Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues

Leimin Tian
School of Informatics, the University of Edinburgh
United Kingdom
s1219694@sms.ed.ac.uk

Johanna D. Moore
School of Informatics, the University of Edinburgh
United Kingdom
J.Moore@ed.ac.uk

Catherine Lai
School of Informatics, the University of Edinburgh
United Kingdom
clai@inf.ed.ac.uk

ABSTRACT

Emotions play a vital role in human communications. Therefore, it is desirable for virtual agent dialogue systems to recognize and react to user's emotions. However, current automatic emotion recognizers have limited performance compared to humans. Our work attempts to improve performance of recognizing emotions in spoken dialogue by identifying dialogue cues predictive of emotions, and by building multimodal recognition models with a knowledge-inspired hierarchy. We conduct experiments on both spontaneous and acted dialogue data to study the efficacy of the proposed approaches. Our results show that including prior knowledge on emotions in dialogue in either the feature representation or the model structure is beneficial for automatic emotion recognition.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**;

KEYWORDS

affective computing, emotion, multimodal, dialogue, LSTM

ACM Reference Format:
Leimin Tian, Johanna D. Moore, and Catherine Lai. 2017. Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues. In *Proceedings of 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3139491.3139497

1 INTRODUCTION

Emotion is an important part of information conveyed in dialogue. Human speakers recognize the emotions of their conversational partner and express their own emotions throughout a conversation using multiple modalities, such as tones of speech and hand gestures. Therefore, it is important for a virtual agent to recognize and react to the human user's emotions during a dialogue in order to achieve better interaction experience. However, human emotions in dialogue are subtle and complex, and recognizing emotions automatically from spoken dialogue remains a challenging task.

We identify a lack of prior knowledge as one factor limiting the performance of current emotion recognizers. Effectiveness of an emotion recognizer is largely influenced by what features it uses to represent the raw signals, and how these features are modelled in the recognition model. Recently, Deep Neural Networks (DNNs) are on the rise with studies showing that better feature representations and model structures can be learnt automatically by a DNN. In particular, the Long Short-Term Memory Recurrent Neural Network (LSTM) have attracted growing interest in emotion recognition research. However, compared to other recognition tasks, databases for emotion recognition are small in size because of the expensiveness of emotion annotation. The small amount of available training data is often insufficient for optimizing a complex DNN. Thus, instead of optimizing the emotion recognizer from scratch, we are motivated to identify better features and model structures inspired by prior knowledge on human emotions in dialogue. In particular, we propose features representing occurrences of DISfluency and Non-verbal Vocalisation (**DIS-NV**) in speech, and a HierarchicaL (**HL**) fusion strategy that uses more abstract or global features at higher layers of its hierarchy. To study the effectiveness of the proposed approaches, we conduct emotion recognition experiments on two databases of English dialogue: the AVEC2012 database of spontaneous dialogue, and the IEMOCAP database of acted dialogue. Here emotions are defined with the dimension of **A**rousal (activeness), **E**xpectancy (certainty), **P**ower (dominance), and **V**alence (positive/negative).

2 DIS-NV FEATURES

State-of-the-art features used for recognizing emotions in spoken dialogue have been focused on acoustic characteristics and lexical content of speech. However, such features can be noisy and contain information beyond emotions conveyed in the speech. These features, inherited from speech processing and sentiment analysis, often overlook the context of the emotion recognition task which, in this case, is a spoken dialogue rather than a monologue or a tweet. Psycholinguistic studies have suggested that disfluencies are indicators of speaker uncertainty and level of conflict in dialogue, while non-verbal vocalisations such as laughter are universal cues of human emotions. Therefore, we are motivated to extract features representing occurrences of three types of disfluency (filled pause, filler, stutter) and two types of non-verbal vocalisation (laughter, audible breath) in speech for recognizing emotions in spoken dialogue. Table 1 contains results on the spontaneous AVEC2012 database [1]. As we can see, DIS-NVs are effective predictors of emotions, especially for the Expectancy dimension which relates to speaker uncertainty. We also find that DIS-NVs contain emotion-related

ISIAA'17, November 13, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5558-2/17/11…$15.00
https://doi.org/10.1145/3139491.3139497

information in addition to the lexical content and speech acoustics.

Table 1: Recognizing emotions in spontaneous dialogue

Correlation Coefficients	A	E	P	V
DIS-NV	**0.250**	**0.313**	**0.288**	**0.235**
Lexical (PMI)	0.152	0.216	0.220	0.186
Acoustic (LLD)	0.014	0.038	0.016	0.040

To study the robustness of the DIS-NV features, we conduct cross-corpora experiments on the AVEC2012 and the IEMOCAP databases [3]. Our statistical analysis indicates missing dialogue complexity in acted dialogue. For example, laughter with negative Valence, such as embarrassed laughter, is missing in the IEMOCAP database. There is also more pronounced acoustic variation in acted dialogue, which suggests exaggerated acting. Moreover, there are fewer and shorter DIS-NVs in the acted IEMOCAP dialogue, which limits the effectiveness of DIS-NV features for recognizing emotions in the IEMOCAP database. These fundamental differences between spontaneous and acted dialogue suggest that in order to build virtual agents that can better recognize user's emotions "in the wild", we need to collect data in a more natural environment and build emotion recognizers trained with spontaneous dialogue.

3 HL FUSION

In current emotion recognition research, the LSTM model has achieved leading performance, and combining multimodal information is shown to yield improvements. There are two main fusion strategies used in current multimodal emotion recognition models: Feature-Level (FL) fusion that concatenates all features before performing recognition, and Decision-Level (DL) fusion that uses outputs of unimodal models to make final recognition decision. However, it is difficult to incorporate inter-modality differences in FL fusion, while intra-modality differences are overlooked in the decision making module of DL fusion. Therefore, we propose a HierarchicaL (HL) fusion strategy which uses a knowledge-inspired hierarchy for modelling both inter- and intra- modality differences. The HL fusion incorporates features that are more abstract or describe data at a bigger time interval at higher layers.

To study the effectiveness of HL fusion, we build multimodal LSTM models combining the DIS-NV features with benchmark acoustic and lexical features using FL, DL and HL fusion, respectively. We conduct emotion recognition experiments on both the AVEC2012 and the IEMOCAP database [2]. As shown in Table 2, multimodal models using the proposed HL fusion consistently outperform those using FL and DL fusion, and improve state-of-the-art performance of emotion recognition from spoken dialogue. However, the recognition performance is limited by a lack of training data.

Table 2: HL fusion for multimodal emotion recognition

F1-Score(%)	A	E	P	V
On the spontaneous AVEC2012 database				
FL	60.1	68.1	74.8	71.7
DL	56.6	63.3	73.5	68.0
HL	**61.8**	**69.2**	**76.2**	**72.4**
On the acted IEMOCAP database				
FL	55.2	#	50.8	47.2
DL	51.6	#	49.7	46.8
HL	**61.7**	#	**52.8**	**51.2**

4 CONCLUSIONS AND DISCUSSION

Our work indicates that it is beneficial for emotion recognition models to incorporate prior knowledge. Our results also illustrate that data aspects, especially dialogue type, can greatly influence the performance of recognizing emotions in spoken dialogue. Moreover, our analysis on DIS-NV in spontaneous and acted dialogue contributes to the understanding of the relationship between emotions in spoken dialogue and dialogue phenomena. Beyond recognizing emotions in spoken dialogue, we also apply the DIS-NV features and HL fusion to predict audience's emotions induced by movies, and find that the efficacy of the proposed approaches can be generalized to other emotion-related tasks [4]. The major limitation of our work and many state-of-the-art studies of emotion recognition is that most reported results have low performance with small differences between various approaches. Thus, it is unclear whether or not improvements in intrinsic emotion recognition experiments will translate to improvements in interaction quality when applying the emotion recognition models to a virtual agent dialogue system. In the future, we would like to study this by applying our emotion recognition models to a virtual agent dialogue system and conduct extrinsic experiments.

REFERENCES

[1] Johanna Moore, Leimin Tian, and Catherine Lai. 2014. Word-level emotion recognition using high-level features. In *Proceedings of 15th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'14)*. Springer, Kathmandu, Nepal, 17–31.

[2] Leimin Tian, Johanna Moore, and Catherine Lai. 2016. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. *Proceedings of 2016 Spoken Language Technology Workshop (SLT'16)* (2016), 565–572.

[3] Leimin Tian, Johanna D Moore, and Catherine Lai. 2015. Emotion recognition in spontaneous and acted dialogues. In *Proceedings of 6th International Conference on Affective Computing and Intelligent Interaction (ACII'15)*. IEEE, Xi'an, China, 698–704.

[4] Leimin Tian, Michal Muszynski, Catherine Lai, Johanna D Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2017. Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same?. In *Proceedings of 7th International Conference on Affective Computing and Intelligent Interaction (ACII'17)*. IEEE, San Antonio, Texas, USA, 28–35.