A Language-Resources Approach to Emotion: Corpora for the Analysis of Expressive Speech

Nick Campbell
Acoustics & Speech Processing Department,
Spoken Language Communication Research Laboratory,
Advanced Telecommunications Research Institute International,
Keihanna Science City, Kyoto 619-0288, Japan.
nick@atr.jp

This paper presents a summary of some expressive speech data collected over a period of several years and suggests that its variation is not best described by the term "emotion". Further, that the term may be misleading when used as a descriptor for the creation of expressive speech corpora. The paper proposes that we might benefit from first considering what other dimensions of speech variation might be of more relevance for developing technologies related to the processing of normal everyday spoken interactions.

1. Introduction

Spoken language has been extensively studied through the use of corpora for several decades now, and the differences between the types of information that can be conveyed through written texts and those that are signalled through speech are beginning to be well understood.

The paralinguistic information which is perhaps unique to speech communication, is largely carried through modulations of prosody, tone-of-voice, and speaking style, which enable the speakers to signal their feelings, intentions, and attitudes to the listener, in parallel with the linguistic content of the speech, in order to facilitate mutual understanding and to manage the dynamics of the discourse [1].

The different types of information that are signalled by different speaking styles are also well understood and are beginning to be modelled in speech technology applications. The more formal the speech, the more constrained the types of paralinguistic information that are conveyed.

As an example of one extreme, we might consider a public lecture, where the speaker is (sometimes literally) talking from a script, to a large number of listeners (or even to a recording device with no listeners physically present) and has minimal feedback from, or two-way interaction with, the audience. This type of 'spontaneous' speech is perhaps the most constrained, and most resembles text.

As an example of the other extreme, we might consider the mumblings of young lovers. Their conversation is largely phatic, and the words might carry little of linguistic content but are instead rich in feelings. For them, talk is almost a form of physical contact.

There are many steps along the continuum between these two hypothetical extremes of speaking-style variation. Perhaps they can be distinguished by the ratio of paralinguistic to linguistic content, i.e., the amount of 'personal' information that is included in the speech. The lecture, having almost no personal information and a very high amount of propositional content will result in a very low value of this measure, while the phatic mutterings will score very high.

If we are to collect data that contains sufficient examples of natural spoken interactions along the whole range of this continuum of values, then low-scoring material will prove very easy to collect, but most lovers might object strongly to the suggestion of a recording device intruding into their privacy. Thus, by far the majority of speech corpora that have been used in previous research score very poorly on this scale and as a result the speech that they contain is not very far removed from pure text in its style and content.

2. A Corpus of Expressive Speech

We need more varied and representative corpora if we are to develop future speech technology that is capable of processing the more human aspects of interactive speech in addition to its propositional content. However, the difficulties of doing this are well known. Since Labov, the presence of an observer (human or device) has been known to have an effect on the speech and speaking style of the recorded subject, and unobtrusive recording is unethical, if not already illegal in most countries. Several approaches have been proposed to overcome this obstacle to future research. This section reports one of them, and discusses some of the conclusions that we reached on the basis of that experience.

The JST/CREST Expressive Speech Corpus [2] was collected over a period of five years, by fitting a small number of volunteers with head-mounted high-quality microphones and small minidisc walkman recorders to be worn while going about their ordinary daily social interactions. Further groups of paid volunteers transcribed and annotated the speech data for a variety of characteristics, including speech-act, speaker-state, emotion, relationship to the interlocutor, etc. All the data were transcribed, and about 10% was further annotated. Figure 1 shows a sample of the annotation results, and Table 1 shows some of the categories that were used for annotation. These samples can be listened to at the project web-site, http://feast.atr.jp/non-verbal/. The material is in Japanese, but many of the findings hold for other languages as well. Japanese are people too, and many of the non-verbal speech sounds in this lan-

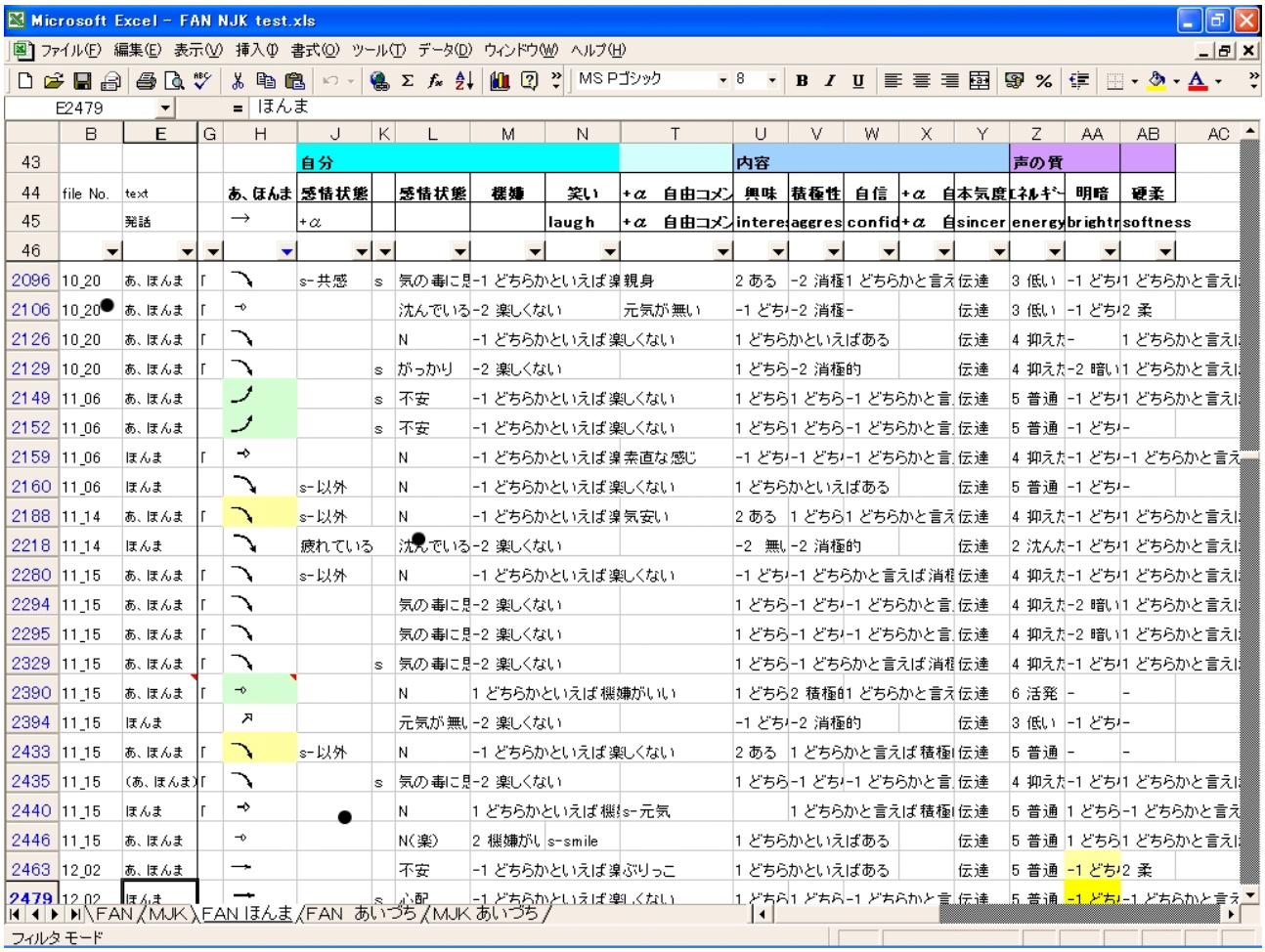

Figure 1: A screenshot of the labelling spreadsheet for the word “honma”. The columns include data described in more detail in Table 1. In this form of labelling, tokens are listened to in isolation, free of contextual influence, while in other forms of labelling they are annotated in time-aligned sequence, taking context into account. By clicking on a filename, the labeller can listen to each sample interactively

guage can be equivalently understood by native-speakers of other languages who have no experience of either the Japanese language or culture. A laugh is a laugh in any language. So is a sigh.

The data in figure 1 represent a few of the approximately 3,500 tokens of the Japanese word /honma/ from one speaker of the corpus. The word functions in much the same way as “really” does in English; both as a qualifying adjective (really hot!) and as an expressive exclamation (really?!). The word is typical of many that are repeated frequently throughout the corpus, and that are used by the speaker more for their discourse effect than for their linguistic or propositional content. No two pronunciations of this word are the same, and each carries subtle affective and interpersonal information that signals many kinds of different states and relationships, as will be described in more detail below.

These words proved most difficult for the labellers to adequately categorise. They function primarily as backchannel utterances, but also serve to display a wide range of attitudinal and affective states. We have reported elsewhere [3] studies that measure the extent to which their function can be similarly perceived by different groups of listeners belonging to different cultural backgrounds and languages. In terms of quantity, more than half of the utterances in the corpus were of this type; short words or simple syllables that occurred alone or were repeated several times in succession, often not appearing at all in a dictionary of the formal language, but forming essential components of a two-way spoken interaction.

3. Annotating the Corpus for Emotion

It is clear that these types of expression carry emotional information. They are very expressive, and revealing of the speaker’s type(s) and degree(s) of arousal. We therefore attempted to label emotion in the corpus data.

A version of the Feeltrace software was implemented (square, rather than round!) and each utterance was assigned a value within the valency/arousal space thus defined. The labellers understood the meaning and validity of these two dimensions, and felt easy about working with the mouse-based software for data entry, but most complained about the work after a short time. They claimed that the framework simply wasn’t appropriate for describ-

ing the different types of variation that they perceived in the speech. They proposed instead the descriptive categories shown in Table 1.

While the speaker was clearly in a given state of emotional arousal during each utterance, the correspondence between what the labellers could determine about the speaker state, from various contextual and expressive clues, and how the speaker's utterance was *performing* in terms of her stance within the discourse, was often very small.

When labelling five-years worth of someone's speech, you become very familiar with that person's mannerisms and even those of their circle of acquaintances. For example, it might be clear from various such clues that the speaker is angry on a given day. Yet the presence or absence of anger in a person may have little or no relationship to the presence or absence of anger in the expression of a given speech utterance. How is this to be labelled in the simple valence/arousal framework?

Specifically, let's examine three such cases: (i) A schoolteacher walks into the classroom and the children continue to be noisy. The teacher gets angry with the children. (ii) The same teacher has been wrongly accused of malpractice during the lunchbreak and continues to teach in the afternoon. She explains to the children the details of the lesson. (iii) The same teacher later in the afternoon as the children persist in being noisy. She gets angry with them again.

In the first case, the speaker expresses anger but does not feel it - she is merely doing her job, and performing an expected role in order to achieve a predictable effect. The children know the rules and soon stop talking. In the second case, the opposite is happening; the person is angry, but her speech is not; as a professional, she continues to speak to the children in the way to which they have become accustomed. In the third case, we have an angry person who is being angry. The effect on the children is immediate. They are afraid.

The three types of speech illustrated above all contain anger, but they differ in whether it is felt or expressed. We could further differentiate by degree of anger, or degree of expression, or both, and with respect to degree of expression, also determine whether "something inside is being let out" or whether the voice is being made to sound as though it is, when in fact inside the feelings may be neutral (whatever that expression might mean).

3.1. Affect and Attitude in the Speech

In view of the above, the labellers felt that it was preferable to work with a three-level labelling system, where (i) facts about the speaker could be distinguished from (ii) facts about the speech, and (iii) separate independent evaluations could be made about the information portrayed by the voice. After some experimentation, the system detailed in Table 1 was proposed.

Level 1 describes the state of the speaker, requiring long-term context, and an estimation of the discourse purpose of the utterance (see details below), the speaker's emotion and mood (these labels are free-input, those in the table being examples), her interest in the discourse, and finally a label to denote labeller-confidence. Numerical labels are forced-choice on a scale of high to low (see lower part of table) with no default or zero settting.

Table 1: Three levels of labelling for describing each utterance, including use of six-level forced-choice tendency scales

Level 1	STATE (about the speaker)
purpose	a discourse-act/DA label (see text)
emotion	happy/sad/angry/calm
mood	worried/tense/frustrated/troubled/...
interest	a 6-point scale from +3 to -3, omitting 0
confidence	a 6-point scale from +3 to -3, omitting 0
Level 2	STYLE (about the speech)
type	speaking-style label (open-class)
purpose	a discourse-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/jealous/sarcastic/flattering/aloof/...
Level 3	VOICE (about the sound)
energy	a 6-point scale from +3 to -3, omitting 0
tension	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	a 6-point scale from +3 to -3, omitting 0

6-point values:	negative	positive
'very noticeable'	-3	3
'noticeable'	-2	2
'only slightly noticeable'	-1	1

Level 2 describes the style of the speech, its type and purpose, and can be estimated from a short-time window (i.e., no context) so that it describes the information available from listening to the isolated speech utterance alone, as distinct from the same utterance situated in a discourse (i.e., we don't care if she is angry or not, but this segment SOUNDS angry). The *sincerity* label describes an important functional aspect of the speech, such as can be distinguished between the verbs 'insisting', 'telling', 'quoting' 'saying', 'feeling', 'recalling', 'acting', 'pretending' etc.

An example from the corpus will illustrate how difficult it can be to assign such apparently simple labels. The speaker, a young woman, says something in Japanese that might translate into: "You're a f***ing pig! I shouted and stormed out of the place!". It was told by the young woman to a sympathetic friend who was laughing with her over the row she and her husband had had the previous evening. On listening to the first few words in isolation, the listener can hear only extreme anger. However, there is no gap in the speech and by the time we reach "stormed out", the speaker is giggling as she speaks, and then finally the utterance ends in real guffaw laughter.

In the example above, we would select 'quoting' (self) rather than 'acting' or 'feeling' for the expletive, and 'feeling' for the laughter at the end, but still have no way to ex-

plain the slide of "emotions" (is that the right word?) from start to end of the utterance, which lasted little more than a second. Fortunately, not all utterances are as complex, and most were satisfactorily assigned a single label for each category in the table.

Manner is a bucket category that includes politeness and sexiness (which are not at all mutually contradictory) as well as childishness, sloppiness, etc to describe the perceived attitude(s) of the speaker towards the listener. This is complemented by Mood and Bias, of which the former indicates the affective states of the speaker, and the latter his or her attitudes.

Level 3 describes the physical aspects of speaker's voice quality and speaking style in perceptual terms.

3.2. Discourse-Act Labelling

In order to describe the purpose or function of each utterance, a decision was first made about its *directionality*, which may be either 'offering' (to the listener) or 'seeking' (from the listener). Utterances were then broadly categorised into seven classes of *discourse intentions*, including Questions, Opinions, Objections, Advice, Information, Greetings, and Grunts. These category labels were determined by necessity as examples of each appeared in the data; the last category accounted for almost half of the utterances in the corpus.

Under the category of *Questions*, we use the following labels: WH Questions, Y/N Questions, Repetition Requests, and Information Requests.

Under the category of *Opinions* we use the following labels: Opinion, Compliment, Desire, Will, Thanks, and Apology.

Under the category of *Objections* we use the following labels: Objection, Complaint.

Under the category of *Advice* we use the following labels: Advice, Command, Suggestion, Offer, and Inducement

Under the category of *Information* we use the following labels: Give Information, Reading, Introduce Self, Introduce Topic, and Closing

Under the category of *Greetings* we use the following labels: Greeting, Talking to Self, Asking Self, Checking Self.

Under the category of *Grunts* we use the following labels: Notice, Laugh, Filler, Disfluency, Mimic, Habit, Response, and Backchannel. Response and backchannel utterances are further subcategorised into the following types: agree, understand, convinced, accept, interested, not convinced, uncertain, negative, repeat, self-convinced, notice, thinking, unexpected, surprise, doubt, impressed, sympathy, compassion, exclamation, listening, and other.

4. Expressive Speech and Emotion

The experience gained from this labelling process has caused us to now rethink our original assumptions. We started off by trying to overcome Labov's Observer's Paradox, hoping that long-term exposure to a recording device would eventually cause the wearer to familiarise with it to the extent that it no longer becomes a hindrance to normal spoken interaction, even of a highly personal kind. This has proven to be the case, as the variety of speech that we have collected well shows.

However, another paradox has arisen in its place. We originally believed that we would be able to capture truly natural and spontaneous emotional speech data by having a microphone active and in place before and while the emotional 'event' took place. Instead, we find that by far the majority of our speech material is NOT marked for emotion as we then conceived it, but that it varies significantly in dimensions better related to affect and attitude, signalling the mood and interest of the speaker, his or her current relations with the listener, and controlling the variable flow of the discourse.

We started out by believing that 'emotion' was the essential component lacking in our speech corpora for technology development, but we now consider that the 'human-dimension' that we were looking for is not best described by the term "emotion" at all. Our data score very highly on the measure of paralinguistic to linguistic content described in the introduction, and are very far from the formal speech of less interactive situations, almost half being non-verbal and affect-related, but they lead us to conclude that the emotional state(s) of the speaker are not always directly expressed, and that social and interpersonal considerations override the supposed link between subjective emotion and displayed affective states. The social aspects of communication take precedence over the blunt expression of feeling, and while the latter can perhaps be determined from an expressive utterance, the multiple levels of information in the former provide a richer source of data to be processed if we are to "better understand the person" through her speech.

5. Conclusion

Since it is of great importance to present experiments with real examples and to have theoretical discussions based on analysis of representative data, it is of fundamental importance to clarify emotional representation, data collection aim and methodology to obtain data. Many corpora of speech are now being designed to maximise the inclusion of emotional samples, so that progress may be made in the understanding of all aspects of human interactions, but because of the difficulty in collecting natural spontaneous materials, actors are being used to simulate the target speaking styles and emotional states. They are undoubtedly very competent and will produce exactly the material that we ask for, but in trying to please us, are they giving us what we really need? In constraining our requests to "emotion" are we not in danger of missing so much more that is perhaps the core of human interpersonal interactions? Our experience with the ESP corpus leads to the conclusion that this might be the case.

References

[1] Campbell,, N., "Getting to the heart of the matter; Speech as Expression of Affect rather than Just Text or Language", pp 109-118, Language Resources & Evaluation Vol 39, No 1, Springer, 2005.

[2] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.atr.jp

[3] Campbell, N., & Erickson, D., "What do people hear? A study of the perception of non-verbal affective information in conversational speech", pp. 9-28 in Journal of the Phonetic Society of Japan, V7,N4, 2004.