Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues

Citation for published version:

Digital Object Identifier (DOI):
10.1145/3139491.3139497

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
ICMI 2017 Satellite Workshop Investigating Social Interactions with Artificial Agents
Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues

Leimin Tian, Johanna D. Moore, Catherine Lai
School of Informatics, the University of Edinburgh
s1219694@sms.ed.ac.uk, J.Moore@ed.ac.uk, 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

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 [4] of spontaneous dialogue, and the IEMOCAP database [1] of acted dialogue. Here emotions are defined with the dimension of Arousal (activeness), Expectancy (certainty), Power (dominance), and Valence (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 [2]. 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 information in addition to the lexical content and speech acoustics.

Table 1: Recognizing emotions in spontaneous dialogue

<table>
<thead>
<tr>
<th>Correlation Coefficients</th>
<th>A</th>
<th>E</th>
<th>P</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIS-NV</td>
<td>0.250</td>
<td>0.313</td>
<td>0.288</td>
<td>0.235</td>
</tr>
<tr>
<td>Lexical (PMI [3])</td>
<td>0.152</td>
<td>0.216</td>
<td>0.220</td>
<td>0.186</td>
</tr>
<tr>
<td>Acoustic (LLD [4])</td>
<td>0.014</td>
<td>0.038</td>
<td>0.016</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Table 2: HL fusion for multimodal emotion recognition

<table>
<thead>
<tr>
<th>F1-Score(%)</th>
<th>A</th>
<th>E</th>
<th>P</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>55.2</td>
<td>#</td>
<td>50.8</td>
<td>47.2</td>
</tr>
<tr>
<td>DL</td>
<td>51.6</td>
<td>#</td>
<td>49.7</td>
<td>46.8</td>
</tr>
<tr>
<td>HL</td>
<td>61.7</td>
<td>#</td>
<td>52.8</td>
<td>51.2</td>
</tr>
</tbody>
</table>

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 [7]. 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


