Emotional speech synthesis for emotionally-rich virtual worlds

Marc Schröder
DFKI GmbH, Saarbrücken, Germany
schroed@dfki.de

Abstract

This paper aims to give a brief overview of the current state of the art in emotional speech synthesis in view of a multi-modal context. After a brief introduction into the concept of text-to-speech synthesis, two approaches to the expression of emotions in speech synthesis are described. The categorical approach models emotions as discrete categories and is able to provide high-quality emotional speech for a few emotion categories; the dimensional approach uses emotion dimensions such as activation and evaluation to model essential emotional properties, leading to more flexible but less specific expressions. Architectural requirements for an audio-visual integration are outlined. Three examples of demonstrators illustrate the types of applications we currently envisage. Finally, the question of validation of a generation system is formulated, and a direction for the development of possible answers is suggested.

1 Introduction

In human-to-human communication, emotions are expressed through a number of channels including facial and bodily expression, behaviour, and speech. Emotion synthesis for virtual worlds will ultimately require an adequate modelling of all these channels. This paper aims to give a brief introduction of current work in the domain of emotional speech synthesis, along with some considerations regarding the embedding of a speech synthesis system in a multi-modal context.

It is assumed here that the emotional state to be expressed is determined by a different processing step, be it a human user, a dialogue script or an affective reasoning engine as part of a planning component.

2 Speech synthesis

A text-to-speech synthesis system [1, 2] consists of two main components: a text analysis part and an audio generation part. The text analysis converts plain text or a speech synthesis markup language document [3] into a phonetic transcription accompanied by prosodic parameters for the specification of intonation, pausing, and speech timing. Natural language processing techniques are used for this task. The audio generation component uses the resulting parametric representation to generate audio data. Signal processing techniques, often in combination with recorded human speech databases, are used by this component.

3 Emotional speech synthesis

In the following, two current approaches to emotional speech synthesis are presented: A categorical approach, modelling emotions as distinct categories, and a dimensional approach, modelling essential emotional properties in terms of emotion dimensions. For a more detailed description of these and other existing approaches, see Schröder [4].

3.1 Categorical approach

In this approach, the emotions are treated as distinct categories, be it basic emotion categories such as anger, joy, fear, sadness, etc. [5] or domain-specific expressive categories. The best results are achieved using unit-selection techniques, which require the recording of an entire speech corpus for each expressive category to express. The amount of work involved naturally limits the number of distinct categories which can be recorded (and thus synthesised): As an example, Iida et al. [6] recorded a full Japanese speech corpus for the three emotions anger, joy, and sadness. Johnson et al. [7] recorded limited domain corpora, with which only certain domain-specific types of sentences can be spoken, for four expressive categories appropriate for the military training domain.

Common to these approaches is the fact that they do not need to, and indeed cannot, explicitly model the acoustic properties of emotions. Instead, the unit selection technique used preserves the expressive speech as it was recorded. Consequently, the resulting speech is natural-sounding for the specific categories which were recorded, but limited to exactly these.

3.2 Dimensional approach

An alternative approach starts with a different representation of emotional states, as emotion dimensions rather than categories. Emotion dimensions [8, 9, 10] are a simplified representation of the essential properties of emotions. Evaluation (positive/negative) and Activation (active/passive) are the most important dimensions, sometimes complemented by Power (dominant/submissive).

The dimensional approach to emotional speech synthesis [11] uses rules to map any point in this two- or three-dimensional emotion space onto its acoustic correlates. This requires an explicit model of the acoustic correlates of emotional states, which can be obtained e.g. by means of a database analysis [12]. The resulting synthesis system is by design highly flexible, thus allowing the gradual build-up of emotions, the synthesis of non-extreme emotions, and the change of emotional tone over time. However, as a simplified representation of the emotion is used, the emotion cannot be expected to be fully specified through the voice. Therefore, complementary sources of information, such as verbal content, the visual channel, and the situational context, will be required for a user to identify the emotional state expressed.

4 Pre-requisites for audiovisual integration

The proper alignment of video and audio requires timing information. As the speech timing is determined during speech synthesis, the speech synthesis system has to make this information accessible, so that it can be sent to the visual generation system. Several types of timing information are potentially useful, such as phoneme durations for the correct rendering of mouth shapes, intonation accent information for gesture and face movements, and pausing information for breathing or idle movements. If the information is not made accessible, as is the case for most off-the-shelf commercial speech synthesis systems, then the only way to obtain an approximation of the required information is to re-infer if from an analysis of the synthesised audio data.

5 Application examples

Examples for the application of the ideas presented here are the NECA demonstrators [13] eShowroom, generating sales dialogues, and Socialite, generating social interactions in an interactive soap opera setting. A more abstract feasibility study is the combination of speech synthesis with a photo-realistic facial animation model [14].

6 Questions of validation

A central question in the generation of emotional speech should be that of validation by users. Their imporession of how natural, believable and convincing a system is will be the ultimate acceptability criterion for the system.

As the research field is still rather in a phase of feasibility demonstration rather than optimisation, this question has not yet been addressed in a satisfactory way. The usual methodology in past research on emotional speech synthesis has been that of identification tasks in which all modalities except speech prosody are artificially kept "neutral", and listeners are requested to identify the emotion category expressed [4].

It may be more appropriate for the application of emotional speech in a multimodal setting to ask users how appropriate the voice sounds given the other channels of emotion expression, such as facial expression, behaviour, verbal content, and situational context. Preference tasks, in which the same visual clip is combined with different versions of a speech utterance, may be a promising approach in this direction [11].

7 Conclusion

Emotional speech synthesis is still in its early stages. The approaches presented here leave a choice between naturalness and flexibility. Future developments will need to aim for combining the benefits of both approaches. For user acceptance, validation paradigms will need to be developed taking into account the multi-channel nature of emotion expression in target applications.

Acknowledgments

This research is supported the the EC project NECA IST-2000-28580. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

References

[1] Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, 1997.

[2] Marc Schröder and Jürgen Trouvain. The German Text-to-Speech synthesis system MARY: A tool for research, development and teaching. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, 2001. URL `http://mary.dfki.de`.

[3] Mark R. Walker and Andrew Hunt. *Speech Synthesis Markup Language Specification*. W3C, 2001. URL `http://www.w3.org/TR/speech-synthesis`.

[4] Marc Schröder. Emotional speech synthesis: A review. In *Proceedings of Eurospeech 2001*, volume 1, pages 561–564, Aalborg, Denmark, 2001. URL `http://www.dfki.de/~schroed`.

[5] Paul Ekman. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *Handbook of Cognition & Emotion*, pages 301–320. John Wiley, New York, 1999.

[6] Akemi Iida, Nick Campbell, S. Iga, F. Higuchi, and M. Yasumura. A speech synthesis system with emotion for assisting communication. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 167–172, Northern Ireland, 2000. URL `http://www.qub.ac.uk/en/isca/proceedings`.

[7] W. Lewis Johnson, Shrikanth S. Narayanan, Richard Whitney, Rajat Das, Murtaza Bulut, and Catherine LaBore. Limited domain synthesis of expressive military speech for animated characters. In *Proceedings of ICSLP 2002*, Denver, Colorado, USA, 2002.

[8] Harold Schlosberg. A scale for the judgement of facial expressions. *Journal of Experimental Psychology*, 29:497–510, 1941.

[9] Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology.* MIT Press, Cambridge, MA, USA; London, UK, 1974.

[10] Roddy Cowie, Ellen Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, Winfried Fellenz, and John Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, January 2001.

[11] Marc Schröder. *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis.* PhD thesis, Institute of Phonetics, Saarland University, Germany, 2003. to appear.

[12] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of Eurospeech 2001*, volume 1, pages 87–90, Aalborg, Denmark, 2001. URL `http://www.dfki.de/~schroed`.

[13] Brigitte Krenn, Hannes Pirker, Martine Grice, Paul Piwek, Kees van Deemter, Marc Schröder, Martin Klesen, and Erich Gstrein. Generation of multimodal dialogue for net environments. In *Proceedings of Konvens*, Saarbrücken, Germany, 2002. URL `http://www.ai.univie.ac.at/NECA`.

[14] Irene Albrecht, Jörg Haber, Kolja Khler, Marc Schröder, and H.-P. Seidel. "May I talk to you? :-)" – Facial animation from text. In *Proceedings of Pacific Graphics 2002*, pages 77–86, 2002.