Emotional Speech Synthesis: A Review

Marc Schröder

DFKI, Saarbrücken, Germany
Institute of Phonetics, University of the Saarland
schroed@dfki.de

Abstract

Attempts to add emotion effects to synthesised speech have existed for more than a decade now. Several prototypes and fully operational systems have been built based on different synthesis techniques, and quite a number of smaller studies have been conducted. This paper aims to give an overview of what has been done in this field, pointing out the inherent properties of the various synthesis techniques used, summarising the prosody rules employed, and taking a look at the evaluation paradigms. Finally, an attempt is made to discuss interesting directions for future development.

1. Introduction

With the intelligibility of synthetic speech approaching that of human speech, the need for increased naturalness becomes more palpable. One of the aspects of naturalness most obviously missing in synthetic speech is appropriate emotional expressivity. This observation has been motivating attempts to incorporate the expression of emotions into synthetic speech for more than a decade, and such attempts seem to have gained popularity in recent years. While advances in other aspects of naturalness of synthetic voices have been made, notably with unit selection techniques, the synthesis of emotional speech still has a long way to go.

In the studies concerned with the expression of emotion in synthetic speech that can be found in literature, an interesting variety of approaches has been employed. This paper will try to give an overview of these studies, and work out the differences and similarities in approaches, techniques and underlying assumptions. First, the studies are presented in groups according to the type of synthesis technique employed, which coincides in many cases with similarities in the approach. Next, prosody rules employed for expressing emotions are reported, and the paradigms used for evaluation are discussed.

Finally, a number of points will be discussed related to possible directions for future development. These are in part inspired by the ISCA Workshop on Speech and Emotion, recently held in Northern Ireland [1], which for the first time brought together researchers interested in speech and emotion from a large variety of backgrounds. This fruitful exchange showed, among other things, that our understanding of the way emotion is expressed in speech can be improved along two axes: On the one hand, the description of the vocal correlates of emotions ("How is a given emotion expressed in speech?"); on the other hand, the description of the emotional states themselves [2] ("What are the properties of the emotional state to be expressed? What is the relation between this state and another state?"). Some implications for future research in the synthesis of emotional speech are proposed in the discussion section.

2. Existing approaches and techniques

The modelling of emotion in speech relies on a number of parameters like, among others, fundamental frequency (F0) level, voice quality, or articulation precision (see 3. below). Different synthesis techniques provide control over these parameters to very different degrees.

2.1. Formant synthesis

Formant synthesis, also known as rule-based synthesis, creates the acoustic speech data entirely through rules on the acoustic correlates of the various speech sounds. No human speech recordings are involved at run time. The resulting speech sounds relatively unnatural and "robot-like" compared to state-of-the-art concatenative systems, but a large number of parameters related to both voice source and vocal tract can be varied quite freely. This, of course, is interesting for modelling emotional expressivity in speech.

Several larger undertakings [3][4][5][6][8][9] have used Formant synthesisers because of the high degree of control that they provide. These include the first ones, from 1989: Janet Cahn's Affect Editor [3][4], and Iain Murray et al.'s HAMLET [5][6]. Both have used DECtalk as a formant synthesis system, providing dedicated processing modules which adapt their input according to the acoustic properties of a number of emotions. In both cases, the acoustic profile for each emotion category was derived from the literature and manually adapted. While the Affect Editor requires the input to be manually annotated, HAMLET processes its input entirely by rule.

Within the VAESS project ("Voices, Attitudes and Emotions in Speech Synthesis"), which ran from 1994 to 1996, emotional expressivity was to be added to a formant synthesiser. Montero et al. [7] report reasonable success for the modelling of three emotions (hot anger, happiness, and sadness) in Spanish using global prosodic and voice quality parameter settings.

Burkhardt, in his 2000 PhD [8][9], has also chosen to use formant synthesis, despite the reduced naturalness, because of the high degree of flexibility and control over acoustic parameters that this technique provides. His systematic, perception-oriented approach to finding good acoustic correlates of emotions for German consisted of two main steps. In a first step, he systematically varied five acoustic parameters known to be related to emotion, without using prior knowledge from the literature about the best parameter values for a given emotion. The resulting stimuli were presented in a perception test, providing perceptually optimal parameter values for each emotion studied. In a second step, these optimal values were taken as the basis for the exploration of a wider set of parameters, inspired from the

literature, and the resulting variants were presented in another perception test.

2.2. Diphone concatenation

In concatenative synthesis, recordings of a human speaker are concatenated in order to generate the synthetic speech. The use of diphones, i.e. stretches of the speech signal from the middle of one speech sound ("phone") to the middle of the next, is common. Diphone recordings are usually carried out with a monotonous pitch. At synthesis time, the required F0 contour is generated through signal processing techniques which introduce a certain amount of distortion, but with a resulting speech quality usually considered more natural than formant synthesis.

In most diphone synthesis systems, only F0 and duration (and possibly intensity) can be controlled. In particular, it is usually impossible to control voice quality.

Fundamental to every attempt to use diphone synthesis for expressing emotions is the question whether F0 and duration are sufficient to express emotion, i.e. whether voice quality is indispensable for emotion expression or not. Interestingly, very different results were obtained by different studies. While [12][14][16][17][19][20] report that synthesised emotions can be recognised at least reasonably well, [13][15] report recognition rates close to chance level. The reason may be that there is no simple general answer: [16] reported that for a given speaker, the relative contribution of prosody and voice quality to emotion recognition depends on the emotion expressed, and [17] has found evidence that this may, in addition, be speaker-dependant. In other words, there seem to be speaker strategies relying mostly on F0 and duration for expressing some emotions, and these can be successfully modelled in diphone synthesis. Whether this is true for all types of emotion is not clear yet.

One approach to emotional speech synthesis with diphones, used by [12][13][14][16][17], is copy synthesis: F0 and duration values are measured for each speech sound in a given utterance (usually an actor's portrayal of an emotion), and used for synthesising the same utterance from diphones. The result is a synthetic utterance with the same F0 and duration values as the actor's speech, but the voice quality determined by the diphones. This technique is suitable for modelling what humans do as closely as possible with the given parameter set. Whether that is the best way to obtain perceptually optimal, believable expressions can be questioned, though: E.g. in the domain of animated characters, it has been observed that features occurring in human expression need to be exaggerated in synthetic expression in order to be believable [26].

A more ambitious approach is the formulation of prosody rules for emotions [10][11][15][18][19][20] (see 3. below for more details).

2.3. Unit selection

The synthesis technique often perceived as being most natural is unit selection, or large database synthesis, or speech re-sequencing synthesis. Instead of a minimum speech data inventory as in diphone synthesis, a large inventory (e.g., one hour of speech) is used. Out of this large database, units of variable size are selected which best approximate a desired target utterance defined by a number of parameters. These parameters can be the same as used in diphone synthesis, i.e. phoneme string, duration and F0, or they could be different. The weights assigned to the selection parameters influence which units are selected. If well-matching units are found in the database, no signal processing is necessary. While this synthesis method often gives very natural results, the results can be very bad when no appropriate units are found.

The feature of unit selection synthesis to preserve the features of the recorded speech very well has been exploited by Iida et al. [21] for the synthesis of emotional speech. For each of three emotions (anger, joy, and sadness), an entire unit selection database was recorded by the same speaker. In order to synthesise a given emotion, only units from the corresponding database are selected. The emotions in the resulting synthesised speech are well recognised (50-80%).

Another, theoretically more demanding approach is to select the material appropriate for the targeted emotion from one database. The equivalent of prosody rules is then used as selection criteria. This has been attempted by Marumoto & Campbell [22], who used parameters related to voice quality and prosody as emotion-specific selection criteria. The results indicated a partial success: Anger and sadness were recognised with up to 60% accuracy, while joy was not recognised above chance level.

3. Prosody rules employed

In the literature concerned with emotional speech synthesis, global prosodic parameters are often treated as universal or near universal cues for emotion. While this claim can certainly be the subject of debate, there seems to be some limited support for it, e.g. [23][24].

At least in formant and time-domain synthesis, prosody rules are at the heart of automatically generated emotional expressivity in synthetic speech. Such rules have been obtained in a number of ways by different authors. [4][6][15][19][20] have extracted rules from literature; [7][11][18][22] have carried out their own corpus analysis; and [9][10] have obtained perceptually optimal values by systematic parameter variation in synthesis.

The types of parameter modelled vary greatly between different studies. All studies agree on the importance of global prosodic settings, such as F0 level and range, speech tempo and eventually loudness. Some studies try to go into more detail about these global settings, modelling e.g. steepness of the F0 contour during rises and falls [4][6][18][20], distinguishing between articulation rate and the number and duration of pauses [4][6][16][18], or modelling additional phenomena like voice quality [4][6][9][15][18][19][22] or articulation precision [4][6][9][15]. A further step is the consideration of interactions with linguistic categories, like further distinguishing between the speech tempo of vowels and consonants [6][15][20], or of stressed and unstressed syllables [6][9][20], or the placement of pauses within utterances [4]. Only rarely taken into account is the influence of linguistic prosodic categories, like F0 contours [9][11], although these have been shown to play an important role in emotion recognition [9][11].

In the following, a short overview of prosody rules is given that have been successfully employed to express a number of emotions. Instead of a reduced summary of all the rules employed in different studies, one successful modelling example per emotion is presented in detail, along with the recognition rate obtained (Table 1).

Emotion Study Language Rec. Rate	**Parameter settings**
Joy [9] German 81% (1/9)	**F0 mean**: +50% **F0 range**: +100% **Tempo**: +30% **Voice Qu.**: modal or tense; "lip-spreading feature": F1 / F2 +10% **Other**: "wave pitch contour model": main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness [4] American English 91% (1/6)	**F0 mean**: "0", reference line "-1", less final lowering "-5" **F0 range**: "-5", steeper accent shape "+6" **Tempo**: "-10", more fluent pauses "+5", hesitation pauses "+10" **Loudness**: "-5" **Voice Qu.**: breathiness "+10", brilliance "-9" **Other**: stress frequency "+1", precision of articulation "-5"
Anger [6] British English	**F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear [9] German 52% (1/9)	**F0 mean**: "+150%" **F0 range**: "+20%" **Tempo**: "+30%" **Voice Qu.**: falsetto
Surprise [4] American English 44% (1/6)	**F0 mean**: "0", reference line "-8" **F0 range**: "+8", steeply rising contour slope "+10", steeper accent shape "+5" **Tempo**: "+4", less fluent pauses "-5", hesitation pauses "-10" **Loudness**: "+5" **Voice Qu.**: brilliance "-3"
Boredom [10] Dutch 94% (1/7)	**F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C, avoid final patterns 5&A and 12

Table 1. Examples of successful prosody rules for emotion expression in synthetic speech. Recognition rates are presented with chance level for comparison. Sadness and Surprise: Cahn uses parameter scales from –10 to +10, 0 being neutral; Boredom: Mozziconacci indicates intonation patterns according to a Dutch grammar of intonation, see [10] for details.

4. Evaluation paradigms

There seems to be a de-facto standard for evaluation of synthetic speech, i.e. a methodology employed by almost everyone. However, whether that method is actually the most suitable may be discussed.

The typical way of evaluating the quality of the resulting synthetic emotional speech is through a forced choice perception test including the emotion categories actually modeled, employing a small number of semantically neutral carrier sentences [4][7][9][11][12][13][14][15][16][17][21][22]. It can be argued, though ([25], p. 615), that this amounts rather to a discrimination task than an identification task, especially when the number of categories involved is small. A forced choice test provides no information about the quality of the stimulus in terms of naturalness or believability. Therefore, a number of studies assess the degree of naturalness, believability or overall preference of the emotion expression in addition to the forced choice rating, often on a five-point scale [4][15][17][21]. In addition, the intensity of the emotion [4] or the synthetic speech intelligibility [21] have been assessed. The advantages of such a forced-choice test are that it is relatively easy to carry out, provides a simple measure of recognition relative to chance level and allows a limited comparison between studies.

Another possibility, especially suited for finding phenomenons not expected by the experimenter, are free response tests [6][17]. A subsequent grouping of the responses into meaningful classes can be performed using validated word lists [6].

An interesting alternative evaluation paradigm was employed by Murray & Arnott [6] and recently adopted by Stallo [20]. First, a number of “distractor” response categories are introduced in the perception test, as well as a category “other”. In addition, semantically neutral as well as semantically emotional texts are used, both synthesised with neutral and emotional prosody. The difference in recognition between the version with neutral prosody and the version with emotional prosody is then taken as the measure for the perceptive impact of the prosody rules. Interestingly, the recognition *improvement* due to prosody was bigger for emotional texts than for neutral texts.

In an audio-visual context, a talking head visually expressing emotion [20] was presented with neutral and with emotional synthetic speech. Subjects rated which version they perceived as more natural, more understandable, etc. The version with emotional speech was clearly preferred.

5. Discussion

Emotional speech synthesis is not yet applicable in many real life settings. A number of structural problems which seem to contribute to that are discussed in the following.

In most studies, a number of between three and nine discrete, extreme emotional states are modelled. The often implicit assumption that the expression of a few basic or primary emotion categories is most important to model, and that other emotional states can somehow be derived from that, has been questioned by Cowie [2]. He argued that systems should be able to express less intense emotions more suitable for real life applications. For a perception-oriented task such as synthesis of emotional speech, a listener-oriented taxonomy like the FEELTRACE dimensions [27] may be a suitable starting point for describing non-extreme emotional states.

Besides the gradual, global parameter settings such as F0 mean, overall speech tempo etc., it is well known that linguistic categories such as F0 contour can have an effect on emotion perception in interaction with other linguistic information like sentence type [28], [29] (p. 8). Such effects,

most likely language-specific in nature, are not yet appropriately accounted for in emotional speech synthesis.

As pointed out earlier (see 2.), synthesis techniques currently seem to show a trade-off between flexibility of acoustic modelling and perceived naturalness. In order to express a large number of emotional states with a natural-sounding voice, either the rule-based techniques need to become more natural-sounding (see e.g. [30]), or the selection-based techniques must become more flexible [22].

Finally, evaluation techniques should be developed that are more suitable for assessing the appropriateness of acoustic parameter settings for a given communication situation. This might be achieved by moving away from forced-choice tests using abstract emotion words towards tests measuring the perceived naturalness of an utterance given an emotion-defining context.

6. References

[1] Cowie, R., Douglas-Cowie, E., & Schröder, M. (eds.), *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Northern Ireland 2000. Belfast: Textflow. http://www.qub.ac.uk/en/isca/proceedings.

[2] Cowie, R., Describing the Emotional States Expressed in Speech, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 11-18.

[3] Cahn, J. E., *Generating Expression in Synthesized Speech*, Master's Thesis, MIT, 1989. http://www.media.mit.edu/~cahn/masters-thesis.html

[4] Cahn, J. E., The Generation of Affect in Synthesized Speech, *Journal of the American Voice I/O Society*, 8, July 1990, p. 1-19.

[5] Murray, I. R., *Simulating emotion in synthetic speech*, PhD Thesis, University of Dundee, UK, 1989.

[6] Murray, I. R., & Arnott, J. L., Implementation and testing of a system for producing emotion-by-rule in synthetic speech, *Speech Communication, 16*, p. 369-390.

[7] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., & Pardo, J. M., Emotional Speech Synthesis: From Speech Database to TTS, *ICSLP 98, Vol. 3*, p. 923-926.

[8] Burkhardt, F., *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren* [Simulation of emotional manner of speech using speech synthesis techniques], PhD Thesis, TU Berlin, 2000. http://www.kgw.tu-berlin.de/~felixbur/publications/diss.ps.gz

[9] Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 151-156.

[10] Mozziconacci, S. J. L., *Speech Variability and Emotion: Production and Perception*, PhD Thesis, Technical University Eindhoven, 1998.

[11] Mozziconacci, S. J. L., & Hermes, D. J., Role of intonation patterns in conveying emotion in speech, *ICPhS 1999*, p. 2001-2004.

[12] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., Duration and Intonation in Emotional Speech, *Eurospeech 93, Vol. 1*, p. 577-580.

[13] Heuft, B., Portele, T., & Rauth, M. (1996), Emotions in Time Domain Synthesis, *ICSLP 96*.

[14] Edgington, M., Investigating the Limitations of Concatenative Synthesis, *Eurospeech 97*.

[15] Rank, E., & Pirker, H., Generating Emotional Speech with a Concatenative Synthesizer, *ICSLP 98, Vol. 3*, p. 671-674.

[16] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M., Analysis and Modelling of Emotional Speech in Spanish, *ICPhS 99*, p. 957-960.

[17] Schröder, M., Can emotions be synthesized without controlling voice quality?, *Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland*, p. 37-55. http://www.dfki.de/~schroed.

[18] Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., & Longhi, L., Validation of an Acoustical Modelling of Emotional Expression in Spanish using Speech Synthesis Techniques, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 161-166.

[19] Murray, I. R., Edgington, M. D., Campion, D., & Lynn, J., Rule-based Emotion Synthesis Using Concatenated Speech, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 173-177.

[20] Stallo, J., *Simulating Emotional Speech for a Talking Head*, Honours Thesis, School of Computing, Curtin University of Technology, Australia, 2000. http://www.computing.edu.au/~stalloj/projects/honours

[21] Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M., A Speech Synthesis System for Assisting Communication, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 167-172.

[22] Marumoto, T., & Campbell, N., Control of speaking types for emotion in a speech re-sequencing system [in Japanese], *Proc. of the Acoustic Society of Japan, Spring meeting 2000*, p. 213-214.

[23] Chung, S.-J., Vocal Expression and Perception of Emotion in Korean, *ICPhS 99*, p. 969-972.

[24] Tickle, A., English and Japanese Speakers' Emotion Vocalisation and Recognition: A Comparison Highlighting Vowel Quality, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 104-109.

[25] Banse, R., & Scherer, K. R., Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology, 70(3)*, 1996, p. 614-636.

[26] Bates, J., The Role of Emotion in Believable Agents, *Communications of the ACM, 37*, 1994, p. 122-125. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/papers/ba-and-emotion.ps

[27] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M., 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 19-24.

[28] Scherer, K. R., Ladd, D. R., & Silverman, K., Vocal cues to speaker affect: Testing two models, *Journal of the Acoustic Society of America, 76(5)*, 1984, p. 1346-1356.

[29] Andreeva, B., & Barry, W. J., Intonation von Checks in der Sofia-Varietät des Bulgarischen [Intonation of Checks in the Sofia variety of Bulgarian], *Phonus 4, Research Report Institute of Phonetics, University of the Saarland*, 1999, p. 1-13.

[30] Kasuya, H., Maekawa, K., & Kiritani, S., Joint Estimation of Voice Source and Vocal Tract Parameters as Applied to the Study of Voice Source Dynamics, *ICPhS 99*, p. 2505-2512.