Multimodal Databases of Everyday Emotion: Facing up to Complexity

Ellen Douglas-Cowie¹, Laurence Devillers², Jean-Claude Martin², Roddy Cowie¹, Suzie Savvidou¹, Sarkis Abrilian², Cate Cox¹

¹ Queen’s University Belfast; UK ² LIMSI-CNRS, France

e.douglas-cowie@qub.ac.uk, {devil, martin, abrilian}@limsi.fr

Abstract

In everyday life, speech is part of a multichannel system involved in conveying emotion. Understanding how it operates in that context requires suitable data, consisting of multimodal records of emotion drawn from everyday life. This paper reflects the experience of two teams active in collecting and labelling data of this type. It sets out the core reasons for pursuing a multimodal approach, reviews issues and problems for developing relevant databases, and indicates how we can move forward both in terms of data collection and approaches to labelling.

1. Introduction

It is a sound rule that data should be as simple and orderly as they can be. The risk is that simplification may discard core information. As Schlossberg observed, some questions about cars cannot be answered with the brake on [1].

For the last few decades, most research on speech and emotion has adopted a working hypothesis that the essentials are represented in drastically simplified data – mainly short utterances with neutral content produced by actors [2]. But despite a large amount of work, and some consensus, the approach has not translated well into applications [3], suggesting that the traditional simplifications may indeed have discarded essentials.

If so, then there is little option but to address more complex data. Two targets stand out. One is to study more natural data even if it means sacrificing control. The other is to study speech in the context of other channels. The two are linked, because the norm is for speech to be one of several channels that alternate or overlap to signal emotion – others include verbal content, facial expression, gesture, and action. It is revealing that given only the speech from that kind of interaction, the natural reaction is to wonder what was happening on the other channels.

This paper reports progress in constructing databases that present speech in the context of natural interactions where emotion is conveyed via multiple modalities, and using them to explore how the modalities may interact. Two databases are considered, the Belfast naturalistic database [4] and the EmoTV database [5][6].

2. The databases: a brief description

Both databases exploit TV as a source of raw material. The Belfast database is in English and consists mainly of sedentary interactions, from chat shows, religious programs, and discussions between old acquaintances. The EmoTV corpus is in French and also draws material from TV interviews, but uses episodes with a wider range of body postures and more monologue, such as interviews on the street with people in the news. Both cover a wide range of positive and negative emotions and of emotional intensities. The Belfast database consists of 125 subjects. For each subject there are at least two video (audio-visual) ‘clips’ or sequences lasting from 10-60 secs: one is of the subject in a relatively neutral state, the other is of the same subject in an emotional state. The EmoTV corpus consists of 48 subjects. There are 51 video clips or sequences (again audio-visual), each of which shows the subject in an emotional state.

Labelling of emotional content has been completed for the Belfast database and is ongoing for EmoTV. Seven raters rated each clip in the Belfast database on two levels. Dimensional rating used the FEELTRACE tool to record perceived emotional state in terms of two dimensions derived from psychological theory - activation (from highly active to deeply passive) and evaluation (from strongly positive to strongly negative). Details of the method and the theory behind it are given in [7]. Categorical rating involved applying labels from a list of 16 terms (sad, angry, amused etc). EmoTV uses ANVIL as a platform and the coding scheme uses both verbal categorical labels and dimensional labels (intensity, activation, self-control and valence). Refinements include labels for the emotional context, and a coarse temporal description of intensity variation on each emotional segment, and others described below.

Although the databases have much in common, the teams have taken different approaches to studying the contributions of different modalities, the LIMSI team working with the discrete categorical element of EmoTV, the Belfast team working with the continuous dimensional element. These offer complementary insights into the particular contribution of speech to the recognition of emotion in a multimodal context.
3. A category-based approach

In order to study the influence of the modalities on the perception of emotions, two raters used the Anvil tool [8] to annotate all the videos of the EmoTV[5] for perceived emotions in three conditions: 1) Audio only, 2) Video only and 3) Audio-visual.

As a first step towards finding an appropriate set of emotional labels, the two annotators labelled the emotion they perceived in each emotional segment by selecting one label of their choice (free choice). This resulted in 176 fine-grain labels. These were classified into the 14 broader categories that appear on the axis of figure 1.

Even after the reduction to 14 classes, the inter-coder agreements on emotion labels were low. The kappa statistics were: 0.37 for audio and video (281 segments), 0.43 for video only (295 segments), and 0.54 for audio only (181 segments). Contrary to expectation, agreement was lowest in the multimodal case. That is a first indication that the relationship between modalities is not straightforward.

Figure 1 shows how agreement on the emotion categories divided across the channels. The chance estimate is based on the assumption that raters agree a priori on the relative frequencies of the labels, but allocate them at random to particular segments. Conveniently, that means the chance level of agreement for a label equals the a priori probability of assigning it.

The data show that low agreement in the multimodal case is a general phenomenon – the audio-visual channel gave the lowest agreement for half of the categories, and around or below chance in a third. Positive/Negative disagreements were given particular attention, and they vary similarly with modality: again, they are highest, at 11%, in the audio-visual case, intermediate with video only (at 7%), and lowest in the audio only condition, at 3%. The categories are ordered by valence to show a suggestive pattern. Excluding the four leftmost categories, which are rare, agreement is highest at the extreme valences, positive or negative.

These findings were followed up by close examination of each disagreement case. The low kappa measures do not appear to be due to bad labelling but are rather a reflection of the nature of real-life data. A large part of the data consists of blended emotions and contradictory multimodal cues, by example cry to bring relief, “looking contented” despite a deception. Different modalities are often – but not always – associated with different aspects of a complex state (for instance tears with signs of relief, a positive face with deeply negative words).

The prevalence of these complex states is critical both for approaches to labelling and for understanding the contributions of different channels, and so the issue was followed up in a second study on only audio-visual condition. Three labellers used a more sophisticated scheme, one of whose elements was describing complex global pattern of emotion in terms of five categories – Single label, blended, masked, sequential (very rapid transition between overt patterns of expression, which suggests that both underlying states are in some sense present) and cause-effect (one kind of emotional event, eg crying, leads to another, e.g. relief).

The data show that low agreement in the multimodal case is a general phenomenon – the audio-visual channel gave the lowest agreement for half of the categories, and around or below chance in a third. Positive/Negative disagreements were given particular attention, and they vary similarly with modality: again, they are highest, at 11%, in the audio-visual case, intermediate with video only (at 7%), and lowest in the audio only condition, at 3%. The categories are ordered by valence to show a suggestive pattern. Excluding the four leftmost categories, which are rare, agreement is highest at the extreme valences, positive or negative.

These findings were followed up by close examination of each disagreement case. The low kappa measures do not appear to be due to bad labelling but are rather a reflection of the nature of real-life data. A large part of the data consists of blended emotions and contradictory multimodal cues, by example cry to bring relief, “looking contented” despite a deception. Different modalities are often – but not always – associated with different aspects of a complex state (for instance tears with signs of relief, a positive face with deeply negative words).

The prevalence of these complex states is critical both for approaches to labelling and for understanding the contributions of different channels, and so the issue was followed up in a second study on only audio-visual condition. Three labellers used a more sophisticated scheme, one of whose elements was describing complex global pattern of emotion in terms of five categories – Single label, blended, masked, sequential (very rapid transition between overt patterns of expression, which suggests that both underlying states are in some sense present) and cause-effect (one kind of emotional event, eg crying, leads to another, e.g. relief).
Figure 2 summarises the types of global pattern reported. It uses majority voting, i.e., a description is accepted if at least 2 of the 3 raters use it.

On that criterion, the proportion of segments with no agreed label is 21%. The key point is that segments rated as showing a specific non-basic pattern (33%) were nearly as common as segments showing a pure emotion (46%). In that situation, asking raters to assign a single label is not only unlikely to yield agreement: it misrepresents the situation, and pre-empts key questions about the roles of different modalities.

The follow-up study also asked raters to identify the cues that they considered to be important in reaching their rating. Table 1 summarises the results. The main point is that both audio and facial cues were involved in the great majority of segments. There were in fact no cases where only one modality was judged relevant.

<table>
<thead>
<tr>
<th>#segments</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>44</td>
</tr>
<tr>
<td>Eyes</td>
<td>40</td>
</tr>
<tr>
<td>Mouth</td>
<td>31</td>
</tr>
<tr>
<td>Head</td>
<td>29</td>
</tr>
<tr>
<td>Brows</td>
<td>16</td>
</tr>
<tr>
<td>Gestures</td>
<td>10</td>
</tr>
<tr>
<td>Torso</td>
<td>4</td>
</tr>
<tr>
<td>Shoulders</td>
<td>3</td>
</tr>
<tr>
<td>TOTAL</td>
<td>177</td>
</tr>
</tbody>
</table>

Table 1: Frequency with which different cue types were reported relevant to emotion judgments (a cue is accepted if at least 2 of the 3 raters use it).

4. Roles of different modalities in the Belfast database

The Belfast database shows a range of phenomena broadly similar to those which have been described above. However, the dimensional descriptions are in a continuous form, and that offers a way of drawing broad but robust conclusions.

Savvidou [9] selected 20 clips for a study of the contributions made by different modalities. They were chosen to span the full range of emotional states in the database, and to be reasonably representative in terms of the spread of ratings each produced. Four versions of each clip were constructed: Audio-visual (unmodified); Visual only (sound suppressed); Full audio (visual suppressed); and filtered audio (a notch filter applied to render the speech content unintelligible with minimal effect on prosody and voice quality).

12 raters participated, each using the FEELTRACE tool [7] to record their continuous impressions of each version in terms of activation and evaluation. Figure 3 summarises the key results by plotting average ratings based on individual modalities alongside ratings of the audio-visual version (which is taken as the best available estimate of the speaker’s true state).
Considering activation, it is apparent that clips involving extremes of activation are rather similarly rated in all versions – suggesting that signs of extreme activation are present in all the modalities.

However, the visual only version gives rise to ratings that are far from neutrality even when the person’s activation is actually (judging on all the evidence available) around the default level. Hence the auditory channel seems specifically to provide evidence of moderate activation.

Prosody in a broad sense (provided by the filtered version) and words (provided by the full audio) seem to make distinct contributions.

These findings match and reinforce the pattern for neutral labellings shown in figure 1. Both indicate that visual information alone is poor at conveying when people are in a relatively unemotional state – which is a critical kind of information. Agreement on extremes of activation also matches the consistency of EmoTV ratings with respect to despair, anger, joy and exaltation.

Regarding evaluation, the dominant pattern is that clips which are rated very positive in the full versions tend to receive more neutral ratings in the partial versions. The pattern is most marked with the filtered audio material, suggesting that prosody (in a broad sense) does not usually in and of itself convey that a person’s emotional state is highly positive. This in not to say that prosody cannot be adapted to convey highly positive emotion: there is evidence that it can in the form of one case where ratings of audiovisual and audio filtered versions coincide. On the other hand, it appears that in a multimodal setting, highly positive prosody (sacrificing precision to brevity) is to be used sparingly even when the person is actually very positive.

5. Discussion and conclusion

Unpicking the roles of different modalities in everyday emotional communication is a delicate business. We have stressed that it depends on appreciating what everyday emotion is like, and finding tractable ways of representing it. We are actively involved in extending the techniques available and the adequate annotation protocols. The LIMSI team has developed ‘soft’ codings as a way of representing everyday emotion mixes.

The new annotation scheme means that two labels per segment (Major and Minor) can be used to describe an emotional state [5][10]. A new emotional descriptor can be computed considering a soft vector combining Major and Minor emotions rated by coders (weights 2 for Major, 1 for minor). That scheme was applied in the second experiment, and resulted in 90% of the soft vectors producing a “winner” label, which gives better reliability scores than using one single label (agreement for the Major label only was 77%).

Embodied Conversational Agents (ECAs) are both an important application and an invaluable tool for controlling modality combinations. Work is in progress using EmoTV annotations to achieve replay of natural mixed emotional behaviour by ECA [11]. Both teams are also working on the opportunity that the databases provide to make cross-cultural comparisons.

6. Acknowledgements

This work was partly funded by the FP6 IST HUMAINE Network of Excellence (http://emotion-research.net).

7. References