read the upcoming words to determine if a word is a discourse marker. Bi-LSTM is also used in disfluency detection [10] and grammatical error detection [11]. Its ability to capture special patterns in text is suitable for this task.

We train Bi-GRU and Bi-LSTM with randomly initialized word embedding as the input layer. The output vectors from two directions are concatenated for prediction or for feature combination.

### 2.2. Hesitation-Relevant Acoustic Feature Extraction

The hesitation-relevant acoustic feature extraction network is shown in the left part in Figure 1. Both convolutional neural network (CNN) and multi-layer perceptron (MLP) models are tested as the internal neural network. We propose a weight filtering method to combine frame-wise features extracted from CNN or MLP into word-level hesitation-relevant acoustic features.

#### *2.2.1. ComParE Feature Set*

We use Computational Paralinguistics Evaluation (ComParE) feature set as the input features for our hesitation-relevant acoustic feature extraction network. ComParE is the baseline acoustic feature set designed for 2013 INTERSPEECH Computational Paralinguistics Challenge [12]. This feature set is also used in filler events detection [13]. The ComParE feature set is extracted by openSMILE tool and includes 141 features such as Mel-Spectrum, Mel-Frequency Cepstral Coefficients (MFCC), energy, zero-crossing rate, and pitch (F0). These features are frame-wise, and we set frame length as 25 microseconds and frame rate as 10 microseconds.

#### *2.2.2. Convolutional Neural Network*

CNN is successfully used in some tasks in speech such as phoneme recognition [14] and emotion recognition [15]. Its local connectivity and weight sharing property allow CNN to exploit the spatial information of signal data. In our model, CNN is employed on ComParE features with one stride. We flatten the output tensors on frame dimension and then pass all vectors to a fully connected layer to obtain the frame-wise feature vectors.

#### *2.2.3. Multi-Layer Perceptron*

MLP is widely adopted in phoneme posterior estimation for ASR. Prior research has also shown the success of MLP on detecting laughter and filler events [13]. In our model, MLP is employed on each frame. Frame-level features of a frame and its neighboring frames in the sliding window are fed into an MLP with 2 hidden layers.

#### *2.2.4. Frame-Level Features to Word-Level Features*

This method learns hesitation-relevant acoustic features for each word from frame-level feature vectors. After frame-wise features are extracted by CNN or MLP, we merge frame-wise feature vectors into word-level feature vectors by applying a weight filter. The weights for each word are learned from all the frame-level feature vectors in the time span of the word. The weights are computed by the procedure as follows.

For each word $w_i$, the system refers to the forced alignment result and gets the time span of $w_i$. We define the maximum frame sequence length as $T$ and write the time span of $w_i$ as $s_i$. The symbol $s_i$ denotes a vector of length $T$. $s_{i,t} = 1$ when time $t$ is within the time span of $w_i$, and otherwise $s_{i,t} = 0$. Next, the model computes the weight $u_{i,t}$ for each frame-wise feature vector $f_t$ with the formula as follows.

$$u_{i,t} = v^T \tanh(W f_t + b) \quad (1)$$

, where $W, b, v$ are the parameters to learn by the model. Then, we apply the time span mask $s_i$ to zero out frame vectors lying outside the time span, i.e., $s_{i,t} = 0$. A softmax function is applied to let all weights sum to 1. The frame weights are computed with the formula as follows.

$$a_{i,t} = \frac{\exp(s_{i,t}\, u_{i,t})}{\sum_{k=1}^{T} exp(s_{i,k}\, u_{i,k})} \quad (2)$$

Finally, the word-level features vector $c_i$ of word $w_i$ is the weighted sum of frame-level feature vectors.

$$c_i = \sum_{t=1}^{T} a_{i,t}\, f_t \quad (3)$$

The word-level acoustic feature vectors are used for prediction or for feature combination.

### 2.3. Network Combination

Word features and word-level acoustic features are concatenated and passed to a fully connected layer with a rectifier to learn feature combination. Finally, the model performs binary prediction on each word with the combined feature. Word sequence labeling network converges much faster than hesitation-relevant feature extraction network does. Thus, we pre-train these two networks separately and train the whole model after then. The loss function is cross entropy. Parameters are trained with 10-fold cross-validation.

## 3. Experimental Setup

### 3.1. Mandarin Conversational Dialogue Corpus

The Sinica Mandarin Conversational Dialogue Corpus (MCDC8) is adopted as the dataset [5]. This corpus consists of eight sets of conversation. Two speakers participated in each conversation. We choose MCDC8 for our task since the conversational data collected in this dataset match our daily life scenario. This corpus is aimed at collecting conversational data that approximates to daily conversation. Speakers can choose and change arbitrary topics during their conversation.

The conversations in MCDC8 are manually transcribed by professional annotators. Special events and word types such as laughter, long pauses, fillers, particles, and discourse markers are also labeled. Table 2 lists the DMs labeled in MCDC8. The LS content words that each DM corresponds to are shown in the right side. The audio sequences are labeled with Inter-Pausing Unit (IPU), and each IPU is segmented into intervals. We use interval as the unit of utterance. There are 15,901 utterances and 93,317 words in total, where 2,000 utterances are used for test and 10% of the rest data are used for validation.

Table 2: *DMs and the corresponding LS content words in MCDC8.*

| DM | # | LS | # |
|---|---|---|---|
| NA | 911 | | |
| NE | 148 | 那 | 1094 |
| NEI | 79 | | |
| NA_GE | 139 | | |
| NE_GE | 237 | 那個 | 316 |
| NEI_GE | 29 | | |
| SHE_ME | 2 | 什麼 | 350 |
| SHEN_ME | 116 | | |
| ZHE_GE | 69 | 這個 | 219 |
| ZHEI_GE | 1 | | |

### 3.2. Evaluation

We first show the performance of our model without the error produced by ASR, i.e., the golden transcripts are given.

Our model is evaluated with precision (P), recall (R), and F-score ($F_1$). For a memory recall assisting system, precision is more important than recall since users' conversation would not be interrupted by a lot of false alarm notifications.

### 3.3. Evaluation on ASR Results

To test if our system can be completely automatic without the efforts of human transcription, we further evaluate our models on ASR results. The ASR service, iFLYTECH,[1] is adopted for machine transcribing. The word error rate (WER) of iFLYTECH is 35.15%.

Our model relies on both acoustic and transcript information. In this case, the transcript, however, is based on the output of an ASR system where wrong recognition may occur. In general, there are three types of wrong ASR outcomes: substitution, deletion, and insertion. The ASR system does not always recognize the target words correctly. In this paper, we propose two strategies to deal with this issue.

#### *3.3.1. Strict Strategy*

In the strict strategy, substitution and deletion errors occurring on discourse markers are directly counted as prediction error. In other words, wrongly recognized words are regarded as non-discourse-marker. Only the correctly recognized target words are further checked if the prediction for them is correct.

#### *3.3.2. Lenient Strategy*

It is common that the ASR system confuses the target word with other words with similar sounds. Even if target words are wrongly recognized, it is still possible for the system to predict from their acoustic information or the surrounding words. In the lenient strategy, all predictions are checked if they are correct.

## 4. Results and Discussions

Table 3 shows the performance of our models for discourse marker detection on golden transcripts (GT). The baseline model, which is a rule-based one, selects all the words in GT with the lexical forms of DMs. That is, DMs and LS content words are all predicted as DMs. The precision of CNN and MLP is slightly higher than that of the baseline model. Bi-GRU word sequence labeling network outperforms the baseline model and Bi-LSTM.

CNN+Bi-GRU and MLP+Bi-GRU significantly enhance Bi-GRU in terms of precision and F-score. That demonstrates feature vectors learned from our hesitation-relevant acoustic feature extraction network effectively improves the predictions. The recall of the Bi-GRU model is higher than that of the MLP+Bi-GRU model. The reason is that Bi-GRU tends to predict the word as DM when seeing LS because Bi-GRU only depends on transcript information.

Table 3: *Results of discourse marker detection with GT.*

| Feature | Model | P | R | $F_1$ |
|---|---|---|---|---|
| **Text** | Baseline | 0.5021 | 1.0000 | 0.6685 |
| **Acoustic** | CNN | 0.5824 | 0.4849 | 0.5292 |
| | MLP | 0.5598 | 0.5622 | 0.5610 |
| **Text** | Bi-LSTM | 0.7133 | 0.9613 | 0.8190 |
| | Bi-GRU | 0.7174 | **0.9699** | 0.8248 |
| **Text and Acoustic** | CNN+Bi-GRU | 0.7620 | 0.9484 | 0.8451 |
| | MLP+Bi-GRU | **0.8287** | 0.9141 | **0.8693** |

[1] http://www.xfyun.cn

In Table 4, we only evaluate on DMs and LS content words to examine if our models can differentiate them. The precision achieved by MLP is better than that achieved by the two models without acoustic information. In addition, Table 5 shows that MLP is good at predicting LS content words, better than Bi-GRU and MLP+Bi-GRU. This result implies MLP is good at differentiating DMs from LS content words based on acoustic information. In other words, the strength of MLP complements Bi-GRU so that the combination of MLP and Bi-GRU enhances the overall performance.

Table 4: *Results on DMs and LS content words.*

| Feature | Model | P | R | $F_1$ |
|---|---|---|---|---|
| **Text** | Baseline | 0.5021 | 1.0000 | 0.6685 |
| **Acoustic** | CNN | 0.7533 | 0.4849 | 0.5900 |
| | MLP | 0.7844 | 0.5622 | 0.6550 |
| **Text** | Bi-LSTM | 0.7179 | 0.9613 | 0.8220 |
| | Bi-GRU | 0.7313 | **0.9699** | 0.8339 |
| **Text and Acoustic** | CNN+Bi-GRU | 0.7727 | 0.9484 | 0.8516 |
| | MLP+Bi-GRU | **0.8320** | 0.9141 | **0.8711** |

Table 5: *Recall on DMs and LS content words.*

| | DM | LS |
|---|---|---|
| MLP | 0.5622 | **0.8441** |
| Bi-GRU | **0.9699** | 0.6406 |
| MLP+Bi-GRU | 0.9141 | 0.1861 |

We perform McNemar's statistical significance test on two combination networks and Bi-GRU respectively. No matter which MLP or CNN is used, the addition of acoustic features extraction network significantly enhances the performance ($p < 0.001$ and $p < 0.05$). In Table 3 and Table 4, the recall achieved by CNN+Bi-GRU is higher than that by MLP+Bi-GRU. Its characteristic is probably different from MLP+Bi-GRU, so we also perform significance test on MLP+Bi-GRU and CNN+Bi-GRU. The p-value shows that these two models are significantly different ($p < 0.02$). If the system is used for recall assistance, MLP+Bi-GRU is the better choice due to the lower rate of false alarms. If we use the system for study purposes, CNN+Bi-GRU are recommended for capturing as much events as possible.

Table 6: *Recall of MLP+Bi-GRU on different DMs.*

| DM | Recall |
|---|---|
| NA | 0.9821 |
| NE | 1.0000 |
| NEI | 1.0000 |
| NA_GE | 0.9047 |
| NE_GE | **0.7894** |
| NEI_GE | 0.8333 |
| SHEN_ME | 0.8333 |
| ZHE_GE | **0.5454** |

Table 6 shows the result for each DM detected by the MLP+Bi-GRU model. The discourse markers NE_GE and ZHE_GE are most difficult to detect. We further evaluate the detection results under the fully automatic condition. Table 7 shows the results of our model with the strict strategy. The performance drops drastically on all models, especially for recall. The recall achieved by the baseline model shows that only 40.77% of DMs are recognized by ASR. The models with only transcript information are highly affected by the errors of ASR. On the other hand, MLP+Bi-GRU can still achieve a precision of 79.00%, very close to the precision 82.87% achieved by the same model given golden transcripts in spite of the 35.15% WER of the ASR service.

Table 7: *Detection performance using ASR results with the strict strategy.*

| Feature | Model | P | R | $F_1$ |
|---|---|---|---|---|
| **Text** | Baseline | 0.4545 | 0.4077 | 0.3613 |
| **Acoustic** | CNN | 0.6590 | 0.2489 | 0.3613 |
| | MLP | 0.5688 | 0.2660 | 0.3625 |
| **Text** | Bi-LSTM | 0.6854 | 0.3648 | 0.4761 |
| | Bi-GRU | 0.6800 | 0.3648 | 0.4748 |
| **Text and Acoustic** | CNN+Bi-GRU | 0.7166 | **0.3690** | **0.4872** |
| | MLP+Bi-GRU | **0.7900** | 0.3390 | 0.4744 |

Table 8 shows the results under fully automatic condition using the lenient strategy. The recall is expected to increase. However, the noise made by ASR has a strong impact on precision. For the text prediction model, the increase on recall is limited. Nonetheless, the recalls of both models with only acoustic features significantly increase. That verifies the capacity of our feature extraction network for learning hesitation-relevant features. The McNemar's test on the ASR results with the lenient strategy also confirms the capacity of acoustic feature extraction network to improve the performance ($p < 0.02$).

Table 8: *Detection performance using ASR results with the lenient strategy.*

| Feature | Model | P | R | $F_1$ |
|---|---|---|---|---|
| **Text** | Baseline | 0.4545 | 0.4077 | 0.3613 |
| **Acoustic** | CNN | 0.5606 | 0.3175 | 0.4054 |
| | MLP | 0.4857 | 0.3648 | 0.4166 |
| **Text** | Bi-LSTM | 0.5056 | 0.3862 | 0.4379 |
| | Bi-GRU | 0.5084 | 0.3862 | 0.4390 |
| **Text and Acoustic** | CNN+Bi-GRU | 0.5443 | 0.3948 | 0.4577 |
| | MLP+Bi-GRU | 0.5931 | 0.3690 | 0.4550 |

## 5. Conclusions

We propose an approach to discourse marker detection in Mandarin conversation for hesitation events using both word features and acoustic features. The experiments show that our method successfully extracts hesitation-relevant information from audio sequence. We explore the complementary characteristics of the acoustic features and the transcript features, and confirm the combination of acoustic and transcript information significantly enhances the overall performance. Our model achieves a precision of 0.79 under the fully automatic condition. That shows its feasibility for subsequent applications such as memory recall assistance.

## 6. Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-107-2634-F-002-019, MOST-107-2634-F-002-011 and MOST-106-2923-E-002-012-MY3.

# 7. References


[1] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pakaski, and J. Kálmán, "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech using ASR," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015, pp. 2694–2698.

[2] H. Maclay and C. E. Osgood, "Hesitation Phenomena in Spontaneous English Speech," *Word*, vol. 15, no. 1, pp. 19–44, 1959.

[3] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, 1992, pp. 521–524 vol.1.

[4] T. L. Lee, Y. F. He, Y. J. Huang, S. C. Tseng, and R. Eklund, "Prolongation in Spontaneous Mandarin," in *INTERSPEECH 2004 – 8th Annual Conference of the International Speech Communication Association, October 4–8, Jeju, Korea, Proceedings*, 2004, pp. 2181–2184.

[5] Y. F. Liu and S. C. Tseng, "Linguistic Patterns Detected Through a Prosodic Segmentation in Spontaneous Taiwan Mandarin Speech," *Linguistic Patterns in Spontaneous Speech*, pp. 147–166, 2009.

[6] Y. Wang, "A Discourse-Pragmatic Functional Study of the Discourse Markers Japanese Ano and Chinese Nage," *Intercultural Communication Studies*, vol. 20, pp. 42–61, 2011.

[7] S. C. Tseng, "Lexical Coverage in Taiwan Mandarin Conversation," *Computational Linguistics and Chinese Language Processing*, vol. 18, no. 1, pp. 1–18, 2013.

[8] C. K. Lin and L. S. Lee, "Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1263–1278, Sept. 2009.

[9] C. H. Lin, C. L. You, C. Y. Chiang, Y. R. Wang and S. H. Chen, "Rich prosodic information exploration on spontaneous Mandarin speech," *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Tianjin, 2016, pp. 1–5.

[10] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency Detection Using a Bidirectional LSTM," *arXiv preprint arXiv:1604.03209*, 2016.

[11] M. Rei, and H. Yannakoudakis, "Compositional Sequence Labeling Models for Error Detection in Learner Writing," in *ACL 2016 – The 54th Annual Meeting of the Association for Computational Linguistics, August 7-12, Berlin, Germany, Proceedings*, 2016, pp. 1181–1191.

[12] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, Lyon, France, Proceedings,* 2013, pp. 148-15.

[13] G. Gosztolya, "Optimized Time Series Filters for Detecting Laughter and Filler Events," in *INTERSPEECH 2017 – 17th Annual Conference of the International Speech Communication Association, August 20–24, 2017, Stockholm, Sweden, Proceedings*, 2017, pp. 2376–2380.

[14] D. Palaz, R. Collobert, and M. M. Doss, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal Using Convolutional Neural Networks," *arXiv preprint arXiv:1304.1018*, 2013.

[15] Q. Mao, M. Dong, Z. Huang and Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networkss," in *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, Dec. 2014.