How (Not) to Add Laughter to Synthetic Speech

Jürgen Trouvain[1,2] and Marc Schröder[3]

[1] Institute of Phonetics, Saarland University, Saarbrücken, Germany
[2] Phonetik-Büro Trouvain, Saarbrücken, Germany
trouvain@phonetik-buero.de
[3] German Research Center for Artificial Intelligence DFKI, Saarbrücken, Germany
schroed@dfki.de

Abstract. Laughter is a powerful means of emotion expression which has not yet been used in speech synthesis. The current paper reports on a pilot study in which differently created types of laughter were combined with synthetic speech in a dialogical situation. A perception test assessed the effect on perceived social bonding as well as the appropriateness of the laughter. Results indicate that it is crucial to carefully model the intensity of the laughter, whereas speaker identity and generation method appear less important.

1 Introduction

This paper describes an explorative study of how laughter can be modelled in speech synthesis. In the past, the integration of emotional components in synthetic speech has concentrated on changing the tone of voice of the generated speech in order to elicitate distinguishable emotions or emotional nuances [1]. Another means of vocal emotion expression, which has been neglected in speech synthesis so far, are non-speech vocalisations such as laughter with a high communicative and emotional character [2].

One basic problem is to predict *when* to add laughter in synthetic speech (everybody knows examples of laughter in inappropriate situations). Another basic problem is *how* to add laughter in synthetic speech. This requires some knowledge of types of laughter and how they are used in human-human communication. This also requires some knowledge about the acoustic structure of different laughter types and some ideas about "laughter synthesis".

Obvious problems are: can we copy-synthesise laughter with existing speech material or do we need special recordings? Where do we insert laughter with respect to syntactic and/or prosodic sentence structure? What type of laughter should be inserted? How do we verify that the synthetic speech with added laughter triggers an appropriate perception for the listener?

2 Laughter in Human Interactions

2.1 Forms of Laughter

Although voiced, loud and long laughter is probably the dominant proto-typical form when we think of laughter, most laughs in real dialogues are *mild* forms which are often ignored or just "overheard", even by dialogue annotators [3]. Thus, shorter and less intensive laughs seem more appropriate and more realistic for a convincing dialogue. This is in line with [1] who states that emotional nuances rather than "full-blown" emotions are required in a dialogue, be it human-machine, human-human or machine-machine [4], multi-modal or speech-only (e.g. over a telephone line).

Human laughter shows a great repertoire [5, 6]. Apart from fully voiced laughter, other types of laughter are unvoiced with an oral constriction or nasal turbulences, which is also mirrored in words such as "cackle", "giggle" and "chuckle". Forms of laughter which are distinct from the ones just listed are those which occur simultaneous to speech, so-called speech-laughs which can occur more often in dialogues than autonomous laughs [7]. Speech-laughs are distinct from smiled speech which can *audibly* occur more or less intensely over a longer or shorter period of speech.

2.2 Laughter in Dialogues

Laughter usually occurs as a spontaneous action in a social context in conversations. Despite the infectious effect of laughter which can be considered as a *listener* reaction, especially in a group of more than two people, many if not most laughs come from the *speaker* [8]. Ultimately, if one wanted to model laughter in dialogues as naturally as possible, it would be necessary to consider that speech can occur not just as "neutral" speech, but also as smiled speech and with short speech-laughs. In addition, it is not uncommon for laughter to be realised as joint laughter of both dialogue partners.

2.3 Functions of Laughter

Although laughter and humour are often mentioned in the same breath, laughing is not only an expression of exhilaration and amusement, but is also used to mark irony or a malicious intention [2]. An important function of laughter is social bonding. It occurs in the interactions between mothers and their newborns as well as between adults [8, 9], especially when the interacting partners are unknown to each other [7]. Social bonding is also the function considered in this study for interaction with a computer. This is, however, not a straightforward task because "social laughter" is highly determined by cultural norms which depend on factors such as age, sex, and social status of participants and their relationship [9].

3 Pilot Study

Research question. For the present pilot study, we targeted a simple autonomous laugh with a length of two syllables and a mild intensity. We are interested in two questions: First, does the integration of a laugh in synthetic speech in a dialogue situation lead to the perception of higher social bonding for the listener? Second, is the effect achieved appropriately?

Speech material. Two dialogue extracts were designed as ends of appointment making conversations. The translations of the two extracts are:

```
A: Shall we do it this way?
B: Okay <Laugh> Then we see each other on Monday.

A: On Friday, I am only free after twelve. The best thing will be if we
   meet at one on Friday.
B: All right. <Laugh> That should be fine with me.
```

Speech Synthesis. We use the German diphone speech synthesiser Mary [10] with the Mbrola voices "de6" (male) for speaker A, and "de7" (female) for speaker B, which provide full German diphone sets in “soft”, “modal” and “loud” voice qualities [11]. As a *baseline*, the speech without any laughter was generated with Mary and prosodically "fine-tuned". This baseline version was changed by inserting a laugh after the first phrase of speaker B.

The laugh was generated in six different ways[1]. *Versions 1, 2, and 3* considered the laugh as quasi-speech ("hehe"): the duration and F0 values of natural laughter were superimposed onto "modal", "loud" and "soft" diphones taken from the voice “de7”. *Versions 4 and 5* were unprocessed recordings of natural laughter produced by the speaker of the “de7” database, with different degrees of intensity. The last laugh, a very mild one, was taken from a natural speech corpus with a different female voice (*version 6*).

Perception test. 14 German speaking subjects listened to the audio files in randomised order, in a quiet office via PC loudspeakers. After each stimulus the following question had to be answered: “How well do both speakers like each other?” (a 6-point scale between “very well” and “not at all”). After this first round, all stimuli with the exception of the no-laughter version were presented again, but this time with the question: “How well does the laughter fit into the dialogue?” (same 6-point scale)

Results. The results (see Table 1) show that well-selected laughter can indeed increase the perceived social bonding between speakers. Laughter synthesised from diphones, however, was inefficient in this respect, as was intense laughter. Most efficient was the very soft laugh from the “wrong” speaker. This softest natural version was also rated most appropriate, followed by the medium natural version. Diphone-based laughter was considered slightly inappropriate whereas intense natural laughter was clearly rated as completely inappropriate.

Table 1. Evaluation of different versions of laughter, and of the no-laugh baseline. Responses from 6 (very well) to 1 (not at all)

version	diphone-based			same speaker		other sp.	baseline
	1	2	3	4	5	6	
question	modal	soft	loud	intense	medium	mild	no
like each other?	3.5	3.8	3.7	3.9	4.2	4.7	3.7
appropriate?	2.9	2.4	2.5	1.0	4.7	5.4	-

[1] Stimuli and spectrograms of laughs can be found at http://www.phonetik-buero.de/laughter/

4 Discussion and Summary

The results of the present pilot study give first indications that and how laughter can be added to synthetic speech so that listeners have the feeling of higher social bonding. The results also suggest that inappropriate type or intensity of the laugh can destroy the desired effect in this socially sensitive area.

Normally, a mixing of generation methods (here: entire laugh recordings mixed with diphones) would lead to worse acceptance by listeners [12]. However, our findings hint that these factors may be less important than a careful control of the laugh intensity. Surprisingly, even the use of a different voice did not counter this effect.

If we want to integrate more laughter types (e.g. for situations with irony, which was deliberately excluded here), we have to predict forms of laughter which are appropriately scaled in intensity. This may be true for other affect bursts as well [2]. Clearly more basic research on the phonetics and appearance of laughter is needed.

Acknowledgements

This research is supported in part by Saarland University and in part by the EC Projects NECA (IST-2000-28580) and HUMAINE (IST-507422).

References

1. Schröder, M.: Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis, Saarland University, Saarbrücken (to appear).
2. Schröder, M.: Experimental Study of Affect Bursts. Speech Communic., 40 (2003) 99-116
3. Trouvain, J.: Phonetic Methods in Laughter Research (in preparation)
4. Krenn, B., Pirker, H., Grice, M., Piwek, P., van Deemter, K., Schröder, M., Klesen, M., Gstrein, E.: Generation of Multimodal Dialogue for Net Environments. Proc. Konvens, Saarbrücken, Germany (2002)
5. Bachorowski, J.-A. & Owren, M.J.: Not All Laughs Are Alike: Voiced but not Unvoiced Laughter Readily Elicits Positive Affect. Psychological Science 12 (2001) 252-257
6. Trouvain, J.: Segmenting Phonetic Units in Laughter. Proc. 15th International Conference of the Phonetic Sciences, Barcelona, Spain (2003) 2793-2796
7. Trouvain, J.: Phonetic Aspects of "Speech-Laughs". Proceedings Conference on Orality & Gestuality (ORAGE), Aix-en-Provence, France (2001) 634-639.
8. Provine, R.: Laughter. A Scientific Investigation. Faber & Faber, London (2001)
9. Apte, M.L.: Humor and Laughter. An Anthropological Approach. Cornell UP, Ithaca (1985)
10. Schröder, M. & Trouvain, J.: The German Text-to-Speech Synthesis System MARY. International Journal of Speech Technology 6 (2003) 365-377
11. Schröder, M. & Grice, M.: Expressing Vocal Effort in Concatenative Synthesis. Proc. 15th International Conference of the Phonetic Sciences, Barcelona, Spain (2003) 2589-2592
12. Gong, L. & Lai, J.: To Mix or Not to Mix Synthetic Speech and Human Speech? Contrasting Impact on Judge-Rated Task Performance versus Self-Rated Performance and Attitudinal Responses. International Journal of Speech Technology 6 (2003) 123-131