DATABASES OF EMOTIONAL SPEECH

Nick Campbell

ATR Information Sciences Division, Kyoto 619-02, Japan

CREST, JST (Japan Science and Technology)

http://www.isd.atr.co.jp/esp / nick@isd.atr.co.jp

Abstract

This paper presents a personal view of some the problems facing speech technologists in the study of emotional speech. It describes some databases that are currently being used, and points out that the majority of them use actors to reproduce the emotions, thereby possibly falsely representing the true characteristics of emotion in speech. Databases of real emotional speech, on the other hand, present serious ethical and moral problems, since the nature of their contents must, by definition, reveal personal and intimate details about the speakers.

1. OBJECTIVES OF THE PAPER

This paper does not set out to provide an inventory of databases available for the study of emotional speech characteristics; the JASA paper by Murray and Arnott [1], the PHYSTA home page [2], and the web pages of Erlangen University and the Salk Institute [3], for example, provide good overviews of such previous work. Instead, the paper presents a personal account of some of the problems facing researchers who wish to study the speech characteristics associated with different emotions. It approaches the issue from the standpoint of speech technology, rather than that of psychology, and describes work planned under a forthcoming JST-funded five-year project for the study of 'expressive speech phenomena'□which will include the production of a large-scale emotional-speech database. Rather than present new facts or data, the paper sets out some topics for discussion and raises questions; in the hope that some of the issues may be resolved during the three days of the workshop.

2. EXPRESSIVE SPEECH

Linguistic information is all that can be carried by text, but it is only a small part of the spoken message. As humans, when listening to speech, we are sensitive to extra-linguistic information about the identity and the state of the speaker, as well as to paralinguistic information about the speaker's intentions underlying the utterance. This information is largely missing from computer speech synthesis, and current speech recognition systems make no use of it.

In many instances of conversational human communication, the speaker's intention, signalled by the manner of speech, is as important as the text of the utterance, and in social or phatic communication, often more so. As humans, we have become used to processing such extra-verbal information and will presumably expect it when interacting with machines through the medium of voice.

2.1. Acting

Previous studies of emotional speech characteristics that are frequently cited (e.g., [4,5]), particularly those conducted for speech synthesis, have often been based on recordings of actors simulating various emotions under studio conditions. To the extent that the intended emotions can usually be correctly identified by listeners afterwards, these recordings must be considered satisfactory. By using different readings of otherwise identical sentences they allow principled analysis and comparison of the changing acoustic features. However, being acted, it is questionable whether they faithfully represent the characteristics of speech used by ordinary people when they naturally experience similar emotions.

The difference between perceived emotion and expressed emotion is ideally minimised under these artificial data-collection circumstances, though in everyday social interaction, it may often be culturally preferred to *suppress* the overt expression of personal feelings or emotions. On the contrary, it is often preferable in social interaction to express emotions which may *not* be felt. We should therefore take care to distinguish between intended expression of emotion and unintended revelation of speaker-state.

The listeners to recordings of such studio-speech may be able to 'correctly recognise' the *intended* emotion from differences in speaking style etc., but they may also be aware at the same time that the speaker is *consciously* intending to express such an emotion, even to the extent that, being acted, the emotion is not 'felt' or sincere.

If a computer speech synthesiser were to emulate such speaking-style characteristics successfully, then it may be liable to misrepresent the intentions of its user. For example, if a disabled person for whom speech synthesis is the sole means of

verbal expression, was to use the synthesised voice to express genuinely-felt pleasure (or anger), then the listener might be able to 'hear' that the voice was only expressing acted pleasure (or anger), and is liable to (*mis*-)respond accordingly.

We therefore face a dilemma; there is a need to balance data for controlled scientific analysis and experiments, but the very act of balancing reduces the validity of the data. Speech collected in laboratory environments is likely only to be representative of laboratory speech; and yet speech cannot easily be collected in a methodical way outside of a controlled recording environment. This was not particularly a problem when the need was only to study the linguistic aspects of the speech, but the focus of research is now turning to paralinguistic information, and particularly to the boundary between paralinguistic and extra-linguistic information in speech, where *sincerity* is of greater importance.

It may not be difficult in practice to make the speaker angry when recording a corpus of speech -- indeed, frustration and anger are very common in our everyday interaction with machines -- but is it morally acceptable? The legal and ethical issues of deliberately inducing an emotion for the purposes of collecting scientific data are not clear. How can we provoke people to express controlled utterances sincerely, in order to obtain reliable and representative data for analysis?

2.2. Stimulation

Rather than simulate emotions in order to produce speech data, it may be better to *stimulate* the emotions, and then record the resulting changes in speaking manner and style. The inevitable loss in control over utterance content must therefore be made up for by an increase in the number of utterances, so that statistical procedures can be applied in order to produce reliable generalisations from the more spontaneous data.

In an attempt to overcome the problem of acted speech in previous work with Iida [6], we used readings of emotionally biased texts to stimulate the feeling of a particular emotion in the speaker. We limited our preliminary study to three emotions likely to be needed in a communication aid for the verbally impaired (sadness, anger, and joy), and used texts composed and collated by the speaker and read by her as the basis of our work.

It is unfortunate that a female scientist in Japan can face considerable bias and may have to work much harder than male colleagues in a similar situation. By selecting a story about sexual discrimination, I believe the reader was able to tap into her own experience and arouse feelings that genuinely correlated with the expressions of emotion in the text. The anger that is evident in her voice in this reading is subjectively different from that which she produces when reading pairs of balanced sentences. Similarly, the sadness that was generated when reading about the experiences of a young boy with hearing difficulties on his first day at school is not acted, but is experienced by the speaker during the reading.

Subsequent perceptual tests confirmed that listeners were able to correctly identify the intended emotions from the speech, but there is a potential contamination in the data for such tests, since many of the sentences themselves contain verbal cues to the underlying emotional state of the speaker. We overcame this by using concatenative speech synthesis to generate a further set of test utterances which were semantically neutral in content. Segments from each database were re-sequenced without signal processing, using CHATR [7,8], to generate new, semantically neutral utterances, maintaining the prosody and voice-quality characteristics of each original database. Listeners were able to identify the underlying emotion at levels reliably above chance, even when the textual cues were thus removed.

2.3. Elicitation

If the need for a balanced set of utterances can be eliminated or reduced, then spontaneous expression of emotion in speech can be elicited under controlled recording conditions. The cost of losing the rigorous balance in the design of the utterances can be compensated by the use of a much larger corpus which will contain similar, albeit shorter, acoustic sub-sequences that can then be analysed by statistical means.

Generalisations can be made from a larger corpus which make up for the lack of 'balanced pairs' of utterances, but which also require more complex factorial analysis to reveal. The tools for such analysis are now well developed, easy to use, and are freely available in the public domain. A separate invited paper in the same proceedings is devoted to this topic. A problem remains, though, with respect to the data collection.

Several techniques exist for the elicitation of attitudinally biased responses in a dialogue or conversational setting if one side of the conversation is privy to the needs of the experimenter. Talk-show interviewers and socio-linguists are particularly familiar with such techniques. However, there is less known about how to elicit or provoke a sincerely felt emotion, such as fear for example, using similar procedures.

In the implicitly safe and ordered (typically laboratory) environment of the data collection, such strongly-felt emotions are considered 'out-of-place', and the act of deliberately inducing fear (though not joy) is ethically questionable, if not illegal in many countries [9]. Co-operation is required from the speaker when recording, but this may not fit well with, for example, a strongly felt feeling of anger or joy, when the natural response might be to leave the controlling environment and cease co-operation altogether.

An interesting example of such data collection has been provided by the SUSAS group, whose research into "Speech Under Simulated and Actual Stress" [10,11] uses recordings taken on location: speech uttered during an amusement park roller-coaster ride and in helicopter cockpits provides authentic samples of spontaneous fear and anxiety. However, the background noise and physical exertion in some of their recording situations make the speech unrepresentative of many everyday situations. The SUSAS recordings of psychiatric patient interviews provide examples of depression, fear, anxiety, and anger, but cannot be freely distributed.

There is therefore, still considerable need for work to be done on testing ways of eliciting speech, with varying emotions and attitudes, from speakers in laboratory or quiet-room conditions so that analysis of the acoustic characteristics can be performed.

2.4. Found speech

'Found speech', or previously existing speech data which were not produced explicitly for the purpose of scientific experiment, can provide a rich source of emotional material. The famous radio news broadcast of the crash of the Hindenburg is an obvious example of such material, in which the excitement and panic caused by an external event are clear in the voice and of which high-quality recordings have been preserved. The media may not be able to use all the material recorded from witnesses and participants in such emotionally-disturbing events for the final edited broadcasts, but often these recordings are archived and can be accessed for research use. Recordings from talk-shows, phone-in programmes and even opera [12] can provide useful material for analysis and comparison. Unfortunately the copyright and other legal issues regarding the public use of such material limit the freedom of the researcher to make full use of these data, and discourage the owners from making the materials available.

If the identity of the speaker could be concealed, then it is likely that found data could become more readily accessible. Voice distortion devices are frequently used to protect the anonymity of informants on live radio interviews or television broadcasts, yet they preserve much of the original intonation pattern and speaking style information. Speech-segment re-sequencing synthesis, on the other hand, preserves voice-quality information while removing contextual clues. It is possible that an interface which combines these technologies might allow protected access to sensitive speech data, enabling research on both prosodic and voice-quality characteristics, but not simultaneously.

An alternative approach would be to find ways to make such data available for the limited purpose of statistical analysis alone, without revealing the actual sounds themselves to the researchers performing the statistical analysis of the content. However, the problem with such an approach is that a trusted set of listeners would be required in order to verify that the speech data contains samples of a particular emotion type, and this in itself would require a form of public release of the contents; which is what we were trying to find a way to avoid.

3. DATABASES OF SPEECH

There has to date been little co-ordination between the various laboratories with respect to the collection and dissemination of emotional speech corpora, and it is likely that as well as work being duplicated, the same mistakes are also being repeated unnecessarily. Admittedly, each research institute has its own specific reasons for data collection, and the effort of sharing resources may be seen as an unnecessary complication, but other fields of speech research have shown the benefits of such co-ordinated work.

COCOSDA is the International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques. It was originally formed ten years ago to promote collaborative work and information exchange concerning speech input/output, to help develop resources and standards in spoken-language engineering. At that time, most of the effort was devoted to gathering corpora for speech recognition and for synthesis research, but now the focus is broadening as the original goals have been satisfied.

COCOSDA organises meetings once a year, where rapporteurs from the world's main geographical areas meet to announce progress that has taken place in the various regions under the heading of specific database-related technical topic domains. The current areas include Africa, Asia, Europe, Latin America, North America, and Oceania, and the technical topic domains include speech synthesis, speech recognition, corpus annotation tools, spoken dialogue, and the specific needs of local languages. A proposal has been made that the next topic area to include should concern corpora of emotional speech.

There is a good argument to be made for the introduction of such a topic area, given the difficulties involved in the collection of such data, and of the legal and ethical issues concerning its distribution and use. However, unless there is a consensus for concerted action between the laboratories that are actually producing the data, little can be expected to come of it. Perhaps this topic might be discussed in more detail during the Belfast Workshop. Further developments will be reported in the COCOSDA web pages at www.slt.atr.co.jp/cocosda, where comments and suggestions will be welcomed.

4. FUTURE WORK

This section reports on a new project funded by the Japanese Science and Technology Agency under the auspices of *'Information Technology in Life for an Advanced Media Society'*. The five-year 'CREST' type-II project will investigate speech-processing techniques for the understanding and generation of para-linguistic and extra-linguistic information as a communication aid and for spoken-language interfaces. It aims specifically to develop a model of the way that functional (speech-act) differences can be assigned to spoken words and phrases through variations in speaking style.

The *Expressive Speech Processing (ESP) Project* will examine ways in which speakers express meaning, intention, and feelings, over-and-above those encoded in the wording of an utterance. It will start by creating databases of expressive speech to determine the range of variability in the vocalisation of attitudes and emotions in three languages (Japanese, Chinese, and English).

A fundamental part of the research will be to determine the acoustic and prosodic variables that indicate different attitudes and emotions in speech, and to map these to linguistic objects and frameworks, so that a model of the para-linguistic and extra-linguistic information can be obtained with respect to the structural and semantic content of an utterance.

Speech technology applications, both active and passive, will be developed on the basis of these speech corpora in order to improve the ease of interaction between people and machines, and to provide the foundations for a speech-based interface that is sensitive to the state of the user for an advanced information and media society.

The research will consist of determining the optimal features that can be extracted reliably from a speech waveform, and the mapping of these onto known linguistic structures and identifiable speaking-style characteristics, so that the desired intention of the speaker can be expressed or interpreted. The mapping must be two-way, with implementations both in speech recognition and in speech synthesis.

4.1 Data collection and labelling

The project will initially require data derived from both acted speech and natural speech, but by using voice-interactive devices in the data collection, it will gradually build a bank of actual *in situ* responses. It will therefore consist in large part of an investigation into the ways of collecting natural speech data which is spontaneously expressive of varied speaker intentions, without any recourse to acting

Initial experiments (performed under separate ATR-ICP, ATR-NAIST, and ATR-Keio collaborations) have shown that both the recognition of attitude by machine processing and the control of emotion in synthetic speech can to a certain extent be achieved by signal analysis implemented as automatic labelling [13]. Thus the efficient labelling of speaking-style and voice-quality characteristics on a speech signal will form the main part of the CREST ESP research.

Phonemic labelling (or segmental alignment) is already a well-established technology with public-domain software available [14], but it requires additional research in order to improve the detection of mis-labelled speech segments.

Prosodic labelling is currently being actively researched in several laboratories throughout the world, and will soon reach a usable level of maturity [15].

Phonation style analysis has often been performed analytically, with manual intervention, but has not yet been automated; so this element of the research still requires considerable development of data and tools.

By combining these three levels of speech labelling (or annotation) in conjunction with a semantic and syntactic analysis of the corpus text, we anticipate being able model the mapping between acoustic events and different levels of intended meaning.

4.2 Stages of the Research

The three main phases of the research will involve (a) development of tools and data (b) modelling, prediction and evaluation, and (c) application prototyping. These will in turn require collection and labelling of the initial database, training and evaluation of the initial models, preliminary prototyping and lab trials, refinement of segmentation and analysis tools, refinement of models and grammar, revision of database design, second (final) data collection, model retraining and evaluation, application development and field trials.

Key elements of the research include: the use of natural data (eliciting free expression of attitude and emotion without acting), statistical mapping of features to text and to speaker-characteristics, development of signal processing tools for feature extraction and waveform labelling, evaluation of human perceptual criteria and consequent confusion of responses, prototyping and evaluation of a speech synthesis interface using CHATR, prototyping of a speech-understanding interface from combined text and speech input, and implementation of the above in software for commercial application

The procedures willl involve mapping from linguistic and speaker-specific information to intonation parameters for synthesis, with a reverse mapping for identification of speaker-intent. Algorithms will be developed for the implementation of prosodic patterns in synthesis, and for the detection and labelling of prosodic and voice-quality events in the incoming speech.

An intermediate goal is to provide tools for the automatic annotation of speech corpora, and for the construction of emotional speech databases.

5. SUMMARY

If people are to interact with machines by means of voice, for example to buy products or to retrieve information, or if machines are to speak on behalf of people, as communication aids, then it is important that the inevitable misunderstandings should be minimised. Research into the perception of emotion in speech has focused on portrayals of emotions by actors, but there is no guarantee that the speech of an actor carries the same information as speech uttered in a spontaneously emotional way. Because the expression of emotion and feeling is such a characteristic feature of human speech, it is important that machines be developed which can process the affective content as well as the textual content of an utterance. We should therefore be especially careful about the data we collect for the training of such devices.

6. REFERENCES

1. Murray, I. R. and Arnott, J. L. Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *Journal of Acoustic Society of America* 93, No. 2, 1097-1108, 1993.
2. TMR PHYSTA: www.image.ece.ntua.gr/physta
3. http://www.forwiss.uni-erlangen.de/msnutt/emotion and http://www.emotion.salk.edu
4. Higuchi, N., Hirai T., and Sagisaka, Y. (1997). "Effect of speaking style on parameters of fundamental frequency contour". In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg (Eds.) *Progress in speech synthesis*, Springer-Verslag, New-York, 417-428

5. Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998). "Transcription of Prosodic and Paralinguistic Features of Emotional Speech". *Journal of the International Phonetic Association*, 28, 83-94.

6. Iida, A., Campbell, N. and Yasumura, M. Design and Evaluation of Synthesised Speech with Emotion. *Journal of Information Processing Society of Japan* Vol. 40, No. 2, 479-486, 1998

7. Campbell, W. N. and Black, A. W. CHATR a multi-lingual speech re-sequencing synthesis system. *Technical Report of IEICE* SP96-7, 45-52, 1996.

8. Campbell, W. N. Processing a Speech Corpus for CHATR Synthesis. *Proceedings of The International Conference on Speech Processing* 183-186, 1997.

9. Shaver, P., Schwartz, J., Kirson, D. and O'Connor, C. Emotion Knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* Vol. 52, No. 1, 1061-1086, 1987.

10. Hansen J.H.L., and Bou-Ghazale, S. "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database", EUROSPEECH-97, Vol.4, pp. 1743-1746, Rhodes, Greece, September 1997

11. http://cslr.colorado.edu/rspl/stress/susas.html

12. Siegwart, H.., and Scherer, K. R. (1995). "Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi," *Journal of Voice 9 (3)*, 249-260.

13. Greasley, P., Sherrard, C., Waterman, M., Setter, J., Roach, P., Arnfield, S., and Horton, D., "The Perception of Emotion in Speech", in *Proc XXVI International Congress of Psychology,* Montreal, 1996.

14. www.mbrola.org: free public-domain phonemic-alignment software for speech labelling.

15. Sprosig: the special-interest-group for speech prosody: www.isca-speech.org/sprosig

16. www.isd.atr.co.jp/esp Home-page of the CREST (JST) Expressive Speech Processing Project.