

The description of naturally occurring emotional speech

E Douglas-Cowie[†], R Cowie[†] and M Schroeder[‡]

[†] Queen's University of Belfast, UK

[‡] DFKI, Germany

E-mail: e.douglas-cowie@qub.ac.uk, r.cowie@qub.ac.uk, schroed@dfki.de

ABSTRACT

Most studies of the vocal signs of emotion depend on acted data. This paper reports the development of a vocal coding system to describe the signs of emotion in naturally occurring emotion. The system has been driven by empirical observation, not by a priori assumptions based on acted or laboratory data. The data used to develop it is the Belfast Naturalistic Database. The system takes a multi-level approach to coding, starting broad brush and moving through progressive layers to finer resolution. The first level uses broad categories which apply to each clip as a whole. Thereafter it uses a tiered approach, starting with an outer tier of relatively coarse descriptors and progressing through successive tiers to more detailed descriptors associated more precisely with locations in a clip. The coding system shows that there are vocal signs of naturally occurring emotion which have not been picked up before in acted data.

1. INTRODUCTION

This paper is concerned with the vocal markers of emotion in naturally occurring emotional speech. It describes a coding system developed to capture the properties of speech that auditorily appear to signal emotionality. In one of the few and biggest naturalistic emotional speech databases (The Belfast Naturalistic Database, see [1]).

The topic is of considerable relevance to current developments in the speech engineering world. As speech recognition and synthesis systems have become reliable in dealing with fundamental problems, there has been increasing interest in developing more sophisticated systems, including those that take account of the mood and emotion of the speaker. Unfortunately most of our knowledge of the vocal signs of emotion is based on acted or laboratory data (see [1] for review), and there is therefore a real question about whether systems trained on this data would cope adequately with naturally occurring emotional speech. The authors of this paper have been working on the development of an emotional recognition system in the EU PHYSTA and ERMIS projects [2], [3], and were prompted by this concern to develop a database of naturally occurring emotion. The database is audiovisual to allow work on the recognition of both vocal and visual signs of emotion and their interaction.

The vocal coding system is driven by empirical observation of naturally occurring emotional speech, not by a priori assumptions which are based on actors simulating archetypal emotions. A template has been established iteratively, using a representative selection from the database to begin with and then applying it to the whole database.

The approach to coding, both for speech and the associated emotional states is essentially broad-brush. Speech is coded top down for whole emotional episodes (of considerable length) rather than bottom up in a linear fashion, segment by segment. Emotions are coded in terms of broad dimensional representations of emotion (negative versus positive, active versus passive) rather than into specific and discrete categories. This broad brush approach is particularly well-suited to engineering applications, in that it pulls out core features that can be used for modeling or training with some chance of success. Approaches to natural emotional data based on categorical representation of emotion and segment by segment coding of speech tend to result in diverse and complex patterns which do not show statistically robust effects [1].

However the coding systems both for speech and emotion also allow for finer resolution in terms of the time domain of an emotional episode and in terms of the nature of the vocal characteristics (going in progressive stages from outer tier descriptors which are relatively coarse to inner tier descriptors which deal with finer distinctions).

2. THE BELFAST DATABASE

The Belfast Naturalistic Emotional Database consists of 298 audiovisual clips from 125 speakers, 31 male, 94 female. Emotional clips are episodes which appear to provide within themselves at least most of the context necessary to understand a local peak in the display of emotion and to show how it develops over time. For each speaker there is at least one clip showing him or her in a state judged relatively emotional, and also one clip in a state that the selector judged relatively neutral. Clips range from 10-60 secs in length. The clips are stored as MPEG files, with audio data extracted into .wav files

The clips were selected to represent emotional states that occur in everyday interactions as well as more archetypal examples of emotion such as full-blown anger or fear. Two

main sources were used— television programmes and studio recordings carried out by the Belfast team. The television programmes include chat shows, religious programmes, programmes tracing individuals' lives and current affairs programmes. Studio recordings were based on one-to-one interactions between a researcher with fieldwork experience and close colleagues or friends. The aim was to cover topics that would elicit a range of emotional responses.

The psychological coding of the clips is based primarily on a dimensional approach to emotion, although there is also a categorical representation based on a 'basic' emotion vocabulary of 16 terms. Two dimensions, activation and evaluation are known to capture a relatively large proportion of emotional variation. A computer program called FEELTRACE was written to let users describe perceived emotional content of the clips in terms of those dimensions. The space was represented by a circle on a computer screen, alongside a window where a clip was presented. The vertical axis represented activation, the horizontal axis evaluation. Raters used a mouse to move a cursor inside the circle, adjusting its position continuously to reflect the impression of emotion that they derived from the clip. The co-ordinates of the cursor are recorded at regular intervals and averaged over the episode to give a quantitative measure for the episode as a whole (giving a broad brush representation of the emotional state of the speaker), but they can also be displayed to give a quantitative representation of the emotional state at any one time (thus allowing for finer resolution in the time domain). The database clips have been rated by 5 subjects and there is reasonable inter-rater reliability (see [1]).

3. THE VOCAL CODING SYSTEM

3.1 General outline

As noted, the system is driven by empirical observation of naturally occurring emotional speech, and a template has been established iteratively. To ensure that the descriptors allow the main trends to be seen, the system takes a multi-level approach to coding, starting broad brush and moving through progressive layers to finer resolution. The first level uses broad categories which apply to each clip as a whole. Thereafter it uses a tiered approach, starting with an outer tier of relatively coarse descriptors and progressing through successive tiers to more detailed descriptors associated more precisely with locations in a clip. The coding system focuses primarily on the speech signal, but it also records some wider communicative aspects of emotion, particularly where these interact with the speech signal (e.g. gestures such as mouth occlusion may impede the speech signal, and the management of turntaking in emotional environments may impact prosody).

The outer tier of descriptors contains seven broad generic groupings – impaired communication, pitch, timing, volume, voice quality, paralinguistic features and articulatory features. The coder marks whether the clip

contains features (in one or more of these generic groupings) that appear to signify emotionality. The coder does not look for specific emotions, just the presence of emotionality. Each of the generic groupings is associated with three inner tiers. The second tier specifies intermediate categories, the third specifies the precise nature of each category and the fourth specifies degree.

The coding system is being automated so that the coder is presented with the options in an ordered way, and responses can be saved in a format that facilitates analysis.

3.2 Specific descriptors

This section describes the seven generic groupings and their tiers and associated descriptors in more detail. There is some overlap and repetition across the groupings, particularly between the descriptors in the impaired communication grouping and the other groupings. This is because some of the descriptors may function differently in two contexts. They may contribute to impaired communication and thus belong to that generic grouping, but they may also form part of a phonetically coherent grouping of vocal indicators of emotionality. For example slurred articulation may lead to problems of intelligibility (and thus impair communication), but it is also part of a wider coherent grouping of articulatory features which mark emotionality. An automated system should be able to cross reference as appropriate.

Impaired communication

This outer tier descriptor is applied to a clip when emotionality leads to impairment of verbal communication or threatened breakdown of verbal communication. Impairment is judged to take place if there are problems of intelligibility or audibility, or if there are unacceptable levels of disruption to communication, or if there are substantial problems of social acceptability. Impairment to verbal communication may arise from low level problems of a paralinguistic or gestural nature, or from the breakdown of higher level rules governing the organization of discourse and conversation. There may be interactions between the high and low level problems, e.g. sobbing may lead to disrupted discourse. Impairment or breakdown may be of short or long duration.

The second tier is made up of seven categories –paralanguage, volume, voice quality, articulation, discourse structure, conversation management, and gestural behaviour. The third tier is the descriptors associated with each of these categories. Examples of key descriptors are given in brackets below: paralanguage (non linguistic sounds – reflexes and voice qualifications; exclamations and interjections whose primary function is to carry affect rather than propositional content); volume (too soft, too loud); voice quality (whisper, creak, tense voice); articulation (slur, stutter, misarticulation); discourse structure (unnecessary sequential repetition of words or phrases, disruption to word order or syntactic patterns

leading to discontinuity); conversation management (overlap, breakdown into long periods of silence, denying conversational participant a turn by speaking for too long); gestural behaviour (mouth occluded, face averted/head down).

The fourth tier specifies degree on two levels (i) the markedness or intensity of the vocal feature and (ii) the time domain of the feature – global (i.e. occurring across substantial parts of the clip or local (i.e. attached to phrase or word or segment length utterances). This fourth tier is the same for all generic groupings.

Pitch and volume

These are considered together here since the descriptors follow a similar structural organisation. The outer tier descriptors are applied to a clip when emotionality leads to pitch or volume being distinctive or salient auditorily. Distinctiveness can be tied to low level markers (e.g. range, height) or high level markers (particular patterns) and can extend across a substantial part of an episode or be more localized. The second tier is made up of four categories – height, range, variability and patterns. The third tier descriptors specify the precise nature of each category. Key examples of descriptors for both pitch and volume categories are given in brackets below: height (raised, lowered); range (wide, narrow), variability (volume or pitch change salient, level volume or pitch maintained). For pitch, the category ‘patterns’ includes descriptors which specify pitch direction at syllable or phrasal level. For volume, the category ‘patterns’ includes a descriptor which specifies inconsistency (loud and soft alternation). For the fourth tier, see above.

Timing

This outer tier descriptor is applied to a clip when emotionality leads to timing being distinctive or salient auditorily. Distinctiveness can be tied to low level markers (e.g. rate) or high level markers (e.g. patterns of pausing) and can extend across a substantial part of an episode or be more localized. The second tier is made up of three categories – rate, pausing and rhythm. Key associated third tier descriptors are: rate (fast, slow); pausing (frequent pausing, too few pauses, pauses too long, disruptive pauses), rhythm (extended excessive stressing, sporadic strongly stressed syllables). For the fourth tier, see above.

Paralanguage

This outer tier descriptor is applied to a clip when emotionality leads to sounds of a non linguistic nature or function taking over from, interrupting or being superimposed upon words. Instances can extend across a substantial part of an episode or be more localized. The second tier is made up of three categories – reflex sounds, voice qualifications and exclamations whose primary function seems to be to carry affect rather than propositional content. As such, they often have exaggerated prosodic realization (e.g. vowel elongation, exaggerated pitch movement) Associated third tier descriptors are given

in brackets below: reflex sounds and voice qualifications (sniffing, gulping, gasping, uneven breath control, sobbing, laughing, tremulous voice, break in voice); exclamations (aaw, yea, oh, ow, gosh). For the fourth tier, see above.

Voice Quality

This outer tier descriptor is applied to a clip when emotionality is marked by a distinctive voice quality. Instances can extend across a substantial part of an episode or be more localized. The second tier is made up of two categories – overall muscular settings and laryngeal/pharyngeal settings. Third tier descriptors are: overall muscular settings (overall muscular tension, overall muscular laxness); laryngeal/pharyngeal settings (creak, breathiness, whisper). For the fourth tier, see above.

Articulation

This outer tier descriptor is applied to a clip when emotionality is marked by instances of impaired or distinctive articulatory movement or setting. Instances can extend across a substantial part of an episode or be more localized. The second tier is made up of three categories – impaired articulation, articulatory timing and articulatory settings Associated third tier descriptors are: impaired articulation (slur, stutter, misarticulation); articulatory timing (too fast, too slow); articulatory settings (excess sibilance or aspiration, hyperarticulation).

4. THE VOCAL SIGNS OF NATURAL EMOTION: SOME OBSERVATIONS

Preliminary analysis suggests that the descriptors discriminate well between emotional and neutral states. A full analysis of the vocal signs of particular emotional states has not yet been completed, and coding has so far only been carried out by one coder. However, even if we cannot tie the vocal descriptors to particular emotional states at this stage, we can still make some interesting observations about the nature of vocal signs in naturally occurring emotion. The descriptors included in the coding system are all drawn from iterative exploration of the naturalistic data in the Belfast Emotion Database, and, as such, are real enough. A number of interesting observations can be made.

Most importantly, the coding system shows that there are vocal signs which have not been picked up before in the coding or description of acted or laboratory induced emotion. A number warrant mention.

Impaired communication is a category rarely mentioned in the literature. Yet our evidence suggests that it is a key descriptor, particularly for fairly intense emotional episodes [4]. Negative passive emotional states, for example, contain a good deal of impaired communication produced by sobbing and break in the voice. Less obvious perhaps is the fact that intense emotion of various kinds can simply lead to silence and thus result in a breakdown of communication. Intense deep happiness, for example, can

lead to periods of silence as can intense sadness. Intense anger is marked by impaired conversational management, especially overlap, and accompanying raised volume and pitch.

It is not surprising that impaired communication is not highlighted by laboratory or acted data. Much laboratory/acted data is based on monologue (often read) whereas much natural emotion tends to occur in interactive situations. Breakdown of communication is readily visible in interactive contexts when there are failures between participants to be understood or to co-operate conversationally in the expected ways.

Other vocal signs not usually drawn to our attention in acted/laboratory data include an interesting group of articulatory features, particularly impaired articulation and articulatory settings. Slur, stutter and misarticulation are examples of impaired articulation that may not occur in acted data, but, on initial analysis, seem to occur quite regularly in negative active emotional states. Articulatory settings tend not to be mentioned in the literature, except for hyperarticulation, but excessive sibilance occurs frequently in positive passive states (e.g. serenity). Excessive sibilance also often seems to be part of the wider overall muscular setting of laxness.

In terms of pitch, volume and timing which are well recognized as markers of acted emotion, the natural database suggests that there are aspects of these that occur in natural emotion, but are not reported for acted data. Examples are successive repetition of a particular tune shape, alternation of volume levels (raised followed by lowered) and alternation of rate (slow versus fast). These features are all characterized by juxtaposition of patterns (whether similar or contrasting). It may be that such patterns do occur in acted data, but that analysis has focused on more global measurements and missed them, or that they simply do not exist.

5. CONCLUSIONS

There is still some way to go in the identification of the vocal signs of naturally occurring emotion. The vocal coding system described here needs some refinement and must be tested thoroughly. Correlations between emotional states in the database and the vocal signs need to be identified. However the development reported in this paper mark a major step forward in our understanding of a relatively unexplored field.

REFERENCES

- [1] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, "Emotional speech: Towards a new generation of databases", *Speech Communication*, vol. 40. pp. 33-60, 2003.
- [2] <http://www.image.ntua.gr/physta/>

[3] <http://www.image.ntua.gr/ermis/>

- [4] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech", *Speech Communication*, vol. 40. pp. 5-32, 2003.