

Expressing vocal effort in concatenative synthesis

Marc Schröder* and Martine Grice†

* DFKI GmbH, Saarbrücken, Germany
marc.schroeder@dfki.de

† Institute of Phonetics, Saarland University, Saarbrücken, Germany
mgrice@coli.uni-sb.de

ABSTRACT

A new diphone database with a full diphone set for each of three levels of vocal effort is presented. A theoretical motivation is given why this kind of database will be useful for emotional speech synthesis. Two hypotheses are verified in perception experiments: (I) The three diphone sets are perceived as belonging to the same speaker; (II) The vocal effort intended during database recordings is perceived in the synthetic voice. The results clearly confirm both hypotheses.

1 INTRODUCTION

1.1 Motivation: Synthesis of emotional speech

A major problem for the expression of emotion in concatenative synthesis is that the voice quality inherent in the diphones is inappropriate for certain emotions [1]. This can be addressed by treating emotions holistically and making recordings of specific emotion categories [2, 3]. While this approach is straightforward and is likely to lead to highly natural emotion expression, it suffers from a severe shortcoming, namely a lack of flexibility: Only the emotional states which have been recorded can be generated.

We pursue an alternative approach, namely to model the vocal correlates of emotions [4]. This undertaking starts with a decision on how to represent the emotional states themselves [5]. We have chosen to use emotion dimensions [6, 7], a continuous framework for the representation of essential properties of emotional states. The two emotion dimensions which have emerged from a large number of studies are *evaluation* (sometimes called *valence* or *pleasure*) and *activation* (sometimes called *arousal*). The main task in our approach is to find a mapping from a point in the emotion-dimension space to the corresponding acoustic correlates [4], and to realise these in synthetic speech. The question under investigation in this paper is related to the latter aspect.

The emotion dimension best conveyed in speech is activation [8, 9]. Apart from prosodic effects such as raised F0 and speaking rate, activation is known to affect

phonation [10], particularly in terms of vocal effort, caused by higher muscle tension. The spectral change caused by varying vocal effort, to a certain extent corresponding to voice quality, is perceptible [11, 12].

The importance of voice quality modelling for expressing emotions in synthetic speech has been a matter of discussion (summarised in [13]).¹ Voice quality by itself appears to be an indicator of emotional tendency rather than of specific emotions: Gobl and Ní Chasaide [14] found tense and harsh voice qualities to be rated as aggressive and aroused, while breathy, whispery and creaky voice were rated as unaggressive and relaxed. This fits well with our goals formulated above: to express emotional tendencies, primarily the *activation* dimension. We do this by recording different levels of vocal effort.

1.2 Choice of synthesis technology

Our approach to expressing emotions, using a model-based mapping from continuous emotion dimensions onto acoustic correlates, requires fine-grained control over the prosodic parameters of the synthesised voice. This requirement is currently incompatible with state-of-the-art unit selection synthesis systems: They draw their naturalness from *not* interfering with the recorded speech signal, and thus rarely allow for an explicit modelling of prosody. Even conceptually, controlling prosody in such systems is only possible within the limits of the units recorded in the database, unless signal manipulation techniques are used, which invariably lead to a reduction in the quality of the synthesised speech. Rule-based formant synthesis, on the other hand, does allow for a wide array of acoustic parameters to be modelled. For this reason, it has been the technology of choice for a number of emotional speech synthesis undertakings [13]. However, due to lack of naturalness, it has nearly disappeared from the landscape of commercial speech synthesis systems. There are promising new approaches, such as data-driven formant synthesis [15], but these are still in an early development phase.

¹A tentative conclusion is that some emotions are conveyed mainly through voice quality, others mainly through prosody [1, 13].

An imperfect, but viable compromise is the use of diphone synthesis, which allows for fine-grained prosody modelling, with a limited degree of distortion. Since there is so far no means of modelling voice quality in diphone synthesis (although there are some promising developments pointing towards future possibilities in this domain, e.g., [16, 17]), we recorded separate diphone databases for three levels of vocal effort.

2 METHOD

2.1 Diphone recordings

One male speaker of standard German produced a full German diphone set for each of three degrees of vocal effort. For simplicity, we refer to these as “soft”, “modal” and “loud”.

A modified version of the FESTVOX toolkit [18] was used for the voice recordings, prompting the speaker to produce nonsense words on a constant pitch. Recordings were carried out using an AKG D330 BT microphone and a Sony MZ-R900 minidisc recorder serving simultaneously as a session log and as a pre-amplifier. Each utterance was recorded directly onto the harddisk of a computer, and was then played to four experts who verified vocal effort, pitch constancy and phonetic correctness. Immediate re-recordings were requested when necessary.

The data was automatically labelled, hand-corrected, and converted to the MBROLA [19] format. The resulting synthesis voice, referred to as de6, is available² free of charge for non-commercial use.

2.2 Perception test I: Speaker identity

A first perception test was carried out to test the hypothesis that the three diphone sets are sufficiently similar to be recognised as belonging to the same person.

A pair of sentences was designed to share a minimal number of phones, making sure that the second sentence could be a suitable answer to or continuation of the first: (1) “Der Mann ist gut zu erkennen” (The man is easy to recognise, [dɛʁmanʔɪstɡu:tʃu:ʔɛkɛnən]) and (2) “Die Frau ist eher verschwommen” (The woman is rather blurred, [di:fʁaʊʔɪstʔɛrɛfʃvɔmən]). These were synthesised using each of the three diphone sets, along with two additional male voice diphone databases, referred to as de2 and de4, with both normal (115–74 Hz) and raised (161–103 Hz) pitch.

The resulting 20 stimuli (one voice with three levels of vocal effort and two reference voices, all at two pitch levels, two sentences) were paired in all 55 possible combinations in the order: sentence (1) — sentence (2), and sorted into four pseudo-randomisations.

The test consisted of a training session using two female voices (8 stimulus pairs), one block of 20 pairs out of the 55 duplicated for the purpose of habituation, and the test proper. 16 native German speakers (8 naive and 8 expert, 4 male and 4 female each) indicated in a forced-choice task whether the stimuli in each pair were produced by the same speaker or by two different ones. Stimuli were presented over headphones. Each subject was presented one of the four randomisations only.

2.3 Perception test II: Perceived effort

A second perception test was carried out to test the hypothesis that the effort intended during the recordings is perceived in the synthesised material.

12 stimuli from test (I) were used in this test (3 levels of vocal effort * 2 pitch levels * 2 sentences).

20 subjects (13 male, 7 female) participated in this test. They were all native German speakers, and none of them had taken part in the first test.

Stimuli were presented over headphones in a self-paced test, using a graphical user interface, in which subjects rated the stimuli on a continuous scale from “without effort” to “with great effort”. Since the stimuli were amplitude-normalised, subjects were instructed to base their ratings on the “sound of the voice” rather than the “loudness”. The test consisted of a short training phase using 6 stimuli from a female voice, followed by one presentation of each of the 12 stimuli in random order.

3 RESULTS

3.1 Speaker identity

Pairs of sentences with identical settings (same voice, same pitch, same vocal effort) were perceived as the same person in 99.5% of the cases. This value can serve as a reference with which to compare the distortions introduced by varying the different parameters.

With a constant pitch, 79.9% of stimulus pairs differing in vocal effort were recognised as produced by the same person. This highly significant effect (Chi-square, $p < .001$) confirms the central hypothesis in this test, namely that the three diphone databases differing in vocal effort are still perceived as the same speaker.

Several interesting effects were observed which are indirectly related to the research question. One is the differentiation of the different voices, independently of vocal effort. With a constant pitch, the three voices were distinguished from each other 63.2% of the time. The voices differed markedly in their similarity: The de4 voice was easily distinguished from our voice, de6 (83.3% correct with constant pitch), and from the de2 voice (72.9% correct). On the other hand, de2 and de6

²<http://tcts.fpms.ac.be/synthesis/mbrola/dba/de6/de6.zip>

were perceptually very similar (only 46.5% correct, i.e. same/different judgments close to chance level). This similarity is greatest between de2 and the “modal” diphone set of de6 (only 29.2% of subjects think these are different voices), followed by the “soft” (45.8% correct) and “loud” (64.6% correct) diphone sets.

Another interesting observation relates to the effect of pitch on the assignment of speaker identity. This effect was greater than the effect of vocal effort: only 45.5% of the pairs differing only in pitch were rated as the same speakers. This effect also seems to depend on the “distinctiveness” of a voice: While pairs consisting of two de4 or two de6-loud utterances differing only in pitch were correctly identified as the same speaker in 56.3% of the time, this was less the case for de2 (31.3% correct) and for the de6-soft and de6-modal diphone sets (37.5% and 40.6% correct, respectively).

A further interesting observation is the fact that although there were considerable differences across subjects, there was no significant effect of experience with speech synthesis, with one exception: In the same voice/different pitch condition, experts (correctly) rated the stimuli as the same speakers in 57.1% of the cases, compared to only 33.9% for non-experts. This difference is significant (Chi-square, $p = .014$).

3.2 Perceived effort

The mean effort ratings of the stimuli are shown in Figure 1. Intended vocal effort, pitch, and sentence were entered as independent variables in an ANOVA. The main effect of intended vocal effort is highly significant ($F(2, 228) = 53.1, p < .001$). The effect of pitch is not significant ($F(1, 228) = 1.8, p = .18$). There was also a main effect of sentence ($F(1, 228) = 14.3, p < .001$): Sentence (1) was consistently rated as involving more effort than sentence (2). The only significant interaction effect is between intended vocal effort and pitch ($F(2, 228) = 3.8, p = .025$), and appears to be due to the fact that for soft and modal, but not for loud diphones, higher pitch is perceived as expressing more effort.

Post-hoc Tukey t-tests of the intended vocal effort variable showed that all differences between diphone sets were as expected (i.e., effort ratings for soft diphones were lower than those for modal or loud diphones, and effort ratings for modal diphones were lower than those for loud diphones) at a highly significant level (all $p < .001$).

4 DISCUSSION

Both hypotheses have been confirmed: Test (I) has shown that perceived speaker identity is little affected by intended vocal effort as manifested in the three diphone sets. Test (II) has confirmed that the vocal ef-

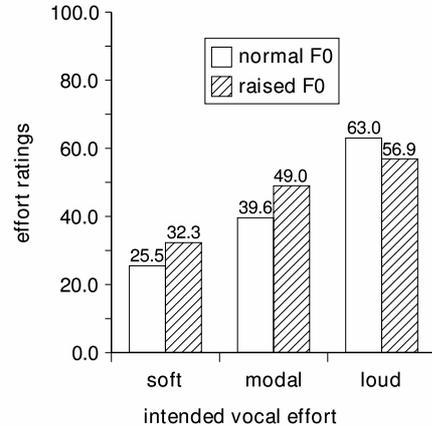


Figure 1: Effort ratings, on a scale from 0 (no effort) to 100 (maximum effort).

fort is perceived as intended. These results indicate the appropriateness of our general approach to expressing vocal effort in concatenative synthesis.

It is true that the method we used for recording the different levels of vocal effort is not well-controlled, but relies on the speaker maintaining a constant level of effort and on the expert listeners perceiving deviations. This method is particularly vulnerable to slow changes of voice quality during a recording session. It would be desirable to have access to technology verifying the consistency of a recorded utterance’s voice quality with the previously recorded data. However, we are not aware of any such software being publicly available.

The fact that there were differences in effort perceived across the two sentences could be interpreted as a random difference, introduced during the recordings of the diphone inventories used. While this interpretation cannot be excluded, it seems somewhat unlikely given the fact that the effect is present for each of the three diphone sets.

This study also reveals a general problem regarding the use of digital signal processing in concatenative speech synthesis, namely the distortions introduced: A modification of pitch slightly beyond the range typically used in non-emotional synthesis (but still moderate in view of emotional speech) caused speaker identity ratings to drop to around or below chance level. This raises the question of the usability for emotional speech of diphone synthesis, or indeed of any concatenative synthesis based on digital signal processing.

As mentioned above, the two available alternatives, formant synthesis and unit selection synthesis, suffer from equally severe problems. It would appear, therefore, that any approach to emotional speech synthesis will currently be confronted with import limitations of the selected synthesis technology. In our case, this has the practical consequences that we should take care to

avoid large abrupt shifts in F0 across sentences, and to use markedly distinct voices when synthesising spoken dialogues.

5 CONCLUSION

This paper has reported on the creation and perceptual evaluation of a new German diphone voice with three full diphone sets expressing different levels of vocal effort. Despite distortions from the MBROLA signal manipulation algorithm, the vocal effort is perceived as intended, and the diphone sets are perceived as belonging to the same speaker.

ACKNOWLEDGMENTS

We would like to thank Cordula Klein for her contributions towards the design of the diphone database, and Stefan Baumann, Cordula Klein and Uta Panten for recording and labelling it.

This research is supported by the EC Project NECA IST-2000-28580.

REFERENCES

- [1] J. M. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enríquez, and J. M. Pardo, "Analysis and modelling of emotional speech in Spanish," in *Proc. 14th ICPHS*, San Francisco, USA, 1999, pp. 957–960.
- [2] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A speech synthesis system with emotion for assisting communication," in *Proc. ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 167–172, <http://www.qub.ac.uk/en/isca/proceedings>.
- [3] M. Bulut, S. S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesiser," in *Proc. 7th ICSLP*, Denver, Colorado, USA, 2002.
- [4] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001, vol. 1, pp. 87–90, <http://www.dfki.de/~schroed>.
- [5] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication Special Issue on Speech and Emotion*, vol. 40, no. 1–2, pp. 5–32, 2003.
- [6] H. Schlosberg, "A scale for the judgement of facial expressions," *J. Exp. Psychol.*, vol. 29, pp. 497–510, 1941.
- [7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
- [8] B. Tischer, *Die vokale Kommunikation von Gefühlen*, vol. 18 of *Fortschritte der psychologischen Forschung*, Psychologie-Verlags-Union, Weinheim, Germany, 1993.
- [9] K. R. Scherer and J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motiv. Emot.*, vol. 1, no. 4, pp. 331–346, 1977.
- [10] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychol. Bull.*, vol. 99, pp. 143–165, 1986.
- [11] I. Lehiste and G. E. Peterson, "Vowel amplitude and phonemic stress in American English," *JASA*, vol. 31, pp. 428–435, 1959.
- [12] A. Eriksson and H. Traunmüller, "Perception of vocal effort and speaker distance on the basis of vowel utterances," in *Proc. 14th ICPHS*, San Francisco, USA, 1999, pp. 2469–2472.
- [13] M. Schröder, "Emotional speech synthesis: A review," in *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001, vol. 1, pp. 561–564, <http://www.dfki.de/~schroed>.
- [14] C. Gobl and A. Ní Chasaide, "Testing affective correlates of voice quality through analysis and resynthesis," in *Proc. ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 178–183, <http://www.qub.ac.uk/en/isca/proceedings>.
- [15] R. Carlson, T. Sigvardson, and A. Sjölander, "Data-driven formant synthesis," Progress Report 44, KTH, Stockholm, Sweden, 2002, <http://www.speech.kth.se/qpsr/tmh>.
- [16] H. Kasuya, K. Maekawa, and S. Kiritani, "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics," in *Proc. 14th ICPHS*, San Francisco, USA, 1999, pp. 2505–2512.
- [17] A. Tassa and J.S. Liénard, "A new approach to the evaluation of vocal effort by the PSOLA method," *WEB-SLS, The European Student Journal of Language and Speech*, 2000, <http://www.essex.ac.uk/web-sls/papers/00-01/00-01.html>.
- [18] A. Black and K. Lenzo, "Festvox: Building synthetic voices, edition 1.6," Tech. Rep., Language Technologies Institute, Carnegie Mellon University, PA, USA, 2002, <http://www.festvox.org>.
- [19] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesisers free of use for non commercial purposes," in *Proc. 4th ICSLP*, Philadelphia, USA, 1996, pp. 1393–1396.