SOME NOTES ABOUT RESEARCH AND DEVELOPMENT AT KTH

Rolf Carlson

Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

The Department of Speech Communication and Music Acoustics of The Royal Institute of Technology is a three-professor department with chairs in: Speech Communication, Music Acoustics, and Hearing Technology. The activities within the department are multi-disciplinary with a concentration on research.

The speech communication group is the biggest within the department with more than 20 researchers. The work covers a wide variety of topics, ranging from detailed theoretical developments of speech production models through phonetic analyses to practical applications of speech technology.

Contacts with other research institutions in Sweden and internationally are well developed. Of major importance in this respect is the scientific journal, the STL-QPSR (Speech Transmission Laboratory, Quarterly Progress and Status Report) which is published three or four times a year in 1000 copies distributed to 45 countries.

The department is engaged in European cooperation within the COST-project on speech synthesis and recognition, and has also been invited to the European Community project on Blindness and Technology. Together with the Swedish Telecom, we are engaged in the SAM project (Speech Assessment Methods) in the second phase of ESPRIT. The Swedish Telecom is an important industrial contact for the department. Close technical and scientific contacts exist with Infovox AB which markets multi-lingual text-to-speech and also recognition products based on our research and development.

For many years we have been engaged in studying speaker characteristics both from an analytical point of view and in modelling different speaker and speaking styles. A speech database is under development in which realizations of segmental and prosodic structures are investigated. Variation in coarticulation strategies, reductions and elaborations along the hypo/hyper speech dimension will be studied in context on this database. Studies of positional variants of sonorants and duration models have already been carried out.

The text-to-speech project [8] has recently focused on improved prosodic models and new strategies for segmental synthesis. A new synthesis model has been developed which includes a new model of the voice source. It will increase the possibilities of modelling different speaker characteristics and speaking styles. Methods in the speech recognition project have been influenced by this work. As a consequence the two projects are, in some respect, merging. This has made our speech recognition efforts slightly different from the general trend.

We now emphasize the research on knowledge-based recognition for large vocabularies. The research program "Nebula" [4,5], includes, for example, speech analysis based on auditory models, feature extraction in a parallel network [9], and prediction models based on speech synthesis [1,2,3].

A major effort at the department has been to study voice source characteristics [6]. The voice source has pronounced effects on the overall shape of the speech spectrum. Intra-speaker voice source variation can cause severe spectral distortion and contributes to recognition errors in current speaker-independent as well as in speaker-dependent recognition systems. The high frequency region of the speech spectrum can vary by 20-30 dB relative to the low-frequency region for a single speaker and still more between among speakers. The voice source carries mainly non-segmental information (apart from the voiced/unvoiced distinction). The prosodic information carried by the voice source is important and should not be discarded. This information is lost in many of the current techniques using parameter estimation methods intended to be insensitive to voice source behavior. Since the voice characteristics are changing during an utterance, the adaptation should be part of the recognition process itself. The speed of the adaptation should be faster than in normal acoustic adaptation. Modelling the source of variation rather than the effect on the speech acoustics potentially makes fast adaptation possible.

As more explicit knowledge on speech production is collected and formulated, it is of interest to explore the use of such information in speech recognition research. A description of speech on a level closer to articulation, rather than the acoustic base that is used in present-day speech recognition will make generalization to different speakers easier. The production component in the form of a speech synthesis system will ideally make the collection of training data unnecessary. During the last year, special projects studying speaker-independent recognition based on stored phoneme prototypes have been undertaken [3]. In these experiments, the references are generated during the recognition process itself. Thus, it is possible to dynamically take into account word juncture and word position effects. The synthetic references can be modified to match the voice of the current speaker. The experiments have shown promising results, see Table 1.

a	DTW	natural references	93%
b	DTW	synthetic references from a text-to-speech system	89%
c	Finite state network	synthetic references	88%
d	Finite state network	synthetic references with voice adaptation	96%

Table 1. Results of pilot experiments using synthetic references d) with and c) without adaption. The score indicates correct word identification in an isolated word recognition test. Our text-to-speech system was used as a base-line in b) and human speakers in a).

Artificial neural networks have also been explored as part of a recognition system [9]. We know from linguistic research that phonetic features are a powerful tool to separate phones from each other. A special network was trained to recognize a number of features. The features were recognized with 80% to 95% accuracy. The features were combined with the original spectrum as input to a phone recognizer. The work will be continued with a dynamic assignment of feature strengths.

Most of the speech technology application studies are now made outside the department. Some recent thesis work has, however, been devoted to this, e.g., speech recognition for air traffic controllers and speech synthesis in the process industry. Speech recognition has also been used in a system for environment control for persons with severe mobility impairments. Speech recognition for mobile telephony has been developed in cooperation with Infovox and Ericsson. The noisy environment in the car has called for several modifications to the original algorithms.

ACKNOWLEDGEMENTS

The work is supported by the Swedish Board for Technical Development. I would also like to thank the organizing committee for inviting foreign guests to participate in this interesting workshop.

REFERENCES

1. Blomberg, M. (1989): "Synthetic phoneme prototypes in a connected-word speech recognition system,"in Proc. ICASSP-Glasgow, Vol. 1, Edinburgh, UK

2. Blomberg, M. (1989): "Voice source adaptation of synthetic phoneme spectra in speech recognition," in Eurospeech 89, Paris, Vol. II, CPC Consultants Ltd, Edinburgh, UK

3. Blomberg, M. (1990): "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," Proc. ESCA Workshop on Speaker Characterization in Speech Technology, Edinburgh, UK, June 1990, CSTR, Univ. of Edinburgh

4. Blomberg, M., Carlson, R., Elenius, K., Granstrom, B., & Hunnicutt, S. (1988): "Word recognition using synthesized templates," in Proc. SPEECH '88, Book 4 (7th FASE-symposium), Institute of Acoustics, Edinburgh.

5. Blomberg, M., Carlson, R., Elenius, K., Granstrom, B. & Hunnicutt, S. (1988): "Word recognition using synthesized reference templates.", Proc. Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Hawaii, USA, also in STL-QPSR 2-3/1988, pp. 69-81.

6. Carlson, R., Fant, G., Gobl, C., Granstrom, B., Karlsson, I. & Lin, Q. (1989): "Voice source rules for text-to-speech synthesis", Proc IEEE 1989 Int. Conf. on Acoustics, Speech, and Signal Processing, Glasgow, Scotland

7. Carlson, R. Granstrom, B. & Karlsson, I.(1990): "Experiments with voice modelling in speech synthesis", Proc. of ESCA workshop on Speaker Characterization in Speech Technology, 26-28 June 1990, Edinburgh

8. Carlson, R., Granstrom, B. & Hunnicutt, S.(1991): "Multilingual text-to-speech development and applications", A.W. Ainsworth (ed), Advances in speech, hearing and language processing, JAI Press, London

9. Elenius K. (1990):"Acoustic-phonetic recognition of continuous speech by artificial neural networks", in STL-QPSR 2-3 1990.