Labelling natural conversational speech data

Nick Campbell

ATR Human Information Sciences
“Keihanna Science City”, Kyoto 619-0223
nick@atr.co.jp

Abstract

This paper describes the set of labels to annotate “voice quality” and “expressiveness” in a large corpus of natural conversational speech. It describes a modification to the Trancriber software which facilitates the rapid labelling of conversational speech. It argues that the simplistic notion of a single setting for “emotion” for use in a paralinguistic speech synthesiser may need rethinking.

1. Introduction

There is growing interest in the topic of emotional speech and its applications in speech technology [1], but in this paper we question whether ‘emotion’ is the most appropriate term, since we presume the speaker's *state* to be of less interest than his or her *intentions* and *relationships*. For spoken language processing, we perhaps need to know more about how the speaker relates to the listener, and to the linguistic content of the message, than how the speaker feels at any given time.

In the JST/CREST ESP Project [2] we have to date collected more than 250 hours of unconstrained spontaneous speech from a range of subjects using two collection paradigms. The first is completely uncontrolled for content, with volunteers telephoning each other at weekly intervals to talk freely for half-an-hour per session, for a period of ten weeks [3]. The second employs volunteers who record their daily spoken interactions for extended periods throughout each day [4].

The goal of this work has been two-fold; a) to provide a knowledge-base for research into speech and emotion (or “expressiveness”), and b) to provide a source database for expressive speech synthesis. Our guidelines for the labelling of the speech data were therefore clear: to mark the differences in speaking style and to identify the variants with a unique set of labels. The criterion in case of doubt is whether a given unit (a waveform segment) could be used in place of another given unit in a concatenated synthesised utterance, without changing the perceived meaning of the utterance. “Meaning” is here defined not just in terms of lexical content, but also in terms of paralinguistic information and pragmatic force.

2. Labelling speech characteristics

We distinguish 3 categories of label, to indicate *speaker state, speaking-style,* and *voice-quality* characteristics respectively. Labels are determined subjectively by an experienced labeller, after listening to the speech several times, to indicate how each section of the speech was perceived.

自然音声対話データのラベル化
ニック・キャンベル　ATR 人間情報科学研究所.

Table 1. Labels for describing supra-linguistic information

level 1	STATE	(about the speaker)
purpose	a speech-act/CA label (open-class)	
confidence	6-point scale from +3 to -3, omitting 0.	
emotion	happy / sad / angry / calm	
mood	worried/tense/frustrated/troubled/...	
interest	6-point scale from +3 to -3, omitting 0.	
level 2	STYLE	(about the speech)
type	a speaking-style label (open-class)	
purpose	a speech-act label (closed-class)	
sincerity	insisting/telling/feeling/recalling/acting/...	
manner	polite/rude/casual/blunt/sloppy/childish/sexy/...	
mood	happy/sad/confident/diffident/soft/aggressive/...	
bias	friendly/warm/sarcastic/flattering/aloof/...	
level 3	VOICE	(about the sound)
energy	a 6-point scale from +3 to -3, omitting 0.	
softness	a 6-point scale from +3 to -3, omitting 0.	
brightness	a 6-point scale from +3 to -3, omitting 0.	
level 0	labeller confidence	
	marked on a 6-point scale from +3 to -3, omitting 0.	

Table 2. Six-level forced-choice tendency scales

	negative	positive
'very noticeable'	-3	3
'noticeable'	-2	2
'only slightly noticeable'	-1	1

Selection of descriptor labels is by means of a pull-down menu, offering a limited range of choices for each category of label. The speech is defined by the *combination* of these labels, in conjunction with a marked-up transcription of its text.

Statistical methods are currently being applied to learn the mappings between acoustic variation and the subjective category labels, using features automatically extracted from the speech signal [5].

3. Elements of SPEAKER STATE

The following categories are used to describe extra-linguistic aspects of the spoken message. They refer to the state of the speaker as perceived from the wider context of the spoken signal. They do not require knowledge of the speaker, nor of the context of the discourse, but a human annotator can infer much about speaker-listener relationships and the mental and physical state of the speaker from this level of information.

Purpose An open-class conversation-analysis label describing the pragmatic function of this section of the discourse. The

labeller is free to describe what the speaker is trying to achieve.
Confidence A description of the speaker's personal confidence as revealed in the discourse, marked on a 6-point scale.
Emotion This category is deliberately limited to the 4 emotions (happy / sad / angry / calm) that are offered by current emotion-enabled speech synthesisers.
Mood A label describing the speaker's mood (state of mind) using closed-class labels from a list including worried / tense / frustrated / troubled / etc. Used to complete the sentence: "This person sounds ...". (see also 'mood' below).
Interest An estimate of the speaker's involvement in the discourse, marked on a 6-point scale.

4. Elements of SPEAKING STYLE

The following categories are used to describe para-linguistic aspects of the spoken message. They refer to the style of the speech as perceived from the limited context of a single utterance. They do not require knowledge of the speaker, nor of the context of the utterance.

Type An open-class descriptor used to associate a bundle of speaking-style labels e.g., Angry1, Angry2, Greeting1, Bored3.
Purpose A closed-class speech-act label describing the illocutionary or pragmatic force of the utterance.
Sincerity A measure of the involvement of the speaker, i.e., the match between feeling and expression, indicated using a scale of insisting / telling / feeling / recalling / acting / reporting / citing/ etc., in the order of strong to weak involvement.
Manner A description of the attitude expressed by the speaker as evident from the speech sounds alone. This may be different from the attitude known to be held by the speaker according to knowledge from a wider context of information. Labels selected from a list including polite / rude / casual / blunt / sloppy / childish / sexy / etc.
Mood Used to complete the sentence: "This speech sounds ...". A description of the mood of the speaker as indicated by the sounds of the current utterance. This may be different from the mood of the speaker estimated from knowledge of the wider context. Labels selected from a list including happy / sad / confident / diffident / soft / aggressive / etc.
Bias An indication of the speaker-listener relationship as it can be distinguished from each utterance. Labels selected from a list including friendly / warm / jealous / sarcastic / flattering / aloof / etc.

5. Elements of VOICE QUALITY

The following categories are used to describe the perceived physical or acoustic aspects of the speech signal. They refer to the controlled qualities of the voice, and can be marked on segments smaller than a single utterance.

Softness An indication of the strain or effort perceived in the voice, measured on a 6-point scale.
Brightness An indication of the perceived brightness of the voice, measured on a 6-point scale.
Energy An indication of the perceived variability and strength of phonation in the utterance, measured on a 6-point scale.

Labeller confidence
This measure is provided in order to allow the labeller to indicate how confident they feel in the choice of labels for each smallest segment of speech. It is not directly related to voice quality, but is marked at the smallest unit size.

6. Tools & Software

We adapted the Transcriber software [6] for labelling the ESP data. The assumption underlying the Transcriber's DTD is that long sequences of audio data will typically contain several *episodes*, each consisting of *turns* by different *speakers* (or musical interludes), each containing one or more utterances.

Since our data are mainly of single-speaker recordings, we were forced to adopt a different underlying organisation for the labelling in order to utilise the richness of the Transcriber's data structure. We adapted the long-term fields (episodes) to label speaker-state, information, and the short-term (turns) for speaking-style. Voice-quality is annotated inline with the text.

This will be described in more detail in the oral presentation of this paper, and our proposed DTD with software patches can be obtained on request from nick@atr.co.jp.

7. Conclusion

Supra-linguistic cues in the speech reveal how the speaker relates to the listener, and to the content of each utterance. So, rather than labelling 'emotion', we prefer to annotate our speech data to indicate these ‘speaker-relationships’, which we consider to be the significant dimensions for a model of speaking style for the next generation of speech synthesis.

Two dimensions of relationship are presently considered necessary; “commitment”, and “friendliness”. The first (i.e., a content-relationship) governs the expressed or revealed sincerity of the speaker, including the expression of emotion, and revelation of attitudinal bias. This dimension distinguishes the social roles that the speaker might be assuming from signs revealing the speaker's inherent attitudinal and emotional states. The second (i.e., a listener-relationship) governs the formality and the degree of familiarity that can be expressed in the speech. The precise details are culture-specific, and depend on inherent rank, age, sex, and familiarity differences, etc., but the speaker can manipulate this dimension freely within pre-determined limits on a case-by-case or day-to-day basis.

Acknowledgements
This work is supported partly by a grant from the Japan Science & Technology Corporation and partly by aid from the Telecommunications Advancement Organisation of Japan.

References

[1] Campbell, W. N., “Databases of Emotional Speech”, in Proc ISCA ITRW on ‘Speech and Emotion’, pp. 34-38, 2000.
[2] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp
[3] Campbell, W. N., “The Recording of Emotional speech; JST/CREST database research”, in Proc LREC 2002.
[4] Campbell, W. N., “MD vs DAT”, in Proc ASJ, Spring 2002.
[5] Mokhtari, P, & Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.", in Proc LREC 2002.
[6] Transcriber: this free software can be downloaded from www.etca.fr/gip/Projects/Transcriber.