

CAN EMOTIONS BE SYNTHESIZED WITHOUT CONTROLLING VOICE QUALITY?

Marc Schröder

Abstract

The present study addresses the question whether it is in principle feasible to convey emotion in synthesized speech using a restricted parameter set which can usually be controlled in concatenation based synthesizers. Using copy synthesis, the prosodic structure of one sentence uttered with five emotional expressions (anger, joy, fear, sadness, and neutral) was transferred to synthetic stimuli. Perception tests show that for some synthetic stimuli, the high recognition rates for the corresponding natural stimuli are almost reproduced, while for other stimuli the emotional information is lost. In a free association perception test, a tendency towards the perception of the unintended category "disappointment" was found that only varied to a limited extent across stimuli.

Die vorliegende Untersuchung widmet sich der Frage, ob Emotionen prinzipiell mit einem begrenzten Parametersatz vermittelt werden können, wie er üblicherweise in konkatener Synthese zur Verfügung steht. Mittels Kopiersynthese wurde die prosodische Struktur eines Satzes, der mit fünf emotionalen Ausdrücken (Wut, Freude, Angst, Traurigkeit, und neutral) produziert worden war, auf synthetische Stimuli übertragen. Perzeptionstests zeigen, daß für manche Stimuli die hohen Erkennungsraten der entsprechenden natürlichen Stimuli nahezu reproduziert werden, während für andere Stimuli die emotionale Information verlorengeht. In einem Perzeptionstest zur freien Assoziation wurde eine Tendenz zur Wahrnehmung der unbeabsichtigten Kategorie "Enttäuschung" festgestellt, die nur bedingt zwischen den Stimuli variierte.

1. Introduction

As synthetic speech gets more intelligible, the wish to make it more natural-sounding increases too, motivating attempts to synthesize emotional speech. As Scherer (1996)

says, "much of speech synthesis is flawed by the lack of appropriate affective variation in prosody and voice quality which seems to be required for both intelligibility and acceptability".

However, recent attempts to synthesize emotional speech using a concatenative synthesizer have yielded very contradictory results. Some studies (Vroomen *et al.*, 1993; Edgington, 1997; Montero *et al.*, 1998) report results which show that synthesized emotions can be recognized reasonably well. Vroomen *et al.* (1993) even report a mean recognition rate of 81%, synthesizing seven emotions by copying only duration and intonation onto a monotonous utterance. On the other hand, some articles report recognition rates close to chance level (Heuft *et al.*, 1996; Rank & Pirker, 1998). This background makes it clear that, for results to be interpretable and reproducible, the multitude of factors in experimental design contributing to a result, especially in this field of synthetic emotions, must be closely controlled and reported in detail.

One major problem in every attempt to produce synthetic emotional expressions is voice quality. As Scherer (1986) explained in detail, the physiological processes that accompany emotion influence phonation, and thus voice quality, in numerous ways. Murray & Arnott (1993) state that "voice quality is important for communication of emotion" (p.1099). In the field of speech synthesis, this has recently been confirmed by Montero *et al.* (1998), who found the influence of diphone voice quality on synthetic emotion recognition to be more important than other prosodic cues.

The present study goes back one step and investigatea the basic question of whether it is in principle feasible to convey emotions in concatenation-based synthetic speech in a recognizable and convincing way. A positive answer to that question does not require high overall recognition rates for the set of stimuli used in this study, as we can by no means expect all or even many of our stimuli to be "good examples"; rather, the feasibility would show in the existence of single, highly recognizable and convincing stimuli. Starting from the observation that we do not know which precise acoustic parameters determine the perception of a particular emotion, all abstraction in modeling the prosodic features of emotional expression is intentionally avoided. Instead, the controllable parameters F0, duration and energy are transferred as precisely as possible from emotionally spoken natural utterances of a sentence to synthetic versions of that sentence.

2. Natural speech samples

All investigations were carried out on different versions of the German sentence "Er ist einfach weggefahren!" (*He just [=simply] left [by car]*). This sentence was chosen because on the one hand, it is very unspecific, i.e. it can have very different meanings in different contexts, and on the other hand, because the German word "einfach" has an emotionally intensifying function.

For the four emotions anger (G. "Wut"), joy (G. "Freude"), sadness (G. "Traurigkeit"), and fear (G. "Angst"), frame stories were invented in order to facilitate the speaker's task of simulating the expression of the four emotions. The stories are summarized as follows. For the expression of sadness, the speaker uttered the sentence after a good friend had left without saying good-bye ("Ich kann mich nicht mal mehr verabschieden! Er ist einfach weggefahren!" / *"I can't even say good-bye anymore! He just left!"*). For fear, the speaker had to imagine being a small child, staying with his/her sister alone in a hut in the forest at night, after their father left without explanation ("Papa ist nicht da! Er ist einfach weggefahren!" / *"Daddy isn't there! He just left!"*). For joy, the speaker was informed by phone that he/she got the job he/she hoped for because the only other candidate had withdrawn ("Ich hab' den Job! Der andere hat aufgegeben! Er ist einfach weggefahren!" / *"I got the job! The other one has given up! He just left!"*). For anger, the speaker saw his/her unreliable colleague leave by car at the moment of a really urgent meeting ("Dieser unzuverlässige Idiot! Er ist einfach weggefahren!" / *"This unreliable idiot! He just left!"*)

2.1. Recordings

Three speakers, one trained male speaker with a quasi native level of German (B) and two native German amateur actresses (C, S), were recorded individually in a sound-treated room with a dynamic microphone (positioned with a tripod at about 5 cm to the left of their mouth. The speakers were sitting during recording). Each speaker uttered the sentence in an "emotionally neutral" style first. No frame story was given for the neutral expression. For the four emotions, the speakers silently read the description, and when they felt ready, they uttered the two or three sentences listed above for each emotion. They could repeat their expression until they felt satisfied with it. Each speaker needed two attempts to produce each emotional utterance. Both utterances were used in the perception experiment described in section 2.2.

The utterances of the test sentence were digitized at 16 kHz. All the utterances of each individual speaker were recorded and digitized with the same amplification, making energy comparable between the utterances of that speaker. The utterances of different speakers cannot be compared with respect to their overall energy.

2.2. Perception test I

The quality of the recorded natural utterances as recognizable emotional expressions was verified in a perception test. Two natural utterances per speaker and per expression were presented in a forced choice perception test. Each of the 30 stimuli was presented five times. The order in which the stimuli were presented was automatically randomized at the beginning of each session. The stimuli were presented via headphones. After a single presentation of a stimulus, the subjects had to type the first letter of the expression they recognized (choosing among the five categories used during recording), as well as a digit indicating how sure they were about their choice (from 1 to 4, following the logic of the German school mark system that small values represent good marks: 1=sicher / *sure*, 2=recht sicher / *reasonably sure*, 3=eher unsicher / *rather uncertain*, 4=unsicher / *uncertain*). Four native speakers of German (two male, two female, between 25 and 46 years old) participated in the test.

2.3. Recognition of natural utterances

In the first perception test, the natural stimuli were very well recognized (table 1). In particular, "anger", "joy" and "sadness" were well recognized. As could have been expected, the "neutral" category attracted erroneous judgments, probably serving as a default response. There were only very limited confusions between the four emotions, which indicates that neither the corresponding perceptual categories nor the acoustic characteristics overlap strongly between these emotions. The only *symmetrical* confusion above chance level was between "sadness" and "neutral". This might indicate an acoustic similarity between these two categories. "Fear" was not often chosen, and often considered as "neutral". This might indicate less typical acoustic characteristics for "fear" than for the other emotions. One subject actually never chose "fear" as an answer. A possible explanation could be that the recorded type of fear ("scared-children-in-the-dark") differs from the more aroused variant of "panic-in-the-face-of-an-acute-danger" which subjects might have expected.

Table 1. Confusion matrix in perception test I for the natural utterances (percent values; rows: stimuli; columns: answers)

Answer Stimulus	Anger	Joy	Fear	Sadness	Neutral
Anger	85	5	1	-	9
Joy	3	79	3	3	13
Fear	9	7	52	5	27
Sadness	-	5	2	71	22
Neutral	1	-	9	48	43
Total	98	96	67	127	114

The percentage recognition rates for individual stimuli are shown in table 2. Table 3 gives the average certainty values for each utterance and speaker. In terms of judgment certainty, the overall mean certainty value of 1.8 suggests that the subjects found the emotions rather easy to identify.

Table 2. Recognition rates (%) in perception test I for the individual natural utterances. Bold characters indicate utterances chosen as models for copy synthesis.

Correct	Anger		Joy		Fear		Sadness		Neutral	
Speaker B	80	80	100	40	60	65	15 ¹	45	40	25 ²
Speaker S	80	70	90	90	55	15	100	95	50	60
Speaker C	100	100	90	65	75	40	95	75	55	25

¹ judged as "neutral" in 75% of the cases, and thus used as a "neutral" model for copy-synthesis.

² judged as "sad" in 65% of the cases, and thus used as a "sad" model for copy-synthesis.

Table 3. Certainty values in perception test I for the individual natural utterances. Smaller values indicate greater certainty. Bold characters indicate utterances chosen as models for copy synthesis.

Certainty	Anger		Joy		Fear		Sadness		Neutral	
Speaker B	1.9	1.6	1.2	2.1	2.0	2.2	2.1	2.2	2.1	2.1
Speaker S	1.8	1.5	1.6	1.4	2.4	2.2	1.4	1.6	2.0	2.0
Speaker C	1.2	1.2	1.6	2.1	1.6	2.0	1.7	1.8	2.2	1.8

3. Synthetic Stimuli

3.1. Selection of natural models for copy synthesis

The results of the above-mentioned perception test were used to select appropriate natural utterances ("good examples" for the different expressions) as models for the copy synthesis. Since, for most combinations of speaker and expression, high to very high recognition rates were reached for at least one of the two utterances (see table 2)³, it was possible to choose one utterance per speaker and per expression, giving a total of 15 models for the copy synthesis. When both versions of an emotion from a given speaker had the same recognition rate, the one with the better certainty judgment (table 3) was selected.

3.2. Diphone database

For speech synthesis, a diphone database was created with trained speaker B's voice. The diphones were extracted from two different versions of the test sentence "Er ist einfach weggefahren" itself, monotonously spoken at 95 Hz. The diphones were taken

³ The only speaker/expression settings with the recognition rates of both utterances below 50% were speaker B / sad and speaker B / neutral. Curiously, one sad utterance of speaker B was judged as "neutral" in 75%, and one neutral utterance of speaker B was judged as "sad" in 65% of the cases. These two utterances were also chosen as models for the copy synthesis, representing the best-recognized emotion instead of the intended emotion.

from the test sentence in order to avoid uncontrollable influence of concatenating diphones coming from different transsegmental contexts.

3.3. Prosodic parameter measures for copy synthesis

The selected natural model utterances were analyzed manually, using a CSL 4300B workstation. Segment durations and energy as well as F0 targets (local extremes) were determined for the 15 model utterances. The energy of the loudest vowel in an utterance ranged from 57 dB SPL for sadness to 81 dB SPL for anger.

As all utterances were to be synthesized with the male voice of speaker B, the female speakers' F0 had to be reduced to a male speaker's level. This normalization was done in the following way. For each speaker, an overall mean F0 was calculated. The male speaker's mean F0 was taken as a reference, and a normalization factor was calculated for each of the two female speakers with which to multiply all of their respective F0 values in order to bring their mean F0 to the same level. As human pitch perception is logarithmic, such a multiplication preserves the perceived F0 range in semitones.

All natural utterances had to be synthesized using the same limited diphone database. The following phonetic transcription corresponds to the speakers' most frequent realization of the test sentence: [eãstʔaɪnfaxvɁEkÛ´faÛn]. However, mainly two variants were produced: 1) absence of the [t], and 2) in his emotional utterances, speaker B realized [faÛÂ´n] instead of [faÛn]. In order to represent these realizations as closely as possible with the recorded diphone database, the following adaptations were effected: 1) to simulate the absence of a [t], extremely short durations for the [t] were synthesized (around 20 ms of closure and 10 ms of burst), which apparently lead to a forward masking effect by the preceding [s]; 2) to simulate [faÛÂ´n], the segments [a] and [n] were lengthened each by half of the total duration of [Â´].

3.4. Stimulus preparation

The synthetic stimuli were generated using the CPK synthesis system provided by the Center for PersonKommunikation, Aalborg (Jensen *et al.*, 1998), with which it is

possible to do precise duration modeling at the segmental level, and which offers great flexibility in F0 modeling due to LPC resynthesis⁴.

As the CPK synthesis system does not allow detailed energy modeling at present, the energy levels of vowels and [n] in the synthesized utterances were adapted by hand, using the CSL 'scale' function, in order to reduce the maximum segment energy to the natural utterance's level. The energy of the other segments was not controlled. In an informal test, three utterances (S/sad⁵, C/sad, and C/neutral) were considered not loud enough. Therefore, their overall energy was increased until the vowel energy was at about 65 dB SPL.

Because of the risk of introducing artifacts and reducing the sound quality by this coarse energy manipulation, the same synthesized stimuli, without energy manipulation, were used in the perception test, too. Their amplitude was halved, so that their energy was between that of the loud and the soft energy-manipulated stimuli.

It is important to note that when a speech synthesis system does not allow for energy modeling, this does not mean that this parameter is held constant. Time-domain F0 modeling is actually a source of energy variation. To increase (decrease) fundamental frequency, the low-energy, perceptually less important open phase within a period is shortened (lengthened). This changes the relation between the high-energy closed phase and the low energy open phase within each period. For higher F0, this means increased overall energy, while for lower F0, the energy decreases. While this effect might be negligible for neutral speech, it is not for emotional speech, where the F0 range is much higher (in the present study, frequencies typically range from less than 100 Hz to over 220 Hz, which means a reduction of period duration of more than 50%).

However, in natural emotional speech, greater arousal *is* usually expressed simultaneously by higher F0 and higher energy, while less aroused emotions are expressed by lower F0 and lower energy (Scherer, 1996). Thus, as long as this type of arousal-based emotional expression is to be synthesized⁶, the covariance of energy with F0, discussed above, has the same direction as for natural emotional speech.

⁴ From the diphones uttered at 95 Hz, F0 values as high as 340 Hz could be generated.

⁵ i.e. the synthetic stimulus based on speaker S's sad utterance.

⁶ contrarily to a "frequency code" type of voice use (Ohala, 1996), where a deep voice aims at conveying an impression of great size, and thus strength and power. There, low F0 should logically be accompanied by a louder voice.

3.5. Perception test II

Each of the 15 synthetic stimuli was presented in two versions (with and without energy modeling) in a perception test similar to the one with the natural stimuli (section 2.2.). This time, each version was presented only three times. As in the first test, the listeners had to choose the category of the perceived emotion and then give a rating from 1 to 4. However, this time the rating was not about the certainty of their choice, but about their impression of how "convincingly" the emotion was expressed (1=überzeugend / *convincing*, 2=eher okay / *rather OK*, 3=eher schlecht / *rather bad*, 4=total schlecht / *very bad*). The term "convincing" was explained to the listeners with the example that in TV soap operas, expressions are often not convincing; the listeners agreed that this is true. Five native speakers of German participated (2 male, 3 female, between 25 and 35 years old). They had not taken part in the first test and were not accustomed to hearing synthetic speech.

Several listeners mentioned that they did not always find the given answer categories appropriate, and in particular, that they missed a category "disappointment".

3.6. Recognition of synthetic stimuli

Globally, the synthetic emotional stimuli were recognized worse than the natural utterances, but still above chance level (table 4). Energy modeling was rather unimportant for recognition: There were only very small differences between the confusion matrices for stimuli with and without energy modeling. For this reason, only the overall confusion matrix is shown.

Globally, "joy" and "fear" are very unattractive categories, while "anger", "sadness" and "neutral" were frequently chosen.

As for the natural utterances, "fear" is the least identified among the four emotions. "Sadness" and "neutral" are mutually confused.

Contrary to the natural stimuli, "anger" was interpreted as "sadness", a confusion that virtually never occurs for natural stimuli (Banse & Scherer, 1996; Leinonen *et al.*, 1997). "Joy" was interpreted as "anger", which is also rare for natural stimuli, although confusions between elation (i.e. the joy type portrayed here) and hot anger do occur (Banse & Scherer, 1996). Finally, "fear" stimuli were interpreted as "sadness" more often than as "fear", which is also very untypical.

Table 4. Overall confusion matrix in perception test II for the synthetic stimuli (percent values)

Answer Stimulus	Anger	Joy	Fear	Sadness	Neutral
Anger	39	6	2	42	11
Joy	32	33	4	10	20
Fear	12	6	24	28	30
Sadness	14	-	3	43	39
Neutral	17	4	3	28	48
Total	114	49	36	151	148

More interesting than the global confusions, however, are the recognition rates for the individual stimuli (tables 5-7), because these show differences that allow some conclusions with respect to the fundamental question that had motivated this study.

A first remarkable result is that for "anger" and "joy", stimuli exist that reach recognition rates above 80%, not far from the original natural utterances' recognition rates (C/anger, S/joy). At the same time, the other stimuli in these categories (B/anger, S/anger, B/joy, C/joy) are not recognized, although the recognition rates of the originals were all above 80%. Apparently, by modeling segment duration and intonation, highly relevant cues were retained in C/anger and S/joy, but not in the other anger and joy stimuli. This suggests that for the natural counterparts of the unrecognized synthetic stimuli, unmodeled parameters like voice quality were crucial for the communication of emotion. These results are in line with the existence of different speaker strategies frequently mentioned in the literature about the vocal expression of emotions (e.g., Banse & Scherer, 1996) which in the present case do or do not allow communication of "anger" and "joy" via duration and intonation alone.

Table 5. Recognition rates (%) in perception test II for the individual synthetic stimuli with energy modeling (+en), without energy modeling (-en). The highest recognition rates in each category are printed in bold.

Correct	Anger		Joy		Fear		Sadness		Neutral	
	+en	-en	+en	-en	+en	-en	+en	-en	+en	-en
Speaker B	7	7	13	20	13	20	47	67	33	47
Speaker S	33	20	73	87	13	13	40	60	67	47
Speaker C	87	80	7	0	40	47	33	13	40	53

Table 6. Most frequently chosen category (%) in perception test II for the individual synthetic stimuli. The letters indicate the most frequent answer for each stimulus: a=anger, j=joy, f=fear, s=sadness, n=neutral. The given values are the frequency of that answer. If the correct answer was most frequent, no letter follows the percentage, which is printed in bold.

Stimuli	Anger		Joy		Fear		Sadness		Neutral	
	+en	-en	+en	-en	+en	-en	+en	-en	+en	-en
Speaker B	67s	47s	47a	67a	40n	33n	47	67	40a	47
Speaker S	60s	73s	73	87	60s	40s	40	60	67	47
Speaker C	87	80	53a	47n	40	47	53n	67n	40s	53

Table 7. "Convincing" judgments in perception test II for the individual synthetic stimuli. Smaller values mean "more convincing" (following the logic of the German school mark system).

Convincing	Anger		Joy		Fear		Sadness		Neutral	
	+en	-en	+en	-en	+en	-en	+en	-en	+en	-en
Speaker B	2.7	2.8	2.3	2.4	2.8	2.9	2.3	2.3	2.6	2.5
Speaker S	2.7	2.3	2.3	2.5	3.0	2.9	2.6	2.2	2.6	2.5
Speaker C	2.1	2.3	2.9	2.3	2.5	2.5	2.3	2.1	2.3	2.3

The most frequent answer given for each stimulus (table 6) is interestingly systematic. All poorly recognized "anger" stimuli are interpreted as "sadness", while "joy" stimuli are most often misinterpreted as "anger"; "fear" is taken for "sadness" or "neutral"; and "sadness" and "neutral" stimuli are mutually confused. Only in three cases (B/neutral, C/joy and C/neutral) does the presence of energy modeling change the most frequently chosen category.

For the "convincing" judgments, the overall mean value of 2.5, which is precisely the middle between "convincing" and "not convincing" marks, suggests that subjects found the stimuli generally neither very convincing nor very bad.

3.7. Perception test III

During perception test II, several points indicated that it was necessary to clarify whether the copy synthesis process created new perceptual categories, different from the categories originally intended by the speakers: On the one hand, listeners remarked that they missed a response category "disappointment"; on the other hand, untypical confusions occurred among categories (see 4.2).

For these reasons, a perception test using free association was conducted. Listeners were asked to describe the expression conveyed by each stimulus in their own words. As in perception test II, each answer was followed by an evaluation (on a

scale from 1 to 4) how "convincing" the stimulus was as an example of the described expression. Only the 15 synthesized stimuli without energy manipulation from perception test II were used. Contrary to the preceding tests, each stimulus was presented only once, but could be listened to several times. Eight native German listeners (2 male, 6 female, 25-35 years old) took part in the test. They had not taken part in the preceding tests and did not know the stimuli. Before the test, they were not informed about the originally intended emotion categories.

The answers were grouped into classes by the author, who tried to reuse the original categories: e.g. the answer "leicht verärgert" ("*slightly angry*") was counted as "*anger*", "amüsiert" ("*amused*") as "*joy*" etc. For frequent answers that could not easily be subsumed under one of the 5 original categories, new categories were created.

3.8. *New perceptual categories*

The results of perception test III show that new perceptual categories had indeed appeared. Listeners had a strong preference for the answer "disappointment" (table 8). In fact, each of the 15 stimuli was evaluated as "disappointment" at least once.

Table 8. Relative frequency (%) of given answers in perception test III, grouped into categories. New categories: Disappointment, Surprise, Resignation, Lack of Understanding.

Anger	Joy	Fear	Sadness	Neutral	Disapp	Surprise	Resign	L.Und.	others
13	6	2	14	10	31	4	6	6	8

This tendency towards the interpretation of stimuli as "disappointment" was strongest for stimuli B/fear, S/fear and S/sadness (table 9). Another interesting observation is that the best recognized stimuli in perception test II were most frequently evaluated as the intended category: C/anger and S/joy. In the same way, two out of three neutrally intended stimuli were evaluated as being neutral. Chance level was 10%.

Table 9. Most frequent answer (%) for each synthetic stimulus in perception test III. Stimuli recognized as intended are printed in bold

Convincing	Anger	Joy	Fear	Sadness	Neutral
Speaker B	38 disappoint	38 disappoint	50 disappoint	38 anger	25 resignation
Speaker S	38 disappoint	38 joy	50 disappoint	50 disappoint	50 neutral
Speaker C	63 anger	38 disappoint	38 sadness	38 disappoint	38 neutral

The mean "convincing" mark of 2.2 was better than in perception test II; in particular, the stimuli originating from speaker B were judged as more convincing.

Table 10. "Convincing" judgments for each synthetic stimulus in perception test III. Smaller values mean "more convincing" (following the logic of the German school mark system)

Convincing	Anger	Joy	Fear	Sadness	Neutral
Speaker B	2.0	1.8	1.9	2.0	2.4
Speaker S	2.1	2.0	2.1	2.9	2.9
Speaker C	1.6	2.1	2.0	2.4	2.1

Manifestly, all synthetic stimuli share cues leading to a tendency towards the perception of disappointment. These cues are only partly modified by the modelled prosodic parameters. One of these cues might be the text of the test sentence, which was felt as unappropriate for joy by listeners in all three perception tests. Another possible influence comes from speaker properties in the diphone database, in particular the (breathy) voice quality. Also, influences stemming from the synthesis method cannot be excluded.

4. Conclusion and outlook

It has been shown that synthetic stimuli do exist for which only F0 and duration were modeled and which can be identified reliably as the intended emotion. These stimuli are judged relatively "convincing". On the other hand, the recognition rate of natural utterances that serve as models for copy synthesis is not a reliable predictor for the recognizability of emotions in the synthesized utterances. Different speaker strategies for expressing an emotion seem to involve duration and intonation to a highly varying extent.

All synthetic stimuli showed a tendency to be perceived as "disappointment". Apparently, factors varying only to a limited extent across stimuli were responsible for

this. In particular, it is possible that the diphone voice quality had an important influence; this should be the subject of a subsequent study.

Energy modeling, at least in the form in which it has been applied here, seems to be a nearly negligible factor for modeling emotions in synthetic speech. The reason for this might be, as Holmberg *et al.* (1988) have shown, that in natural speech, changes in loudness are accompanied by changes in voice quality, in the sense that when a voice becomes louder, it also becomes harsher. In our synthetic stimuli, soft voice quality, as a cue to low "speaker effort", might have annulled the effect of higher energy in subjects' interpretation of the stimuli.

The consequences of these results for developing a strategy for the implementation of emotions in a concatenative synthesizer might be the following: It seems possible to express emotions via the limited parameter set of duration and intonation, as the recognition accuracy of over 80% for certain stimuli shows. However, special care has to be taken to focus on modeling a special subset of speaker strategies that involve voice quality as little as possible. Modeling energy in the way it has been done here does not seem to make much sense. It would be interesting, however, to try a more precisely controlled energy modeling in combination with voice quality cues to speaker effort. In order not only to increase recognition but also to express emotions in a "convincing" way, it is probably necessary to control voice quality.

More work should be done to be able to further interpret the results of the present study. For example, the perception test results should be put in relation to the acoustic parameters. Interpreting that relationship can lead to hypotheses about how to efficiently express emotions in synthetic speech, which could serve as guidelines for future studies.

Acknowledgments

The author is particularly grateful to Jürgen Trouvain for very valuable support and stimulating discussions. Thanks also to Jaques Koreman for his feedback, as well as to the speech synthesis group at the Center for PersonKommunikation, Aalborg, especially Claus Nielsen, for providing the CPK synthesis system.

References

- Banse, R., & Scherer, K. R. (1996). Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology*, 170 (3), p. 614-636.
- Edgington, M. (1997). Investigating the limitations of concatenative synthesis, *Eurospeech '97*, Rhodes.
- Heuft, B., Portele, T., & Rauth, M. (1996). Emotions in time domain synthesis, *ICSLP '96*, Philadelphia.
- Holmberg, E. B., Hillman, R. E., & Perkell, J. S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, *Journal of the Acoustic Society of America*, 84(2), p. 511-529.
- Jensen, J., Nielsen, C., Andersen, O., Hansen, E., & Dyhr, N.-J. (1998). A speech synthesizer with modeling of the Danish "stød". *Proc. IEEE Nordic Signal Processing Symposium (Norsig '98)*, p. 121-124.
- Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M. (1997). Expression of emotional-motivational connotations with a one-word utterance, *Journal of the Acoustic Society of America*, 102 (3), p. 1853-1863.
- Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., & Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS, *ICSLP '98*, Sydney, Vol. 3, p. 923-926.
- Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustic Society of America*, 93 (2), p. 1097-1108.
- Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice, *ICSLP 96*.
- Rank, E., & Pirker, H. (1998). Generating emotional speech with a concatenative synthesizer, *ICSLP '98*, Sydney, Vol. 3, p. 671-674.
- Scherer, K. R. (1986). Vocal Affect Expression: A Review and a Model for Future Research, *Psychological Bulletin*, 99, p. 143-165.
- Scherer, K. R. (1996). Adding the affective dimension: A new look in Speech Analysis and Synthesis, *ICSLP 96*.
- Vroomen, J., Collier, R., & Mozziconacci, S. (1993). Duration and intonation in emotional speech, *Eurospeech '93*, Berlin, Vol. 1, p. 577-580.