Speech Recognition in the Waxholm Dialog System

Mats Blomberg, Kjell Elenius & Nikko Ström*
Department of Speech Communication and Music Acoustics
KTH, Box 70044, S 100 44 Stockholm, Sweden

ABSTRACT

The speech recognition component in the KTH "Waxholm" dialog system is described. It will handle continuous speech with a vocabulary of about 1000 words. The output of the recogniser is fed to a probabilistic, knowledge-based parser, that contains a context-free grammar compiled into an augmented transition network.

INTRODUCTION

The KTH dialog demonstrator application, *Waxholm*, gives information on boat traffic in the Stockholm archipelago. It was first presented at the Eurospeech '93 conference (Blomberg et al. 1993). The system contains modules that handle speech recognition, speech synthesis and graphic information such as pictures, maps, charts, and time-tables.

Initially speech data have been collected using a Wizard-of-Oz technique, in which the recognition component was replaced by a human. The recorded speech is used for training and testing the recogniser.

SPEECH RECOGNISER

The speech recognition component will handle continuous speech with a vocabulary of about 1000 words. The work on recognition has been carried out along two main lines: artificial neural networks and a speech production oriented approach. Since neural nets are general classification tools, it is quite feasible to combine the two approaches.

Speech production approach

Our system uses a speech synthesis technique to generate spectral prototypes of words in a given vocabulary, see Blomberg (1991). A speaker-independent recognition system has been built according to the speech production approach, using a formant-based speech production module including a voice source model. Whole word models are used to describe intra-word phonemes, while triphones (three-phoneme clusters) are used to model the phonemes at word boundaries. An important part of the system is a method of dynamic voice-source adaptation. The recognition errors have been significantly reduced by this method.

Artificial neural networks

We have tested different types of artificial neural networks for performing acoustic-phonetic mapping for speech signals, see Elenius & Takács (1990), Elenius & Blomberg (1992), Elenius & Tråvén (1993). The tested strategies include self-organising nets and nets using the error-back propagation (BP) technique. The use of simple recurrent BP-networks has been shown to substantially improve performance.

* Names in alphabetic order.

Lexical search

The frame based outputs from the neural network form the input to the lexical search. There is one output for each of the 40 Swedish phonemes used in our lexicon. Each word in the lexicon is described on the phonetic level. The lexicon may include alternate pronunciations of each word. The outputs are seen as the aposteriori probabilities of the respective phonemes in each frame. We have implemented an A* N-best search using a simple bigram language model. In a second stage the speech production approach mentioned above will be used to reorder the N-best list according to speaker specific criteria. A tight coupling between the parser and the recognise is a long-term goal in the project. This will naturally influence the search algorithms.

ACKNOWLEDGEMENT

This work has been supported by The Swedish National Language Technology Program.

REFERENCES

Blomberg, M. (1991): "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," Speech Communication, Vol. 10. pp 453-462.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993): "An experimental dialog system: WAXHOLM," Proceedings of Eurospeech '93. pp. 1867-1870.

Elenius K. and Takács, G. (1990): "Acoustic-phonetic recognition of continuous speech by artificial neural networks," STL-QPSR 2-3, Technical Report, Dept. of Speech Comm., KTH, 1990. pp 1-44.

Elenius, K. & Blomberg M., (1992): "Experiments with artificial neural networks for phoneme and word recognition," Proceedings of ICSLP 92, Banff, Vol. 2, pp. 1279-1282.

Elenius K. & Tråvén H. (1993): "Multi-layer perceptrons and probabilistic neural networks for phoneme recognition," Proceedings of Eurospeech '93. pp 1237-1240.