Speech Intention Classification with Multimodal Deep Learning

Yue Gu(✉), Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic

Department of Electrical and Computer Engineering,
Rutgers University, New Brunswick, NJ, USA
{yue.guapp, Xinyu.lilll8, scl624,
jz549, marsic}@rutgers.edu

Abstract. We present a novel multimodal deep learning structure that automatically extracts features from textual-acoustic data for sentence-level speech classification. Textual and acoustic features were first extracted using two independent convolutional neural network structures, then combined into a joint representation, and finally fed into a decision softmax layer. We tested the proposed model in an actual medical setting, using speech recording and its transcribed log. Our model achieved 83.10% average accuracy in detecting 6 different intentions. We also found that our model using automatically extracted features for intention classification outperformed existing models that use manufactured features.

Keywords: Multimodal intention classification · Textual-acoustic feature representation · Convolutional neural network · Trauma resuscitation

1 Introduction

Human-computer interaction (HCI) is becoming more prevalent in daily living, appearing in applications ranging from navigation systems to intelligent voice assistants. There has been significant research focused on understanding speech, an essential vehicle for human communication. However, machines have faced difficulty extracting the intention of human speech, in part because words may carry different meanings in different contexts. For example, "it is snowing" could either be a comment, reply, or question depending on the inflection and punctuation. To understand the actual meaning and detect the speaker's intention, machines must be able to make these distinctions.

The definition of *speech intention* in this paper is different from the general action-based intention recognition [1]. The goal in this paper is to identify the actual purpose of a speaker's verbal communication in the trauma scenario (e.g. "Directive" for commands/instructions to the medical team or patient). Trauma-related language contains rich information regarding medical operations, cooperation, and team performance that can be used for the detection of medical processes and operation workflows. To identify the actual status of a surgical activity (e.g. the activity is started or finished), it is essential to precisely estimate the intention from human speech. Hence, this intention classifier may be used in a larger system for recognizing clinical activities or verbal procedures.

M. Mouhoub and P. Langlais (Eds.): Canadian AI 2017, LNAI 10233, pp. 260–271, 2017.
DOI: 10.1007/978-3-319-57351-9_30

One approach to estimating speech intention, used by many current systems such as Microsoft LUIS, is to analyze only the sentence syntax. However, this approach ignores valuable acoustic information such as pitch contour, stress pattern, and rhythm. In many applications, such as the emergency medical field, speech tends to be short and unstructured, with its meaning hidden in the vocal delivery. For example, the utterance "pain in belly" may either be a report to another care provider or a question to the patient, which is impossible to tell given the text alone. Without much syntactical structure, an intention classifier cannot afford to disregard the speech's acoustic features.

To address these intention detection challenges, we propose a multimodal deep learning structure for sentence-level intention classification using both text and audio data. The model uses manually transcribed speech text and its synchronized speech audio as input. We first preprocess the text data with word embedding [2] and extract mel-frequency spectral coefficients (MFSC) [3] from the audio. These inputs are fed to two independent convolutional neural networks (ConvNets) for feature extraction [3, 4]. The extracted features are then fused and used for the intention classification decision.

To demonstrate the model, we collected text and audio data from an actual trauma room during 35 trauma resuscitations, which are full of short, stressed sentences. The dataset contains 6424 transcribed sentences with their corresponding audio clips. The 6 intentions we used were: Directive, Report, Question, Clarification, Acknowledgement, and Response/Reply. With an 80-20 training-testing split, the model achieved 83.10% average accuracy, outperforming the text-only and audio-only models, and other similar models from other research. Our experiments also showed that ConvNets provide better-performing features than manufactured features such as fundamental frequency, pitch/energy related features, zero crossing rate (ZCR), jitter, shimmer, and mel-frequency cepstral coefficients (MFCC).

The contributions of our work are:

- A deep learning multimodal structure capable of automatically learning and using textual and acoustic features for intention classification.
- Detailed analysis and comparison of different modality and features that is commonly used for similar topic.
- A case study with actual medical application scenario, which indicated both the efficiency and drawbacks of the proposed system, this can be used as our future implementation reference as well as other similar applications.

The paper is organized as follows: Sect. 2 introduces related work, Sect. 3 describes our dataset, Sect. 4 details the model architecture, Sect. 5 presents the experimental results, Sect. 6 discusses model limitations, potential extensions, and concludes the paper.

2 Related Work

Previous research defined "intent" as understanding spoken language by converting sentences into representations of meaning [5]. Recently proposed methods in this field have used semantic analysis [6] and language understanding [7]. However, these methods only considered textual features, and ignored the acoustic information.

Because the same words may carry different intentions depending on the manner of speech, these acoustic features are critical to understanding the underlying meaning. Several different approaches were introduced using various features to help identify the meaning of human speech. Sentiment analysis tried to distinguish positive and negative attitudes [8]. This approach showed strong performance on document-level sentiment classification, but its binary (positive or negative sentiment) categories made it unhelpful in speech understanding. As an improvement, acoustic-based speech emotion recognition (SER) has been proposed [9, 10]. Different emotions reflect the speaker's attitude, but are too abstract for many scenarios and unhelpful to intention recognition. Therefore, we borrowed strategies from sentiment analysis and emotion recognition, and identified the intentions in our problem domain.

To combine the various feature types for understanding the underlying meaning, multimodal structures were introduced in related research fields. Early research demonstrated that the joint use of audio and text features achieved better performance in speech emotion recognition [11]. 2100 sentences with text and audio were collected from broadcast dramas. 33 original acoustic features and some specific emotional words were manually defined for predicting 7 emotions. However, the manufactured acoustic features and the specific keywords impeded their generalizability. A multimodal structure consisting of audio, visual, and textual modules was introduced to identify emotions [12]; manufactured features were selected from each module and the system was evaluated on three different emotion datasets (Youtube dataset, SenticNet, and EmoSenticNet). The results showed that using the visual features improved performance, but the system was still based on manually selected features. Recently, more complex multimodal systems using ConvNet-based visual and textual features were shown to outperform manufactured ones in emotion recognition [13, 14]. A ConvNet structure was used as a trainable textual feature extractor [13, 14], and another ConvNet structure extracted the visual features [14]. These systems were tested on the Multimodal emotion recognition dataset (USC IEMOCAP), multimodal opinion utterances dataset (MOUD), and Youtube dataset. Both systems demonstrated the power of ConvNet feature extraction. However, they still partially relied on manufactured acoustic features reduced by principal component analysis. The previous work also indicated that their system, although multimodal, relied heavily on text input [14]. We assume that the problem arose because they used manufactured acoustic features instead of automatic ones extracted by a ConvNet. To overcome this shortcoming, our proposed model uses ConvNet feature extractors for both the text and audio. Considering the limitations in the trauma scenario, it is not possible for us to capture the visual information from the human faces, so our multimodal system uses only text and audio. To demonstrate its robustness, we also tested it with inaccurately transcribed text input from a speech recognition engine.

3 Dataset

3.1 Data Collection

Our dataset was collected during 30 resuscitations at a level-1 trauma center in the north-eastern US. In medical settings, data collection devices must preserve privacy

and not interfere with medical tasks. Therefore, instead of using wearable Bluetooth microphones, we used a hands-free SONY ICD PX333 Digital Voice Recorder. It was placed on the Mayo equipment stand roughly 2.5 feet away from the trauma team leadership group of three clinicians, who mostly remained stationary during the resuscitation. We recorded on a mono channel with 16000 Hz sampling rate.

3.2 Transcribing and Labeling the Raw Data

We manually segmented the recorded audio data into sentence-level clips. To keep the audio quality, clips with overlapping speech or strong background noise (e.g. patient crying, electronic equipment noise) were removed. The final dataset contained 6424 audio clips. Each audio clip was then manually transcribed (unpunctuated) and labeled with an intention (Table 1). The "intention" represents the speaker's original purpose of saying the utterance. Six different intentions were defined based our dataset. For example, "Q" represents the speaker intends to inquire for information, "DIR" means the speaker plans to give an instruction or command to someone, "RS" indicates the speaker is responding to an inquiry. Table 2 shows example sentences with their intentions.

Table 1. Speech intention classes in the trauma resuscitation

Class	Intentions	Frequency
DIR	Directives (task assignment/instruction/command)	1172
RS	Response to an inquiry or request for information	934
Q	Inquiry/request for information	1256
RP	Report (report on patient status or results of an activity)	1045
CL	Clarification (request for retransmission of information)	1044
ACK	Acknowledgement	973

Table 2. Examples of speech intention from trauma resuscitation

Class	Sentence
DIR	Get her fluids going
RS	I don't know she's got three IVs on left side
Q	What's the IV access
RP	Bilateral TMs are clear
CL	You mean on the left
ACK	Yes alright I will

4 The System Structure

Our system structure is composed of four modules (Fig. 1). The data preprocessor formats the input text and audio into word vectors and Mel-frequency spectral coefficients (MFSC) maps for feature extraction. The feature extractor uses two ConvNets to learn and extract textual and acoustic features independently. The feature fusion layer balances and merges the extracted features through a fully-connected (FC) structure.

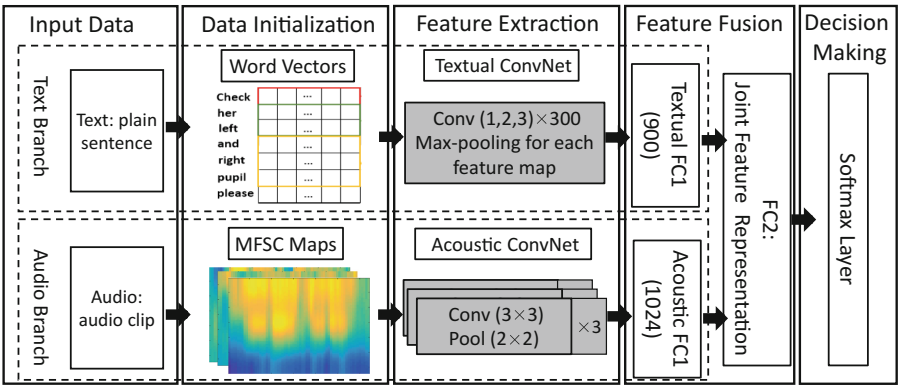


Fig. 1. The structure of our intention recognition model. FC = fully connected. The numbers in parentheses represent the number of neurons in each network layer.

The decision maker uses softmax (multinomial regression) to perform the multiclass classification.

4.1 Data Preprocessor

To generate the input layer for feature extraction, we used different initialization strategies for text input and audio input. We preprocessed the text into word vectors [2, 15, 16] and the audio into MFSC maps [3]. Word vectors are low-dimensional mappings describing semantic relationships between words [2]. A sentence is then a matrix with the word vector sequence as its rows [2]. Two different strategies exist for the initialization of the word vectors [2, 16]. We selected Mikolov's method [2] for its good performance on the semantic and syntactic learning for words. The word2vec embedding dictionary, trained on 100 million words from Google news [2], is the most commonly used word embedding tool. Since our sentences have varying lengths (between 1 and 26 words), we zero-padded all sentences to the longest length as suggested by others [4]. Each word was embedded into a 300-dimensional word vector using the word2vec dictionary, and unknown words were initialized randomly. Sentences were, therefore, 26×300 matrices (sentence length by word-vector length).

In human speech, different energy intensities and the variance of energy in different frequencies may reflect speaking manners. We represented the audio with time-frequency energy maps. Instead of the Mel-frequency cepstral coefficients (MFCCs) commonly used in speech recognition, we used Mel-frequency spectral coefficients (MFSC) to avoid the locality-compromising discrete cosine transform (DCR) [3]. We extracted the static, delta, and double delta MFSC for each audio for use as individual input channels. We empirically divided the 0–8000 Hz into 64 frequency bands. Each input frame contained 1 s of data (sampled at 40 ms with 50% overlap, following previous work [3, 17]), generating $64 \times n$ MFSC maps for an n-second clip. All MFSC maps were rescaled to 64×256 with bicubic interpolation. The initialized word vectors and MSFC maps were used as the input layers for the text branch and audio branch, respectively. As mentioned, the dimensions of the inputs data were 26×300 for the text branch and $64 \times 256 \times 3$ for the audio branch.

4.2 Feature Extractor

We used two ConvNets for feature extraction. One-dimensional filters were implemented for textual feature extraction (Fig. 2), following prior work [4]. As suggested in [4, 15], we applied multiple convolutional filters with different widths (1, 2, and 3) to capture phrases of varying length (Fig. 2). We chose these rather short filter widths because in our problem domain (trauma resuscitation) most of the speech is not in the form of full and grammatically correct sentences, but is instead in the form of short phrases. We empirically found that one convolutional and one max-pooling layer is sufficient for our application, similar to [4]. From each of the three filter sizes, 300 feature maps were then max-pooled, producing a 900-dimensional textual feature vector (comparable in size to the 1024-dimensional acoustic feature vector).

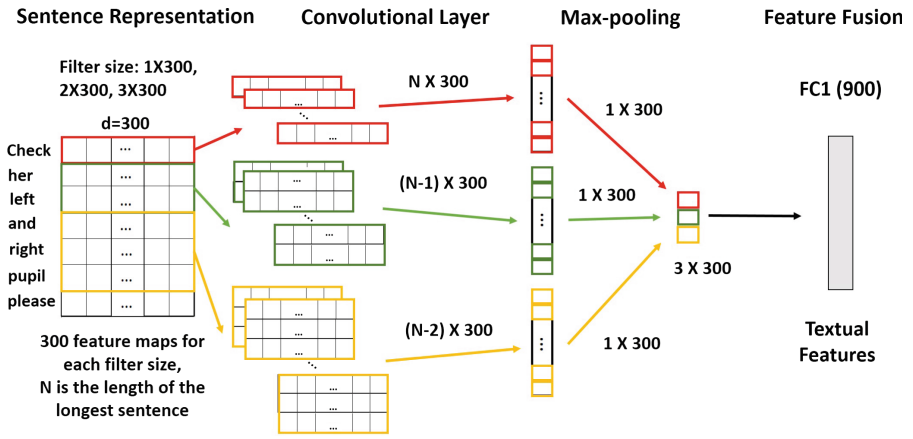


Fig. 2. Textual feature extraction ConvNet (CNN_T)

An eight-layer ConvNet structure was implemented to extract acoustic features from MFSC maps (Fig. 3). As suggested for VGG Nets [3, 17], we used 3×3 convolutional and 2×2 max-pooling filters with zero padding and chose $3 \times 3 \times 32$, $3 \times 3 \times 64$, $3 \times 3 \times 128$, and $3 \times 3 \times 128$ as the kernel sizes and the number of

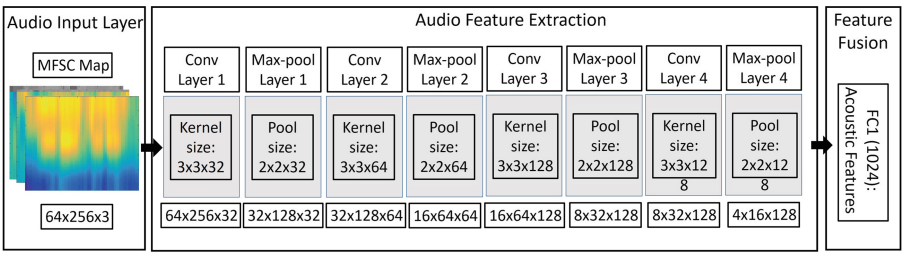


Fig. 3. Acoustic feature extraction ConvNet (CNN_A)

feature maps are determined empirically. Although deeper networks tend to perform better, we stopped at eight layers due to the hardware constraints.

4.3 Feature Fusion Layer

Two independent fully-connected layers concatenate the features from CNN_T and CNN_A (Fig. 1). To balance the two modalities, they were represented with 900 and 1024-dimensional vectors respectively. A fully-connected (FC) layer fuses the text and audio features into a joint feature representation ready for the final decision (Fig. 1). There are two main reasons for us to select feature-level fusion instead of decision-level fusion. First, because we used ConvNet structure to extract both the acoustic features and textual features, we do not need to apply the implementation method designed to deal with the heterogeneous data in feature-level fusion (such as the SPF-GMKL in [14]), which makes the feature-level fusion reasonable and convenient. Second, feature-level fusion outperformed decision-level fusion in previous research [14]. We decided to follow those successful implementations in similar applications.

4.4 Decision Making

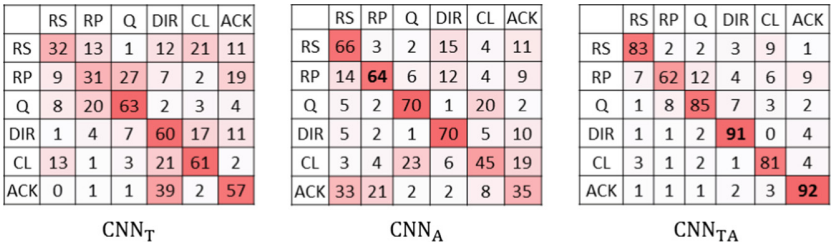
Predicting intention is multiclass classification, so we used a softmax decision-making layer to predict the intention class based on the 1924-dimensional feature vectors from the feature fusion layer (last layer in Fig. 1). Our experiments showed that having a deep neural network just before the softmax only slightly improved the performance. Considering the hardware costs, we used only a softmax layer in our decision-making module.

4.5 Implementation

We used Keras, a high-level TensorFlow-based neural network library [18], for model training and testing. As suggested previously [3, 17], we used the rectified linear unit activation function for all convolutional layers. We initialized the learning rate at 0.01 and used Adam optimizer to minimize the loss value [17]. Dropout function and 5-cross validation was applied during training to avoid overfitting.

5 Evaluation Results

We first compared the multimodal performance (CNN_{TA}) with that of text-only (CNN_T), and audio-only (CNN_A) models (Table 3). In our unimodal experiments, we used the same input data as during the multimodal training. As expected, the CNN_{TA} outperformed the other two because it exploited the strengths of both. The accuracy of the CNN_{TA} was 83.1% which is much higher than the CNN_T with 57.36 and CNN_A with 59.64. Our analysis of the confusion matrices (Fig. 4) found the following:



	RS	RP	Q	DIR	CL	ACK
RS	32	13	1	12	21	11
RP	9	31	27	7	2	19
Q	8	20	63	2	3	4
DIR	1	4	7	60	17	11
CL	13	1	3	21	61	2
ACK	0	1	1	39	2	57

CNN_T

	RS	RP	Q	DIR	CL	ACK
RS	66	3	2	15	4	11
RP	14	64	6	12	4	9
Q	5	2	70	1	20	2
DIR	5	2	1	70	5	10
CL	3	4	23	6	45	19
ACK	33	21	2	2	8	35

CNN_A

	RS	RP	Q	DIR	CL	ACK
RS	83	2	2	3	9	1
RP	7	62	12	4	6	9
Q	1	8	85	7	3	2
DIR	1	1	2	91	0	4
CL	3	1	2	1	81	4
ACK	1	1	1	2	3	92

CNN_{TA}

Fig. 4. The confusion matrix for different model structure (number is in percentage).

- The CNN_T differentiated well between classes with distinct contents (e.g. Q, CL). Both Q and CL have interrogative phrases such as "what is," "how about," and "do you" that appear very infrequently in the remaining classes.
- The CNN_A distinguished well between classes with different speaking manners (e.g. Q, DIR, RS, RP). Using the acoustic features improved the accuracy rate of Q and DIR. This was expected, as their acoustic features are different despite similar phrases and vocabulary. The performance for RS and RP was also strong, despite the text-level content variation, due to their relatively fixed speaking manners.
- The CNN_{TA} significantly outperformed the unimodal models, indicating that text or audio alone is insufficient for intention classification. Both datatypes compensate the weaknesses of each other, making their joint feature representation more useful for understanding the underlying meaning and intentions in speech.
- The accuracy on RP was comparatively low. This might have been because RP sentences vary widely in content, and are acoustically similar to other classes. After further analysis of the data, we found that the speakers often increased the pitch at the end of each sub-report during a summary report at the end of each process phase, in order to emphasize certain point. This made RP sound like a Q or CL. When ground truth coding for RPs, humans usually exploit contextual information (i.e. nearby sentences), indicating that context may improve intention classification.

Considering the infeasibility of manual transcription in speech applications, we also tested the CNN_{TA} model with text generated from automatic speech recognition (CNN_{SA}). Specifically, we used the text generated by the *Microsoft Bing Speech API* along with the original audio clip as the input to our system. The Bing API could not achieve human accuracy for our medical dataset due to the noise, and had a 26.3% word error rate. However, our results showed that despite the transcription inaccuracies, CNN_{SA} predictions were only 6% less accurate than those of CNN_{TA} (Table 3). This indicates that word-level accurate text is not as essential to intention classification as loose sentence structure and keywords, which is a significant finding. Using text-only model to detect activity or motion of trauma team members is very hard because high ambient noise made speech recognition difficult. Our experiment showed that including the acoustic features improved the accuracy of the intention recognition and the multimodal system is not very sensitive to the textual features. Hence, our model could be used for capturing speech intention in noisy environments such as trauma room.

Table 3. Comparison of model accuracies

Model	Input data	Accuracy (%)
CNN_T	Manually generated script	57.36
CNN_A	MFSC map	59.64
CNN_{TA}	Manually generated script and MFSC map	**83.1**
CNN_{SA}	Machine generated script and MFSC map	77.21
HSF	Manufactured features	48.73

We also compared the accuracies of individual or combined modalities using ConvNet-learned versus manufactured features. A model with human-selected feature extraction (HSF) using the same data split and softmax configuration was trained on several widely used manufactured features including fundamental frequency [19], pitch related features [20], energy related features [21], zero crossing rate (ZCR) [21, 22], jitter [21], shimmer [21], and Mel-frequency cepstral coefficients (MFCC) [22–24]. As suggested in [19, 20], we applied the statistical functions including Maximum, Minimum, Range, Mean, Slope, Offset, Stddev, Skewness, Kurtosis, Variance, and Median for these features. We normalized all the manufactured acoustic features using z-score normalization [22] and fed them into a softmax layer. The results showed that the features extracted by our multimodal structure's led to significantly better performance than the manufactured features (Table 3). In fact, even CNN_A alone outperformed HSF, further demonstrating the effectiveness of ConvNet-based feature selection. Although one could fine-tune the manually-selected features [21, 22], doing so would be highly laborious compared to automated ConvNet learning.

Finally, we compared our work with some similar research [12–14] (Table 4). We compared our model with several similar models in different application scenarios. This comparison showed that our model is competitive with some [13, 14], and outperformed others [12]. The previous model achieving the best performance [14] had a much simpler application than ours: it only classified 447 instances into two categories (positive/negative sentiment). Although MKL [13] achieved 96.55% accuracy in binary sentiment analysis (positive or negative), it relied heavily on visual features. Its accuracy dropped to a comparable 84.12% given only text and audio. In addition, their

Table 4. Comparison of our intention recognition model to similar research

Model	Input data	Classes	Accuracy (%)
Multimodal sentiment [12]	Manually selected video, audio, text features	3 classes	78.2
Convolutional MKL [13]	Text, manually selected acoustic features, video	2 classes 4 classes	84.12 76.85
Deep CNN [14]	Text, manually selected acoustic features, video	2 classes	88.6
CNN_{TA} CNN_{SA}	Manually generated script, MFSC map Machine generated script, MFSC map	6 classes	83.1 77.21

binary classification problem is much simpler than our six-class classification application. In fact, our six-class system even outperforms MLK's four-class emotion detection implementation.

6 Discussion and Conclusion

Although applicable and competitive, our proposed model can be improved in two aspects. Using wearable microphones would improve audio quality, which would increase the automatic speech recognition accuracy and lead to better predictions in the automated transcription. Using an LSTM to learn contextual features would also better discover features in text data [16, 25].

There are also inherent limitations to our current model. Our model performed well in scenarios such as trauma resuscitation, where there are few unessential words in relatively short sentences. However, it is difficult to detect the intention from fully-formed sentences, since the speaker may use multiple speaking manners in one sentence. This shortcoming may be solved with contextual intention recognition, which will be part of our future research. Lack of speaker-independency is another limitation. Even though there were work shifts in trauma room, our dataset only had voices from a limited set of persons and we did not consider speaker independency during training the model. Evaluating the model performance on the unknown voices should be further researched. Another limitation is the necessary removal of overlapping speech, which occurs very frequently. Around 28% of trauma speech has heavy overlapping speech. Then, to further improve speech intention recognition in the trauma environment, we must also consider solutions to the cocktail party problem.

To conclude, we restate this paper's contributions to the field:

- A framework that learns textual and acoustic features for intention classification.
- A comparison and analysis of our system performance on intention classification using multimodal, unimodal, and human-selected features.
- A system application in an actual medical setting that can be used as a reference for future study.

Acknowledgment. The authors would like to thank the trauma experts of the Children's National Medical Center, United State involved in this work. We also would like to thank three anonymous reviewers for their valuable comments and suggestions.

References

1. De Ruiter, J.P., Cummins, C.: A model of intentional communication: AIRBUS (asymmetric intention recognition with Bayesian updating of signals). In: Proceedings of SemDial 2012, pp. 149–150 (2012)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)

3. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1533–1545 (2014)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
5. Wang, Y.-Y., Deng, L., Acero, A.: Spoken language understanding. IEEE Sig. Process. Mag. **22**, 16–31 (2005)
6. Tur, G., Mori, R.D.: Introduction. In: Spoken Language Understanding. pp. 1–7 (2011)
7. Williams, J.D., Kamal, E., Ashour, M., Amr, H., Miller, J., Zweig, G.: Fast and easy language understanding for dialog systems with Microsoft language understanding intelligent service (LUIS). In: Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2015)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retr. **2**, 1–135 (2008)
9. Ayadi, M.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**, 572–587 (2011)
10. Minker, W., Pittermann, J., Pittermann, A., Strauß, P.-M., Bühler, D.: Challenges in speech-based human–computer interfaces. Int. J. Speech Technol. **10**, 109–119 (2007)
11. Chuang, Z.J., Wu, C.H.: Multi-modal emotion recognition from speech and text. Comput. Linguist. Chin. Lang. Process. **9**(2), 45–62 (2004)
12. Poria, S., Cambria, E., Howard, N., Huang, G.-B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174**, 50–59 (2016)
13. Poria, S., Iti, C., Erik, C., Amir, H.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM (2016)
14. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing (2015)
15. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing (2015)
16. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment TreeBank. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). vol. 1631, p. 1642 (2013)
17. Li, X., Zhang, Y., Li, M., Chen, S., Austin, F.R., Marsic, I., Burd, R.S.: Online process phase detection using multimodal deep learning. In: 2016 IEEE 7th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) (2016)
18. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado G.S.: Tensorflow: large-scale machine learning on heterogeneous distributed systems (2016). arXiv preprint arXiv:1603.04467
19. Busso, C., Lee, S., Narayanan, S.: Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Process. **17**, 582–596 (2009)
20. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Sig. Process. Mag. **18**, 32–80 (2001)
21. Kotti, M., Paternò, F.: Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. Int. J. Speech Technol. **15**, 131–150 (2012)

22. Wang, K., An, N., Li, B.N., Zhang, Y., Li, L.: Speech emotion recognition using Fourier parameters. IEEE Trans. Affect. Comput. **6**(1), 69–75 (2015)
23. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Sig. Process. **28**, 357–366 (1980)
24. Kamaruddin, N., Wahab, A., Quek, C.: Cultural dependency analysis for understanding speech emotion. Expert Syst. Appl. **39**, 5115–5133 (2012)
25. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing, vol. 1: Long Papers (2015)