WAYS OF EXPLORING SPEAKER CHARACTERISTICS AND SPEAKING STYLES

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

In the exploration of speaking style and speaker variability we make use of a multi-speaker database and of a speech production model. A recent version of this model includes a variable voice source and a more complex modelling of the vocal tract. Systematic variation in speech synthesis has been used as a tool to explore possible style and speaker dimensions. Preliminary listening experiments have been carried out with the aim to investigate whether it is possible to describe different synthesis samples according to different attitudinal and emotional dimensions.

1. INTRODUCTION

An increasing amount of knowledge concerning the detailed acoustic specification of speaking styles and of speaker variability is presently accumulating. The ultimate test of our descriptions is our ability to successfully synthesize such voices [1]. A better understanding will also have an impact on several applications in speech technology. A systematic account of speech variability helps in creating speaker adaptable speech understanding systems and more flexible synthesis schemes.

Why introduce emotional content in speech synthesis? Firstly, to increase naturalness and intelligibility of a spoken text. Speaking style variation and the speaker's attitude to the spoken message are also important aspects to include. However, if the attitude can not be convincingly signalled, it is better to stick to a more neutral, even non-personal machine-like synthesis.

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.

2. NEW TOOLS

In the exploration of speaking style and speaker variability we make use of a multi-speaker database. In our speech database project we have started to collect material from a variety of speakers, including professional as well as untrained speakers [5]. The structure of the database makes it possible to extract relevant information by simple search procedures. It is thus easy to retrieve information on the acoustic realization of a linguistic unit in a specified context. Earlier studies have concentrated on linguistic structures rather than paralinguistic descriptions.

We aim at explicit descriptions that are possible to test in the framework of a text-to-speech system [3]. A recent version of the speech production model of the synthesis system includes a variable voice source and a more complex modelling of the vocal tract [4]. This synthesis model gives us new possibilities to model both different speakers and speaking styles in finer detail. The necessary background knowledge, however, is in many respects rudimentary. We will here show one example of analysis-synthesis applied on emotive speech.

3. ACOUSTICS OF EMOTIONS

In acoustic phonetic research most studies deal with function and realization of linguistic elements. With a few exceptions, e.g. [7,8], the acoustics of emotions have not been extensively studied. Rather, studies have dealt with the task of identifying extralinguistic dimensions qualitatively and sometimes also quantify these by using e.g. scaling methods. Spontaneous speech has been used as well as read speech with simulated emotional expressions. Judgements have been made by the researchers' ear and also by a variety of listening tests, using untrained and trained listener groups.

An interesting alternative is to ask the listener to adjust presented stimuli to some internal reference, such as joy, anger etc. This is typically done by using synthetic speech, which cannot be too poor in quality if emotions should be conveyed. Recent experiments using DECtalk has been reported by Cahn [2].

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 a study by Öster & Risberg [6], female joy and fear were more easily confused than for male voices, 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.

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 perceiveing the speech signal without reference to visual cues. Depending on the contextual references it is thus easy to confuse anger with joy, fright with sorrow, etc.

4. SPEECH ANALYSIS

We have analysed readings by two actors who were portraying different emotions by reading a fixed set of sentences in different ways: with anger, joy, fear, sadness, surprise and also in a neutral tone of voice. This material has already been used by Öster in the investigation referred to above [6], with the aim of investigating the possible differences in ability to perceive emotion acoustically, as shown by two listener groups, young hard-of-hearing subjects and young normal hearing subjects.

We specifically analysed pitch, duration and segmental qualities and also made synthetic matchings of a number of these sentences trying to extract the relative importance of the different acoustic cues.

One example from the database can be seen in Figure 1, where two versions of the Swedish sentence "De kommer på torsdag" (They will arrive on Thursday) pronounced by a male actor in an angry and a joyful mode are shown. Numerous differences can be observed. For this particular "angry" utterance the pitch is lower and more even than the "happy" utterance. The voicing is also stronger and somewhat irregular especially in the first vowel (probably the false vocal cords are also involved).

For some of the sentences it was obvious that the two actors made use of a number of extra factors such as sighs, voice breaks and jitter, lip smacks, etc, which often contributed 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.

As a point of reference we have also initiated a small study on spontaneous speech from radio interviews. This speech often contains passages that are extremely compressed or expanded. These effects are difficult to make use of in speech synthesis applications. Nevertheless, it is a good reminder of just how diverse and flexible the speech signal appears in real-life communication.

5. VALIDATION BY SYNTHESIS

Different analysis-by-synthesis techniques show great promise in deriving data for the synthesis of different voices, styles and emotions. Specifically, we investigated an interactive production paradigm. We asked subjects to sit at a computer terminal and change the horizontal (X) and vertical (Y) position of a point within a square on the screen by means of a mouse. The X and Y values can be used in a set of synthesis rules,

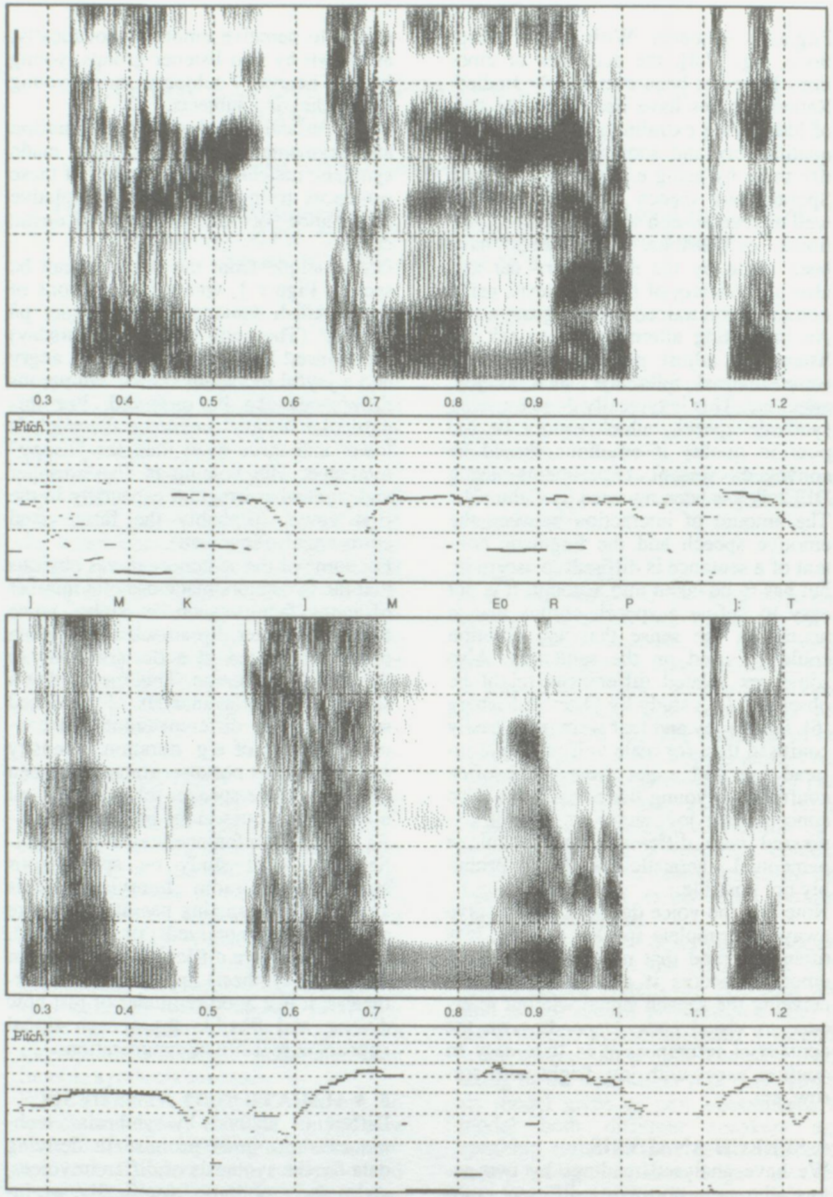

Figure 1: Spectrograms for two emotions imitated by an actor reading the sentence "De kommer på torsdag." /dɔm 'kɔmər po 'tu:ŝda/. Only the underlined part is displayed Top: angry voice, Bottom: happy voice. In the pitch plot, horisontal dotted lines are 50 Hz apart starting at 100 Hz.

changing e.g. different aspects of the voice. In this way we set up a number of rules that changed e.g. pitch deviations, intensity dynamics or voice source parameters of a synthesized sentence. The subjects were asked to try different combinations of these parameters by moving the mouse and reporting on the impression that the synthetic sentence made on them in terms of e.g. emotional content.
In Figure 2 a result from such an experiment is shown. The X dimension corresponds to the slope of the declination line where a low coordinate value, (left), corresponds to a rising contour and a high value, (right), corresponds to a falling contour, with the midpoint in the sentence kept at a constant pitch value. The Y dimension is the pitch dynamics, where the low end corresponds to small pitch movements and the top to larger local pitch excursions. The tested sentence is the same as in Figure 1, i.e. a linguistically quite neutral statement. Obviously, the variations suggest several different attitudes to our listeners. The task appeared quite managable to the subjects, who responded with a fair degree of consistency. We are pursuing this line of experiments further including also voice source variations.

voice break	happy optimistic		content self-assertive
worried	neutral	sure	determined
threatening	caution	disappointed angry	
unnatural	indifferent		compliant

Figure 2. Example of free verbal responses in a speech synthesis production experiment with four subjects. See text for tested dimensions.

6. FINAL REMARKS

In this contribution we have indicated some ways of exploring speaker characteristics and speaking style using the speech database and synthesis environment at KTH. The work is still at a very preliminary stage. The presented example from emotive speech suggests that the described technique is useful also for other speech dimensions. Future applications of the gained knowledge are to be found in next generation speech synthesis and speech understanding systems.

ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish National Board for Technical Development, The Swedish Council for Research in the Humanities and Social Sciences, and the Swedish Telecom.

REFERENCES

[1] Bladon, A., Carlson, R., Granström, B., Hunnicutt, S. & Karlsson, I. (1987): "Text-to-speech system for British English, and issues of dialect and style", *European Conference on Speech Technology, vol. 1*, Edinburgh, Scotland.
[2] Cahn, J. E. (1990): "The generation of affect in synthesized speech", *Journal of the American Voice I/O Society, vol. 8*, pp. 1-19.
[3] 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
[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] Carlson, R., Granström, G. & Nord, L. (1990): "The KTH speech database", *Speech Communication, vol. 9*, pp. 375-380.
[6] 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.
[7] 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.
[8] Williams, C. E. & Stevens, K. N. (1972): "Emotions and speech: some acoustical correlates", *JASA vol. 52*, pp. 1238-1250.