



SPEECH RECOGNITION BASED ON A TEXT-TO-SPEECH SYNTHESIS SYSTEM

Mats Blomberg*, Rolf Carlson*, Kjell Elenius*, Björn Granström* and Sheri Hunnicutt*

ABSTRACT

A major problem in large-vocabulary speech recognition is the collection of reference data and speaker normalization. In this paper we propose the use of synthetic speech as a means of handling this problem. An experimental scheme for such a system will be described.

NEBULA

During the last years many experiments have been carried out at our department concerning different aspects of speech recognition and speech perception. At the same time work on speech synthesis has been pursued. In this paper we will briefly describe a speech recognition scheme, NEBULA, which combines results and methods from these efforts into a coherent system. According to this scheme, some experiments have been undertaken. Preliminary results of an isolated word recognition experiment will be presented and discussed.

THE PARALLEL FRONT END

Inspired by the parallel nature of human speech perception we have developed a special framework to formulate and explore the front end of our recognition system (ref 1). The speech is represented as a continuous flow of information in multiple channels. This makes it possible to use diverse analysis mechanisms which can be simple but should work in a coordinated structure. The system is used to formulate the lower levels of speech analysis including spectral transformations, lateral inhibition, temporal onset/offset effects, and a variety of phonetic-cue detectors.

Using conventional signal processing techniques we have earlier tried some of the proposed auditory representations in the context of a speech recognition system (ref 2). Based on one of these models, the DOMIN model, we are currently implementing a new primary analysis module that will enhance spectral peaks and suppress valleys.

THE LEXICAL COMPONENT

The front end explores the descriptive power of cues, and uses multiple cues to analyze, classify, and segment the speech wave. These classifications are used during the lexical search (ref 3). Additional information from a prediction system (ref 4) is also used in the lexical selection part of NEBULA. As a result of this component we get a selection of possible words, a cohort.

* Names in alphabetic order.

Department of Speech Communication and Music Acoustics Royal Institute of Technology Box 700 14, S-100 44 Stockholm, Sweden.

REFERENCES FROM A TEXT-TO-SPEECH SYSTEM

The phonetic component of a text-to-speech system is used to create references from the cohort. The synthesis system has been described elsewhere (ref 5) and is also presented at this conference (ref 6). These references are sent to the identification and verification part of NEBULA. Speech synthesis has been explored in other speech recognition projects, for example in (refs 7, 8).

THE IDENTIFICATION COMPONENT

There are presently two types of recognition techniques available for NEBULA. One is a whole-word pattern-matching system based on filter bank analysis, cepstral transformations and non-linear time warping, described in more detail in (ref 9). As in other systems of this kind, a separate training session is needed to establish acoustic reference data. In our system, the reference material is provided by the rule synthesis system. It will be possible to have the references generated during the recognition process and then take into account word juncture and word position effects, which is not easily achievable in conventional word based speaker-trained systems. This is the method used in the present experiments.

The second method is based on phonetic recognition using a network representation of possible realisations of the vocabulary. The acoustic analysis is the same as in the previously described method, but the phonetic decisions are based on comparisons to a library of synthetic allophones. The network approach enables handling of optional pronunciations. On the other hand, non-stationary parts of the speech wave may be better represented by a more detailed description of the time evolution of the utterance, as in the first method. A combination of the two methods would enable the advantages of both techniques to be used.

THE EXPERIMENTAL COHORT

In the current experiment, the lexical search was simulated and a 26-word cohort was chosen which was of the type 'VCVCCVC. In this case, the suggested preliminary analysis only discriminates between vowel and consonant and identifies the stressed syllable. The cohort is drawn from a corpus consisting of the most frequent 10,000 words of Swedish. The word structure is, in most cases, a compound word with a bisyllabic initial morph. The structure is rich enough to expose a variety of deviations among the human speakers from the norm pronunciation used in the synthesis. These deviations generally occur across the compound boundary. Both deletions like [øvar] → [øva] and insertions or hypercorrect pronunciations occur, e.g, [otæʂtor] → [otærstor]. 37 such deviations were identified among the total of 26*10 words recorded. This cohort constitutes a rather hard recognition vocabulary. Within the cohort there are many examples of morphological overlap as can be seen from the list of words in table 1. One word pair differs in only one consonantal segment.

Table 1 Words in the experimental cohort

1 obekväm.	2 uppenbar.	3 ingenting.	4 ovanför.
5 innanför.	6 inifrån.	7 inomhus.	8 utanför.
9 utifrån.	10 utomhus.	11 uppifrån.	12 egendom.
13 enighet.	14 äventyrs.	15 äventyr.	16 överskott.
17 övergick.	18 öppenhet.	19 uteblev.	20 uteslöt.
21 övergår.	22 övergett.	23 återstod.	24 återstår.
25 återger.	26 återkom.		

SPEECH MATERIAL AND PRELIMINARY RECOGNITION RESULTS

Ten male subjects participated in the experiment. The test vocabulary was recorded in a normal office room with additional noise from a personal computer. The subjects were asked to read the words from a list with little instruction except to pronounce each word separately. These recordings were used as input to the pattern matching verification component of NEBULA together with synthesized versions of the cohort. The synthesis was used to build the references. No adjustments were done to the synthesis in this first stage. 74.6% of the test words were identified correctly.

In addition to the synthesis, each speaker was used to create references for the other speakers. All the human speakers served better as reference than the synthesis, and ranged from 79.1% to 93.6% correct with an average value of 89.5%.

TEMPORAL NORMALIZATION

The significantly better results when human references are used was not a surprise. It is well known that text-to-speech systems still need more work before they reach human quality. However, the result was encouraging. At an early point we noticed discrepancies in the durational structure of the synthetic and the human speech. This created errors despite the fact that dynamic programming was used. The synthesis system creates reductions if the duration of a phoneme is short. If, on the other hand, the duration is long, the steady state targets are reached. Differences in segment duration will then cause spectral differences that cannot be eliminated by the time warping procedure. The durations for one speaker was measured and the durational framework for each word was imposed on the synthesis. The result showed a significant improvement, reaching 81.5% correct, slightly better than our worst human speaker.

SPECTRAL NORMALIZATION

One important component of the NEBULA system is the normalizing process which should adjust the system to a certain speaker. In our last experiment, we analysed the vowels for one speaker and adjusted the synthesis to have the same formant targets. The consonant targets, all amplitudes and the smoothing algorithms were kept constant. Some major coarticulation errors were observed but were not corrected. The recognition result was slightly lower than with durational normalization, 79.6%. On the other hand, the

number of errors for the measured speaker went down from 4 to 2. The same effects were observed when adapting the vowel formants to another speaker. The reduction in errors for the target speaker was expected. We had also hoped for a reduction of errors for all speakers, since the subjective overall speech quality increased. This was not the case, however. A more detailed acoustic analysis of the synthesis in comparison with the speaker will be presented at the conference.

CONCLUDING REMARKS

In this experiment, which is part of a long term knowledge based speech recognition project, NEBULA, we have taken the extreme stand of comparing human speech to predicted pronunciations on the acoustic level with a straightforward pattern matching technique. This has given us valuable feed-back on both the prediction/synthesis component and the matching algorithm and also some information on how these components interact when exposed to a variety of speakers.

REFERENCES

1. Carlson, R., Granström, B., & Hunnicutt, S. (1985): "A parallel speech analyzing system", STL-QPSR 1.
2. Blomberg, M., Carlson, R., Elenius, K., & Granström, B. (1984): "Auditory models in isolated word recognition," Conference Record, IEEE-ICASSP, San Diego.
3. Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1986): "Phonetic properties of the basic vocabulary of five European languages: implications for speech recognition", Conference Record, IEEE-ICASSP, Tokyo.
4. Hunnicutt, S. (1986): "Lexical Prediction for a Text-to-Speech System", Communication and Handicap: Aspects of Psychological-Compensation and Technical Aids, E. Hjelmquist and L.-G. Nilsson, editors, Elsevier Science Publisher B. V. (North Holland).
5. Carlson, R., Granström, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module," Conference Record, IEEE-ICASSP, Paris.
6. Bladon, A., Carlson, R., Björn Granström, B., Hunnicutt, S. and Karlsson, I. (1987): "A text-to-speech system for British English, and issues of dialect and style ", This Conference.
7. Woods, W.A., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhol, J., Nash-Webber, B., Schwartz, R., Wolf, J. and Zue, V. (1976): "Speech Understanding Systems - Final Technical Progress Report," Report No 3438, BBN, Cambridge, Ma, USA.
8. Bridle, J.S. and Chamberlain, R.M. (1983): "Automatic Labelling of Speech using Synthesis-by-Rule and Non-Linear Time Alignment," 11th I.C.A. Toulouse Satellite Symposium, Vol 9.
9. Blomberg, M., Elenius, K., Neovius, L., Tjernlund, P. (1987): "Speech recognition in mobile applications", This conference.