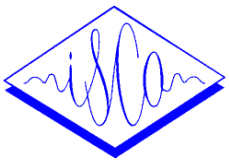

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

EXPERIMENTS WITH EMOTIVE SPEECH - ACTED UTTERANCES AND SYNTHESIZED REPLICAS

*Rolf Carlson, Björn Granström and Lennart Nord**

Dept of Speech Communication & Music Acoustics, Royal Institute of Technology, KTH,
Box 70014, S-10044 Stockholm, Sweden. / Phone 46 8 7907847, Fax 46 8 7907854
*names in alphabetic order

ABSTRACT

We will report on our current work to model extralinguistic features in speech. As a starting point we analysed sentences representing different emotions produced by actors. In another project, this material had been evaluated by a listener panel. Sentences with successfully acted emotions were selected for further analysis. The analysis corroborated earlier findings concerning speech tempo and fundamental frequency contours. We also found differences in segmental phonetic realizations, partially correlated with speaker efforts, such as energetic angry and restrained, sad speaking styles. The parametric representation of the analysis was simplified to conform with our rule based text-to-speech system. Several manipulated synthetic versions were compared and evaluated for perceived emotions. The experiment showed that emotions are signalled through a complex interaction of segmental and prosodic cues. In addition to the normally used parameters like fundamental frequency dynamics and level, speaking rate and segmental realization we also observed supporting non-phonetic sounds that enhanced the perceived emotional quality. We will in our presentation give example of the synthetic stimuli and discuss potential problems in the modelling of emotional speech in the present framework.

10.21437/ICSLP.1992-220

I. INTRODUCTION

A number of studies have reported on the acoustics of emotional speech. See for example [1]. Apart from pitch patterns and segmental durations, other kinds of signal properties have been identified as carrying information about emotions, such as intensity variations and certain coarticulatory variations. Absolute values in pitch, rate and loudness as well as ranges and variation properties are used by speakers.

Speech synthesis has for decades been used as the tool to study the significance of the acoustic description of speech. Originally, our main concern has been the intelligibility aspects of speech, with a concentration on segmental and prosodic quality. As our synthesis models develop and the general quality of the synthesis improves, the possibility to model more subtle aspects of speech increases. Also, the expanded use of speech synthesis devices in diverse applications has made the demand for greater flexibility more pronounced. The use of speech synthesis as a speech prosthesis is a typical application, where, ideally, the full flexibility of natural speakers could be used. In such applications the possibility in speech to convey attitudes and emotions is important.

Until recently, the accumulated knowledge has not been used in speech synthesis. Murray, Arnott, & Newell, [2], report on preliminary experiments with DECtalk to create six different emotions by modifying both global settings, like speech rate, and average pitch and emotion specific "prosodic rules" for primarily segment duration and pitch contours. Cahn, [3], has reported on similar experiments, also using DECtalk. She has created a so-called "affect editor", where combinations of user-accessible parameters are used to create different speaker affects or attitudes. Applications in speaking prostheses are envisaged, but the general need in a wider context is also observed.

The experiments using DECtalk show that even commercially available synthesizers to some extent lend themselves to experiments with style variations. Parametric synthesizers, like the one in DECtalk, have considerable flexibility compared to, e.g., a standard LPC based diphone system. In some of the products based on formant synthesizers, voice variation is one feature. In such systems there are several possibilities to vary the speech. On the global level there are possibilities to vary speech tempo, amplitude, and voice parameters like amount of aspiration, mean pitch and pitch dynamics, vocal tract length, general speed of articulation, etc. Some of these parameters have been combined to define different voices. The simulation of some voices are, however, still not very convincing. In a research tool more flexibility is requested, both from the production model itself and the detailed control of the associated parameters. Especially, the voice source model has been improved in several systems, [4, 5, 6].

In the rule components of the KTH system, [7], all sorts of effects could be modelled, i.e., as long as they are rule governed. The conditioning factors for these rules could either be given as analysis by the system, such as syntax or some measure of predictability of words, as commands to the system, or it could be given as extra information in the input text string. One example of the latter is the emphasis control in the present system. By adding a number before a word in the phonetic string we can over-rule such things as default sentence stress assignment and function word reduction and also force different degrees of emphasis.

The present investigation is a small study where we restrict ourselves to only a few "emotion labels". We are evaluating an experimental technique that is partly based on an automatic formant extraction software, [8], a pitch extraction program, and a source matching procedure. These program facilities help us to generate parameter tracks for the synthesis samples that are further manipulated and mixed with each other.

This methodology is to be regarded as a complement to a method used in a previous study, [9]. In that study, synthesis parameters were manipulated by the subjects, who were also requested to verbally describe the resulting emotions in synthesized sentences.

II. SPEECH MATERIAL

In this study we used a sentence material that had been recorded for another project and evaluated by different listener groups by Öster and Risberg, [10]. They investigated whether there were any differences between hard-of-hearing and normal hearing subjects in their ability to distinguish between emotions signalled by the speech wave. The advantage of using this speech material was that we could benefit from the perceptual evaluation already performed. These sentences were simple statements that were read by two actors with underlying emotional content. The original material consisted of six sentences, each read with six emotions. The credibility of the readings

was evaluated by an adult listening panel with 23 listeners. In the study, the confusions of the individual sentences were analysed. We could hence chose sentences that showed a minimum of confusions for the basic emotions "happy, sad, angry" and "neutral" (one error or less, if "happy" and "positive" are equated). Such a sentence: "Sommarlovet börjar sent i år" (The summer vacation starts late this year) was analysed. Figure 1, shows two spectrograms with the emotions "happy" and "sad". Several differences can be seen, primarily in the use of duration and fundamental frequency. Note the slower tempo for the sad utterance with the prolonged /s/ and /t/ in /se:nt/. Also the pitch movement is monotonous in the sad utterance and high and more dynamic in the happy sentence. Some of these differences were reported in [10].

Duration measurements on the material from the male subject showed in general that the sad utterances were produced at lower rate, while the happy and angry utterances were produced with more energy and at slightly faster rates. The neutral statements were the most rapid ones, a partial reason being that there were typically no emphasized words in these utterances.

The durations for the four versions of our test sentence, read by the male speaker, were found to be, for the neutral reading 2030 ms, for the happy emotion 2650 ms, for the sad 3070 ms and for the angry emotion 2740 ms.

The mean pitch values for the four sentences were: neutral 139 Hz, happy 154 Hz, sad 131 Hz and angry 148 Hz. Moreover, the sad utterance was rather monotonous, while the happy and angry showed a more dynamic pitch contour.

Several phonetic differences were observed. The angry and to some extent the happy utterances showed a more energetic articulation, with more abrupt vowel onsets, trilled /r/ allophone, as opposed to vowel-like /r/.

This material contains extra-linguistic signals such as sighs, breathing sounds, lip smacks, etc. were also used by our speakers, however, those sounds were not systematically investigated.

III. ANALYSIS BY SYNTHESIS

In an effort to distinguish the different acoustic factors that the actors made use of to convey emotions, we have used speech synthesis. In our first attempts we have concentrated on one male speaker and phonetic variations rather than extralinguistic sounds like sighs, laughter etc. Voice breaks and other changes in the voice source were also interesting, especially as our speech synthesis voice source is flexible and should account for part of this variation.

The synthesized replicas were produced by using an analysis program that automatically extracted formants, [8]. Some hand-corrections of formant traces were needed, especially for the lively voice samples, that is, the sentences with strong emotional content. As part of the synthesis software, the parameter tracks can be superimposed on spectrograms of real speech and matchings can be made. An analysis program, was used to derive the pitch contours. Matchings of voice source parameters were also tried in an attempt to improve the synthesis. Examples of sentences synthesized this way will be presented at the conference

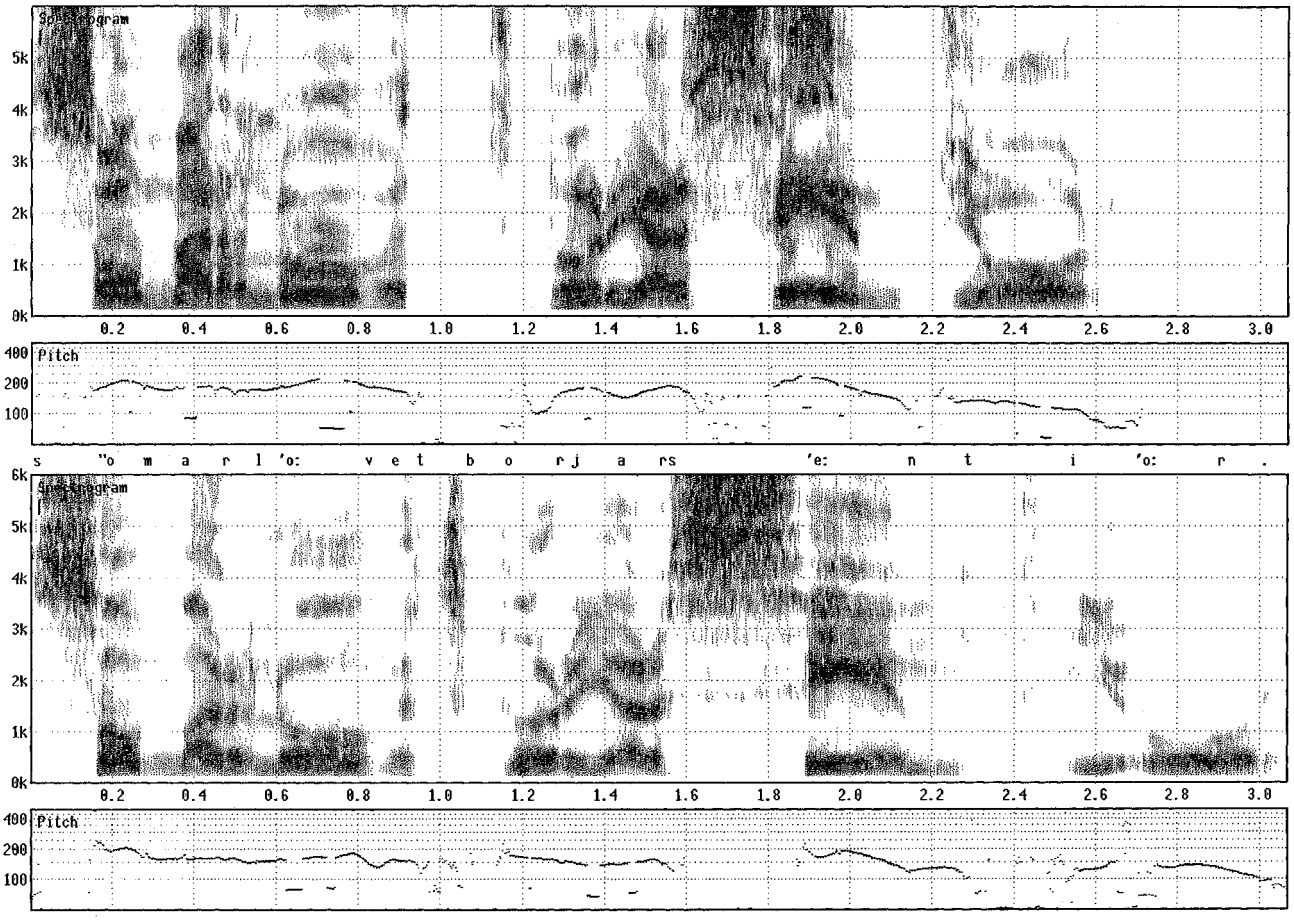

Figure 1. Spectrograms for two emotions acted by a male speaker, reading the sentence "Sommarlovet börjar sent i år" /s"omarl'o:vet børjar s'e:nt i o:r/. Top: happy voice, Bottom: sad voice. In the pitch plot, horisontal dotted lines are 50 Hz apart starting at 100 Hz

Including the flexible voice source, the "LF-model", described in [5], in the parametric analysis, was not a straightforward task, when we tried to match local variations in the sentences with emotional qualities. The reason for the difficulties was that much of the spectral mismatch between the real speech and the synthesis was due to factors other than the voice source quality, such as local appearances of zeroes, addition of noise, etc. Therefore, our use of the voice source mainly served the purpose of creating a general perceptual impression of a somewhat aperiodic voicing at certain places. A difference in the impression of the natural and the synthetic voice was that the synthetic voice phonates with fewer disturbances, and with weaker low-frequency energy.

After having produced replicas of four utterances, we also made new synthesized versions of sentences by mixing the acoustic elements. The original four sentences: neutral, happy, angry and sad, were mixed with the pitch contour of the others. To do this, time alignments were first made of the sentences, by making adjustments of the segmental durations. By the stretching and the compression of the parameter tracks, some disturbance of the phonetic quality was introduced. We decided, however, not to pay any attention to these effects, as they were considered small. Part of our interest was to study how segmental articulation affected the perceived emotion. Our philosophy was thus, that we did not want to change the articulations of the sentences and still be able to overlay a pitch contour. Especially, stretching the neutral sentence to match the sad sentence, which was of longer duration than the rest of the sentences made some segment alignments quite drastic.

IV. LISTENING TESTS

4.1 Experimental design

Stimuli set: According to the principles described in the previous section a set of fifteen stimuli was produced. The basic happy, sad and angry utterances were synthesized (HH, SS, AA). The neutral sentence was time aligned to the happy, sad and angry sentence (Nh, Ns and Na). The next group of stimuli consist of the sad, angry and neutral sentences time aligned to the happy utterance and incorporating the happy pitch contour (SH, AH and NH). In the same fashion two more groups were produced, based on the sad and the angry sentence (HS, AS, NS and HA, SA, NA). A test list of fifty items was produced. It consisted of the fifteen stimuli randomized three times complemented with five dummy samples.

The listening test was designed as a forced choice test, where eighteen subjects listened to the synthetic stimuli played from a tape and had to give one response for each stimulus sentence, presented twice. The emotion labels that were allowed were: "neutral, angry, happy, sad". Listeners were also asked to put down a number between one and ten for each stimulus, indicating how clearly the perceived emotion was signalled. Most of the listeners had previously been exposed to synthetic speech.

4.2 Results

Figure 2 gives an overview of which emotions were signalled. by the copied and manipulated synthesized sentences.

Two of the basic emotions (AA, SS) were quite convincingly conveyed, sad 95% and angry 80%; while the response "happy" was not considered very strong, (HH 42%), 25% being chance level.

The group of sentences with a neutral spectral trace, aligned to the happy, the sad and the angry utterances, (Nh, Ns and Na) were consistently regarded as neutral utterances (ca 80%). Just manipulating the durations did not appear to affect the emotional categorization. There seem to be a weak tendency for the stretched sentences to give sad impressions (10-15%) for these stimuli. The original "sad" sentence is actually the longest utterance in our material. Sadness is often described as related to a slow speech tempo. Given the general spread of the sad responses there is an alternative explanation: the general quality of the synthesis could affect the sad response. In support of this view we can observe that "Na" is the least sad stimuli in the whole set.

The following group of stimuli in Fig 2, with the happy pitch contour superimposed on the spectral specifications of sad, angry and neutral sentences, aligned with the happy sentence. show predominant happy responses only for the version based on the neutral utterance (NH). For the two other versions, SH and AH, the angry responses are more numerous. Observe that the original happy synthesis (HH) was the least convincing of the three emotions, also often confused with anger. A possible explanation for this could be the general excited character of both anger and happiness, at least as produced by our talker.

The next group of stimuli, with the sad pitch contour superimposed on the happy, angry and neutral parameter tracks, (HS, AS and NS) gave generally high sad responses. The only exception is the version based on the angry utterance with 63% sad and 24% angry responses. As discussed below the explanation could be that the angry utterance had a characteristic energetic articulation.

The last triplet, where the angry pitch contour was superimposed on the parameter tracks for the happy, sad and neutral sentences

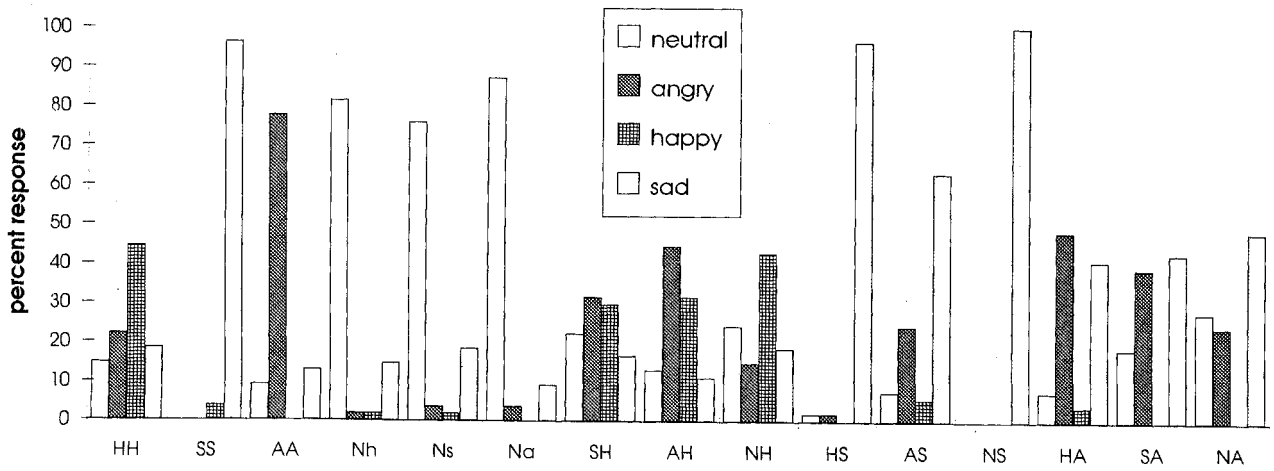

Figure 2. Results from emotional categorization in a listening test using synthetic sentences with varied emotional content. The first letter in the stimulus label indicates the original sentence (H= happy, S= sad, A= angry, N= neutral). The second letter indicates which sentence has contributed the duration and pitch contour in the modified sentences. Lower case indicates modification of duration only. For further explanations of the stimuli labels, see the text.

(HA, SA and NA) gave as many sad responses as the expected angry responses. Note the difference between the original angry synthesis and the mixed stimuli with other segmental content and angry pitch contour, which gave a smaller number of angry responses. Typical features of the original angry synthesis include faster onsets of voiced segments, and a strong, very close articulated /e:/ in the stressed word /se:nt/ ("late"). These phonetic details are not part of the mixed stimuli that have the other sentences as a starting point with only the angry pitch contour from the angry utterance.

Summarizing, it can be said that the emotion for the mixed stimuli, with the timing and pitch taken from other sentences, also were perceived as having the emotion of the sentence that contributed with the pitch contour and segmental durations. Especially, the sad emotion was successfully conveyed by stimuli HS and AS, that is, sentences that originated from a happy and an angry utterance. Also the angry emotion was strongly signalled by the HA and the SA stimuli, that is, stimuli originating from the happy and the sad utterances. In this last group of stimuli an unforseen tendency occurred, namely that all three stimuli, HA, SA and NA often were labelled as sad. A detailed analysis of the listener responses shows this effect clearly. Some subjects answered sad more often than any other emotion, as if there was an inbuilt quality of sadness in many of the synthetic stimuli.

What is not given in the figure is the degree of naturalness for the sentences, an emotion might be signalled clearly, although the phonetic quality or even the prosody is bad.

Regarding the perceived strength of the emotion in the stimuli, not all subjects gave numbers for all the stimuli. We will therefore not attempt a full analysis of the ratings. Some subjects put high values on very few stimuli, in general those that unanimously were perceived as sad or angry. However, a preliminary evaluation showed a correlation between the ratings and the mean percent perceived emotion category.

V. DISCUSSION

The amount of interaction between the emotive speech and the linguistic content of a sentence is difficult to ascertain, but has to be taken into account. It is not easy to define a speech corpus that is neutral in the sense that any emotion could be used on the sentences. Also some sex related differences might be observed. In the study by Öster & Risberg, [10], they reported that female joy and fear were more easily confused than for the male voice, where instead joy and anger were more often confused by young listener groups. Also concepts like joy, anger etc. can be expressed very differently and a unique perceptual - acoustic mapping is probably not possible. For a discussion of these concepts, see [11]. By introducing active/aggressive as opposed to passive forms of emotions it is evident why researchers sometimes report seemingly different results when testing the signalling of emotions.

In our material there were some biases towards sad and to some extent angry emotions, judging from the listening results. The male speaker often used an articulate, rather stern tone of voice which might have favoured the given responses. Also this particular sentence might often be interpreted as a piece of bad news and should accordingly seldom be judged as having a happy tone of voice. Nevertheless, the results reported by Öster & Risberg, [10], did not point in that direction.

Note that the voice does not always give away the complete speaker attitude. It is often observed that misinterpretation of emotions occurs if the listener is perceiving the speech signal without reference to visual cues. Depending on the situational context it is thus easy to confuse anger with joy, fright with sorrow, etc.

One problem when comparing emotional sentences is that many sentences contain extra factors such as sighs, voice breaks and jitter, lip smacks, etc., which often contribute in a decisive way to the intended emotion. This means that a standard acoustic analysis of produced sentences with different emotional content, in terms of e.g. duration, intensity and pitch, does not discriminate between emotions, if the speaker relies heavily on non-phonetic cues in the production.

VI. FINAL REMARKS

In this contribution we have discussed the analysis of emotions in speech and investigated the implementation of these emotions in speech synthesis. Although we did not succeed in fully mastering the different emotions that we set out to investigate, we still consider the present results promising for a continuation of this work. It is apparent from our experiment that emotions are signalled through a complex interaction of segmental and prosodic cues.

We will continue this project by comparing the present speech material with other speakers' production of the same emotions. Generalizations from these studies will be expressed as production rules and evaluated within the framework of the text-to-speech system.

Several applications can be foreseen, e.g. synthesis as a speaking prosthesis where the user is able to adjust speaker characteristics and emotional content or in translating telephony, where speaker identity ought to be preserved and tone of voice aspects also form part of the communication.

ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish Language Technology Programme and the Swedish Telecom.

References

[1] Williams, C. E. & Stevens, K. N. (1972): "Emotions and speech: some acoustical correlates", *JASA* vol. 52, pp. 1238-1250.

[2] Murray, I.R., Arnott, J.L., & Newell, A.F. (1988): "Hamlet - simulating emotions in synthetic speech," pp. 1217-1223 in (W.A. Ainsworth & J.N. Holmes, eds.) Proc. *SPEECH '88*, Book 4 (7th FASE Symp.), Edinburgh.

[3] Cahn, J. E. (1990): "The generation of affect in synthesized speech", *J of the American Voice I/O Society*, vol. 8, pp. 1-19.

[4] Carlson, R., Granström, B. & Karlsson, I. (1990): "Experiments with voice modelling in speech synthesis", in Laver, J., Jack, M. & Gardiner, A. (eds.), *ESCA Workshop on Speaker Characterization in Speech Technology*, pp. 28-39, CSTR, Edinburgh.

[5] Fant, G., Liljencrants, J & Lin, Q. (1985): "A four-parameter model of glottal flow", *STL-QPSR* 4/1985.

[6] Klatt, D. & Klatt, L. (1990): "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am.*, vol. 87(2), pp. 820-857.

[7] Carlson, R., Granström, B. & Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed), *Advances in speech, hearing and language processing*, JAI Press, London

[8] Carlson, R. & Glass, J. (1992): "Vowel classification based on analysis by synthesis", in this proceeding.

[9] Granström, B., & Nord, L. (1991): "Ways of exploring speaker characteristics and speaker styles," in *Proc. of the XIIth ICPhsS, Aix-en-Provence*, Vol. 4., pp. 278-281.

[10] Öster, A-M. & Risberg, A. (1986): "The identification of the mood of a speaker by hearing impaired listeners", *STL-QPSR* 4/1986, pp. 79-90.

[11] Scherer, K. (1989): "Vocal correlates of emotion", in (Wagner, H. & Manstead, T., eds.), *Handbook of Psychophysiology: Emotion and Social Behavior*, pp. 165-197. Chichester: Wiley.