Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Word recognition using synthesized templates

Blomberg, M. and Carlson, R. and Elenius, K. O. E. and Granström, B. and Hunnicutt, S.

journal: STL-QPSR
volume: 29
number: 2-3
year: 1988
pages: 069-081

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

II. SPEECH SYNTHESIS

A. WORD RECOGNITION USING SYNTHESIZED TEMPLATES*

Mats Blomberg, Rolf Carlson, Kjell Elenius, Björn Granström, & Sheri Hunnicutt

Abstract

With the ultimate aim of creating a knowledge based speech understanding system, we have set up a conceptual framework named NEBULA. In this paper we briefly describe some of the components of this framework and also report on some experiments where we use a production component for generating reference data for the recognition. The production component in the form of a speech synthesis system will ideally make the collection of training data unnecessary. Preliminary results of an isolated word recognition experiment will be presented and discussed. Several methods of interfacing the production component to the recognition/evaluation component have been pursued.

1. 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. The speech recognition scheme, NEBULA, combines results and methods from these efforts into a coherent system. The system is presented in Fig. 1.

1.1. The front end

Using conventional signal processing techniques, we have earlier tried some of the proposed auditory representations in the context of a speech recognition system (Blomberg, Carlson, Elenius, & Granström, 1984). Based on one of these models, the DOMIN model, we are currently working on a new primary analysis module. This peripheral auditory model explores the possibilities for synchrony effects that will enhance spectral peaks and suppress valleys. At the same time, wide band effects will be taken into account.

1.2. Feature extraction

At the output of the auditory model, the speech is represented as a continuous flow of information in multiple channels (Carlson, Granström, & Hunnicutt, 1985). This makes it possible to use diverse analysis mechanisms which can be simple but should work in a coordinated structure. The goal is to formulate the lower levels of speech analysis as a parallel structure. The process could include spectral transformations, lateral inhibition, temporal onset/offset effects, and a variety of phonetic-cue detectors.

*This is an expanded version of a paper presented at the FASE SPEECH'88 meeting i Edinburgh. This paper also includes some new experiments along the same lines.

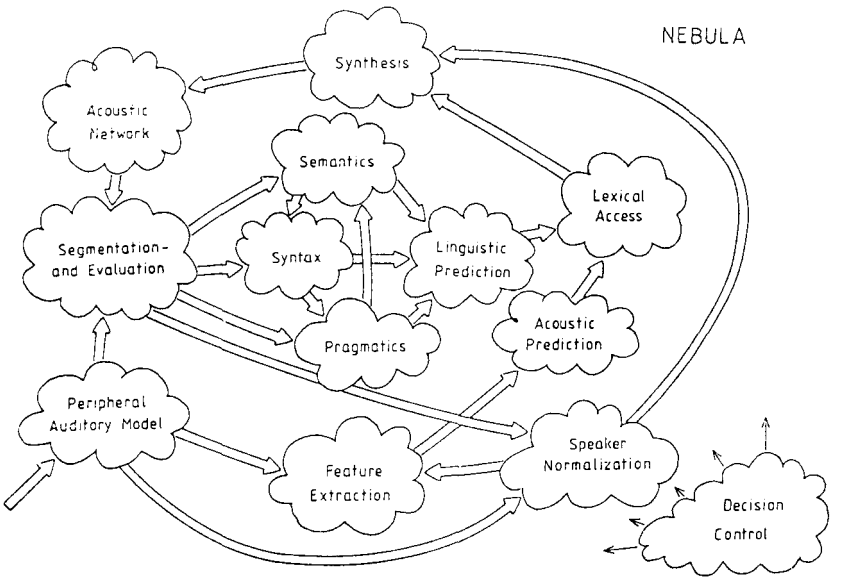

Fig. 1. The speech recognition scheme, NEBULA.

1.3. The lexical component

The low levels of NEBULA explore 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 (Carlson, Elenius, Granström, & Hunnicutt, 1986). Additional information from a prediction system 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.

1.4. High level linguistic processes

The mid-portion of NEBULA is currently represented by a syntactic component of the text-to-speech system, morphological decomposition in the text-to-speech system, and a concept-to-speech system. A special phrase structure grammar is employed which takes account of word order, phrase order, and grammatical information. These parts were originally developed for a different purpose than speech recognition, but we expect them to be applicable in this area as well.

One of several word prediction algorithms has been designed to find cohorts of possible words from partial information generated by a word recognition scheme. Other prediction algorithms are being used in handicap applications, to help persons with a speech or motoric disability or in aphasia rehabilitation. These algorithms are currently being complemented with syntactic and semantic components, Hunnicutt (1986).

1.5. 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 elsewhere (Elenius &

Blomberg, 1986). 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.

1.6. Word 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 (Carlson, Granström, & Hunnicutt, 1982). It is based on rules and has a formant synthesizer as output module. These references are sent to the identification and verification part of NEBULA.

2. Using synthetic templates: previous work

Use of synthetic speech as a reference for aligning natural speech with dynamic programming techniques has been reported by many authors, i.e., Chamberlain & Bridle, 1983; Hunt, 1984; Höhne, Coker, Levinson, & Rabiner, 1983; Woods & al., 1976. The papers by Chamberlain and by Höhne are mainly concerned with the time-warping aspects of mapping long utterances (sentences) to each other. Hunt cites four reasons besides the obvious one of improving speech synthesis for the research in this field. First he mentions the analysis by synthesis based technique as a good method for extracting formant frequencies, which seem to be better for indicating phonetic identity than the gross spectral shape, often used in speech recognition. Another property of synthetic speech is that it can modified to match the voice of the current speaker. Synthesis can also be used to exploit knowledge that is available about natural speech such as duration and the context of a word. Finally he discusses the positive effect of the perfect consistency of synthetic speech. It can be used for speaker verification, where the speaker characteristics can be related to synthetic speech. In recognition the consistency may be used to improve separation between words such as stalagmite and stalactite that have phonetically identical parts. After having compared synthetic speech (MITalk) to natural speech in some recognition experiments, he notes that aligning speech between natural speakers gives considerably better results than using synthetic speech. His conclusion is that the synthesis rules must be improved before the synthesis can give comparable results.

3. Some experiments within NEBULA

In 1987 we reported on the ongoing work inside some of the clouds in the NEBULA scheme (Blomberg, Carlson, Elenius, & Granström, 1987). At that time the experiments were run on three different computers and two different kinds of special hardware. This created practical problems and slowed down the continuation of the work.

During the last year, both the synthesis and the recognition software have been implemented on our Apollo work stations. This has opened up new possibilities to make additional experiments along the same lines as before. The interaction between the different modules in NEBULA is now fast and easy. We will in this paper review the earlier work and report some results from a new test series. This paper also includes additional experiments that were not included in the written version that is published in the proceeding of the SPEECH'88 FASE symposium (Blomberg, Carlson, Elenius, Granström, & Hunnicutt, 1988).

3.1. Test vocabulary and subjects

In all experiments reported in this paper, the lexical search was simulated. In this case, the suggested preliminary analysis only discriminates between vowel and consonant and identifies the stressed syllables. A 26-word cohort was chosen which was of the type "VCVCC'VC. It was drawn from a corpus consisting of the most frequent 10,000 words of Swedish. Ten male subjects were asked to read the 26 words from a list with little instruction except to pronounce each word separately. The vocabulary was recorded in a normal office room with additional noise from a personal computer.

The word structure, in most cases, is a compound word with a bisyllabic initial morph. The structure is rich enough to expose a variety of deviations among the subjects and a synthesized reference. These deviations generally occur across the compound boundary. Both deletions and insertions or hypercorrect pronunciations occur and 37 such deviations from the norm were identified among the total of 26*10 words recorded. Within the cohort there are many examples of morphological overlap as can be seen from the list of words in Table 3. One word pair (14 'äventyrs' and 15 'äventyr') differs in only one consonantal segment.

3.2. Preliminary recognition results; EXPERIMENT 1

The recorded speech material was used as input to the pattern- matching verification component of NEBULA, which in the first experiment consisted of the special hardware recognition system developed at KTH. The output from our hardware text-to-speech system was recorded and used to train this system. 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 speakers than the synthesis, and the correct result ranged from 79.1% to 93.6% with an average value of 89.5%.

At an early point we noticed discrepancies in the durational structure of the synthetic and the human speech. Differences in segment duration will cause spectral differences that cannot be eliminated by a time warping procedure, since time dependent coarticulation and reduction rules are active in the synthesis system. The segmental durations for one speaker were measured and the durational framework for each word was imposed on the synthesis. The result showed an increase to 81.5% correct identification, which is slightly better than our worst human speaker. The results from our experiments are summarized in Tables 1 and 2, and Fig. 2.

Reference patterns from	Experiment 1	Experiment 2
Human Speakers	89.5	90.7
Synthesis	74.6	77.2
Synthesis, Duration Adjusted	81.5	81.2
Software Synthesizer	---.-	76.4

Table 1. Result from the recognition experiment 1 and 2.

3.3. A new environment; EXPERIMENT 2

The new series of experiments use the same recordings as before, but different methods are used to create the spectral templates for the identification component.

As a start, the old recordings of both the human speech and the synthesis were digitized, using 16 kHz sampling frequency. The filter bank was simulated with the help of an FFT analysis followed by pooling of the spectral components into 16 bands, from 200 to 5000 Hz, equally spaced along the Bark scale. The recognition system, now running in the Apollo work station, was used to repeat the same experiment as before. The result can be seen in Table 1. The increased accuracy in the analysis gives a slightly better recognition result for the human speakers and the non-normalized synthesis.

In the following experiment, all parts of the text-to-speech system were running in the same computer including the synthesizer. The result of 76.4% correct identification is presented in Table 1 under the name 'software synthesizer'.

3.4. Parameter generated spectrum; EXPERIMENT 3

The next experiment in this series included a different method to generate spectral frames. The control parameters to the synthesizer were used directly to generate the spectral shape. We can then by-pass several problematic areas in the analysis. The interaction between harmonics and formant peaks in the vowel spectrum can be avoided and the fricative noise spectrum is stable. Fig. 3 gives an example of this method. The control frames to the synthesizer are used to generate the spectral representation in Fig. 3a. These spectral slices are transformed into 16 channels corresponding to the output from the filter bank used in the identification part of the recognition system, Fig. 3b. To the left in Fig. 3c is the synthetic reference for the first word, 'obekväm', in the vocabulary and to the right is the analysis of a speaker saying the same word. Several observations can be made. The noise level has a considerable influence on the spectrum. The speaker uses an unvoiced labio-dental fricative instead of the voiced counterpart in the synthesis. We will return to a more detailed analysis later in the paper.

The frame rate in the synthesis is 10 ms while the recognition system uses 25 ms between each observation. The down-sampling was achieved by simply taking the maximum of two sequential spectral slices. The recognition result with this method was 78.4% correct identification. An alternative method to use every other spectral slice gave a better result, 81.1%. Interpolation between two static spectra will give unwanted effects. If a resonance is moving too fast between two frames a double peak will be stored in the spectral representation. As an alternative, we used all synthesis frames and increased the frame rate in the verification part by two. The synthesis gave a slightly higher value 82.4% correct. A repetition of experiment 2 using each of the human speaker as reference

gave a mean of 92.8%. The two worst cases were 84.7% and 88.9%. These results are presented in Table 2 and Fig. 2.

In Fig. 4 the distance between the correct and the best incorrect match is shown for both a typical human reference and the synthesis. We can observe that the distribution is different in at least two aspects. The data points are closer to the diagonal in the synthesis case compared to the human case. This means the synthesis gives a less confident answer even if it is correct. Also the mean distances between the synthesized reference and the test vocabulary are 17% larger compared to the corresponding distances using references by human speakers. To reduce the distances to a comparable or smaller value and to push it away from the diagonal is a challenge for our continued work.

Reference pattern		
Human Speakers		
25 ms and 10 ms sampling:	90.7	92.8
Parameter Generated Spectrum		
20 ms, max of two frames:	78.4	
Parameter Generated Spectrum		
20 ms and 10 ms sampling:	81.1	82.4

Table 2. Result from the recognition experiment 3.

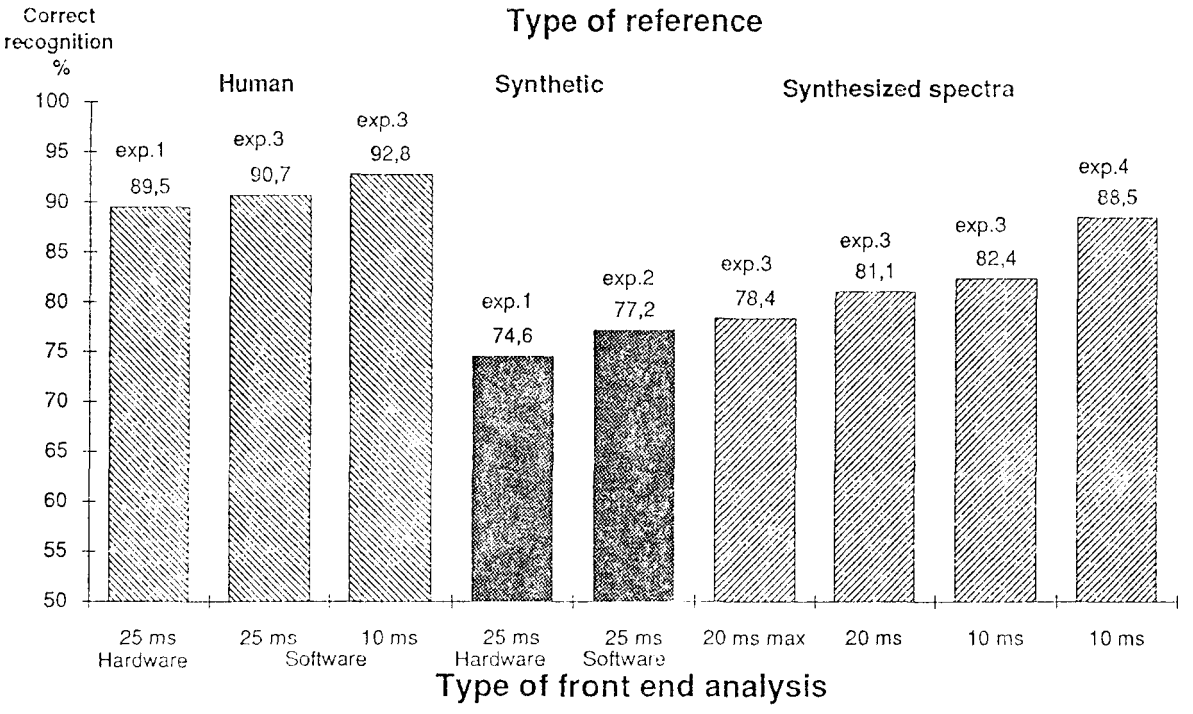

Fig. 2. Recognition results in the experiments using different kinds of references.

A confusion matrix of the experiment 3 (20 ms frames) is shown in Table 3. We can observe that the word 'ingenting' has been over-represented in the responses. A comparison is made in Fig. 5 between the tense vowel /e:/ and the lax /I/ for the same speaker pronouncing words 13:enighet and 3:ingenting. It is obvious that the spectral

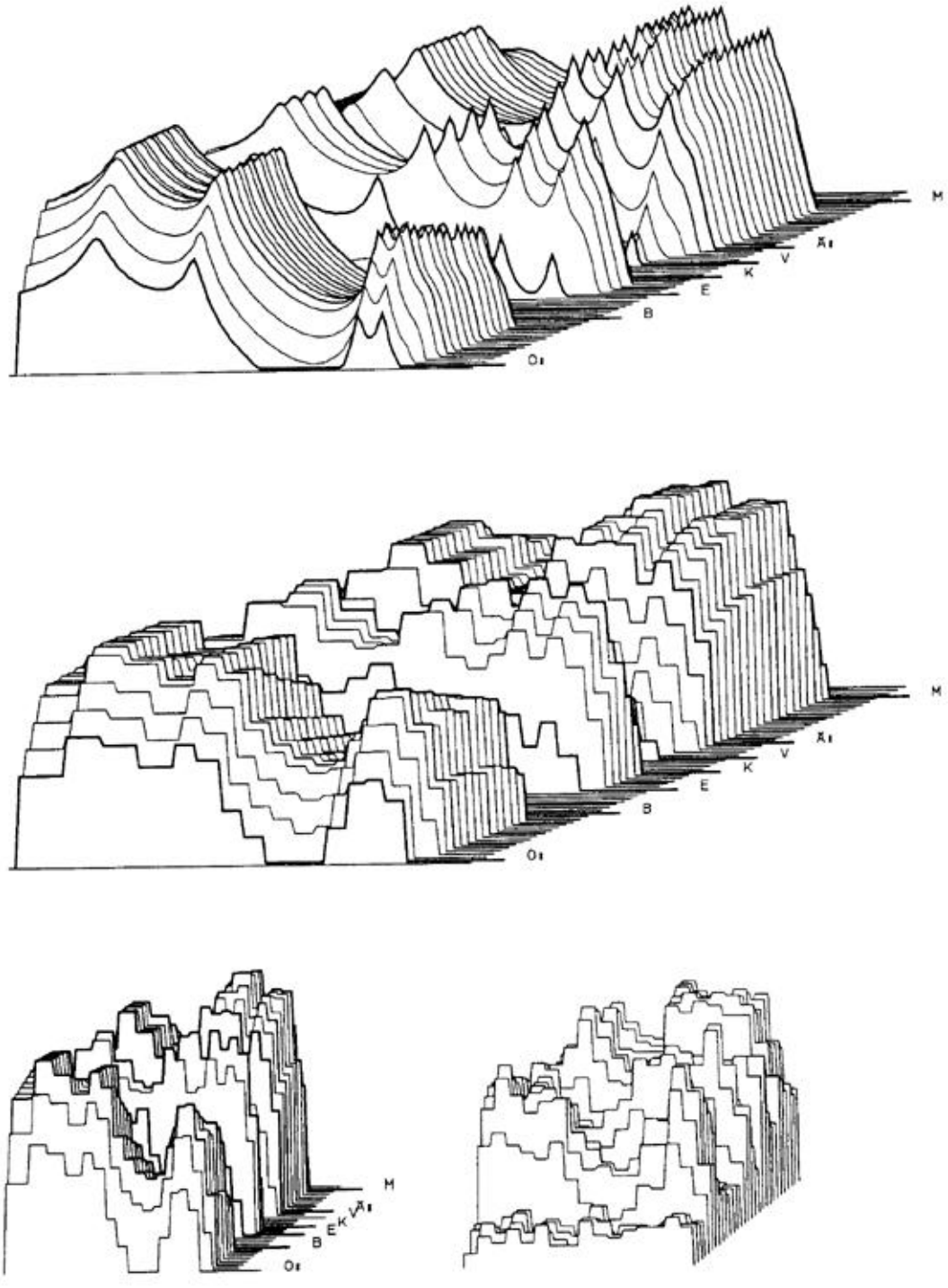

Fig.3. Creation of spectral slices a) and transformation to filter bank representation b). Comparison between synthetic (left) and natural (right) spectral templates for the word 'obekväm'.

shape of the main stressed vowel in this case can not be used as a distinguishing cue. However, relative duration, coarticulation and diphthongization can give supportive information for vowel discrimination.

As a complement to the case study of the /e:/ and /I/ mentioned above we made a statistical analysis of the energy distribution, see Fig. 6. The figures are created by making a two dimensional histogram of the energy/frequency distribution of the observed material. This distribution is then divided in 10% intervals, which are drawn in the graph. The 30%, 50% and 70% are marked with a thicker line. This method is used in our general work with our speech data-base (Carlson & Granström, 1985). The two phonemes were analyzed in initial and final position in the same recorded materials. A small difference in the higher formants can be seen as expected. If we compare the spectrum in initial and final position, we find a small difference in spectral slope which can be referred to the glottal source. Several studies at our laboratory have been dealing with these types of variations (Fant, Liljencrants, & Lin, 1985).

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

Table 3. Confusion matrix for Experiment 3 (20 ms frames).

3.5. Improved synthesis; EXPERIMENT 4

Detailed analysis of the results gave good information on basic errors in the syn-

thesis. First the phonetic transcription was not according to the typical pronunciation of the speakers. The /g/ in 'enighet' was for example deleted by most speakers. This created most of the errors for this word.

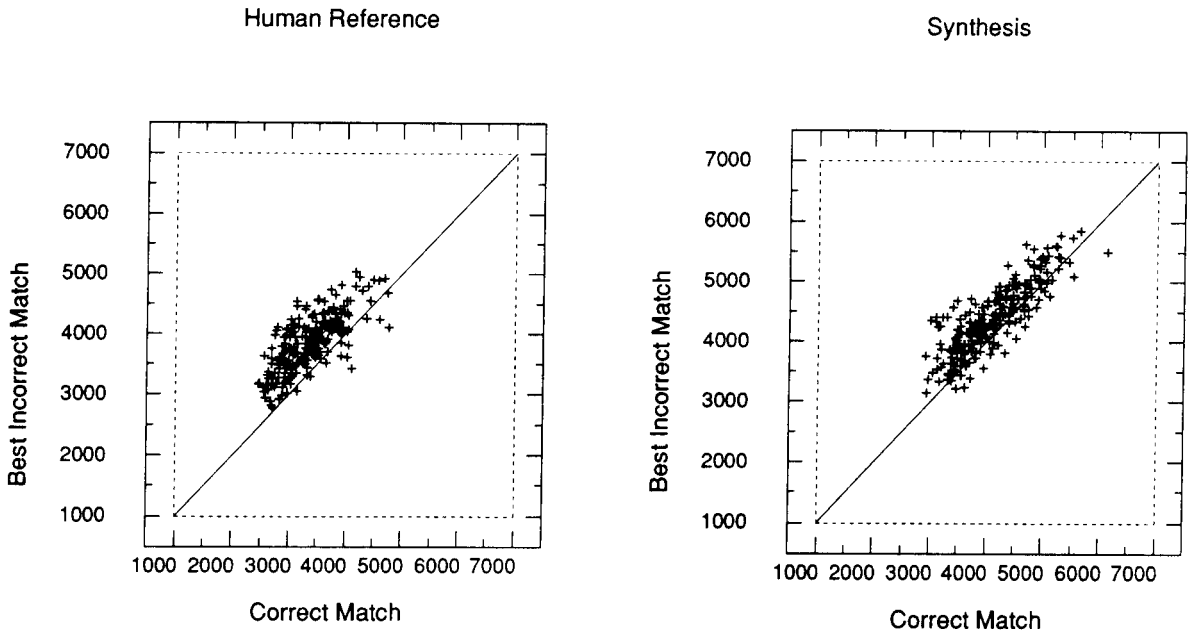

Fig. 4. *The distance between the correct stimulus and the correct reference, and between the correct stimulus and the best incorrect match. Human reference a) and synthetic reference b).*

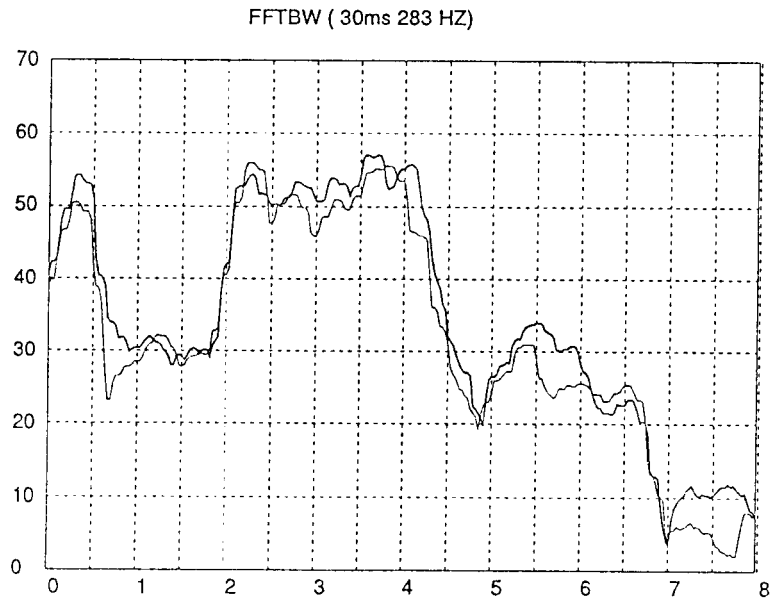

Fig. 5. *Spectral section of /I/ in the word 'ingenting' and /e:/ in the word 'enighet' for one of the speakers.*

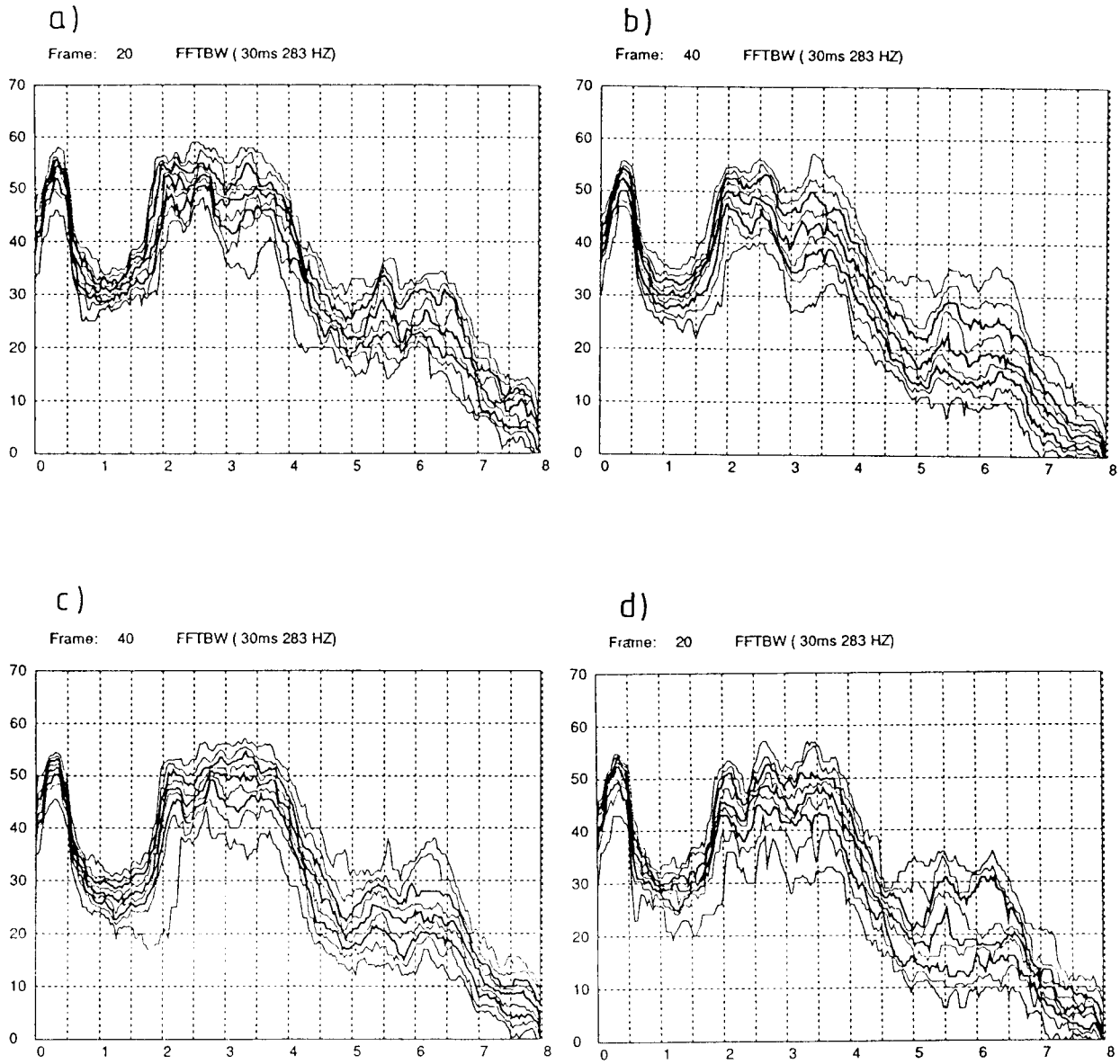

Fig. 6. Energy distribution for initial and final /e:/ (top) and initial and final /I/ (bottom).

The synthesis system used so far is based on smoothed square waves for most of the parameters. This has proven to be a good method for interpolation in many cases. It will automatically create reduction effects when the duration of a segment is short. The method is related to the thinking that the production of speech can be simulated by control step functions smoothed by the muscular/mechanical system of articulators. However, the frequency domain is probably not the correct domain for this smoothing. Articulatory parameters are more natural in this respect. As an alternative we are currently building a slightly different synthesis system where the parameters are specified by target values and the time it takes to reach this value. The movement towards a target can be interrupted by a new target. Unfortunately this method adds new demands on the system. Reduction does not come automatically as before. It has to be described in a more explicit way. On the other hand phonetic knowledge can be more accurately specified.

As a final experiment word references from this new system was tested in a recognition experiment. Several new synthesis aspects were considered. Different allophones for the /r/ phoneme was used in the 'CV' or 'VC' position. The diphthong /e:/ was adjusted. Fig. 7 shows the synthesis spectrum for /e:/ compared to the energy histogram from Fig. 6a. This experiment gave a result of 88.5% accuracy, which is better than the worst human reference and close to the second to last one.

3.6. Analyzing identification results

In order to analyze the recognition results we can use a program giving a display as seen in Figs. 8a and 8b. Before time aligning words by dynamic programming they are linearly normalized to a nominal length of 800 ms - or 80 cepstral frames, since frames in this experiment are calculated at 10 ms intervals. At the top we see the warping function as a thin line. The bold line is a cepstral difference function between the matched words calculated along the warping function. Below this plot we see the three energy functions displaying: 1) the test word (bold), 2) the time warped reference word, and 3) the reference word. Below these we see two 16 band spectral sections of corresponding frames at a point along the time warp. The bold section belongs to the test word and the other is from the reference. The time point is marked by a vertical bar in the warped difference function at the top.

The display makes it possible to interactively analyse why words are misrecognized and what part of the words are mismatched. It also gives a means of understanding what makes the test word more similar to an incorrect reference than to the correct one. In this case we are analysing the test word '**äventyr**' by speaker GF that was erroneously identified as the synthesized reference '**ingenting**'. In Fig. 8a we see the result of matching '**äventyr**' by GF to the synthesized version of the same word and in Fig. 8b we see it matched to the synthesized word '**ingenting**'. The spectral sections are from the first vowel as marked by the vertical bar in the cepstral difference plot. The cepstral distance is larger between 'ae' and the synthetic 'ae' than between 'ae' and the synthetic 'i'. The corresponding plots of spectral sections show that in this case the synthetic 'ae' (8a) has too little low energy compared to the natural voice (the bold line). It should be stressed that during identification the matching is done in the cepstral domain, not in the frequency domain, and that one should be careful about conclusions drawn from looking only at spectral sections.

3.7. Comparison between the experiments

Fig. 2 shows the results from all experiments. It can be seen that various improvements have successively raised the recognition accuracy. The increased frame rate gave a bigger improvement for the human references compared to the synthetic.

4. Concluding remarks

In our paper we have reported on some experiments, which are part of a long term project towards a knowledge based speech recognition system, NEBULA. We have taken the extreme stand in these experiments of comparing human speech to predicted pronunciations on the acoustic level with the help of a straightforward pattern matching technique. 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 results can be regarded as encouraging.

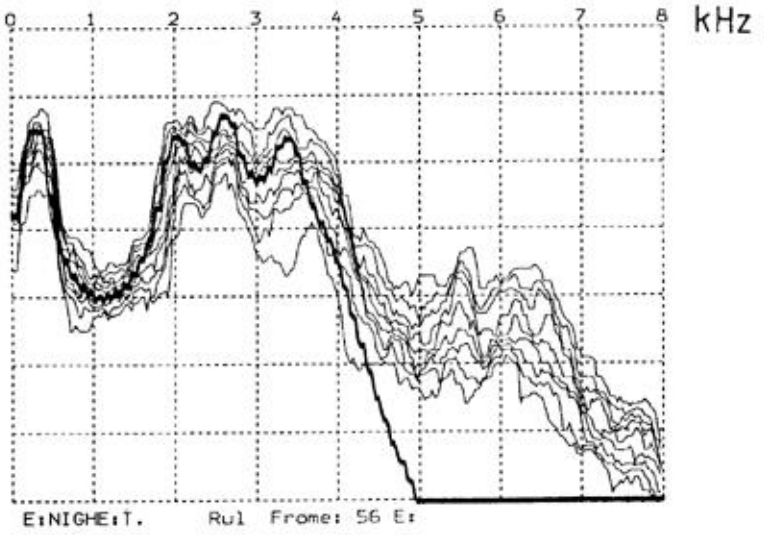

Fig. 7. *Comparison between synthetic /e:/ and energy histogram.*

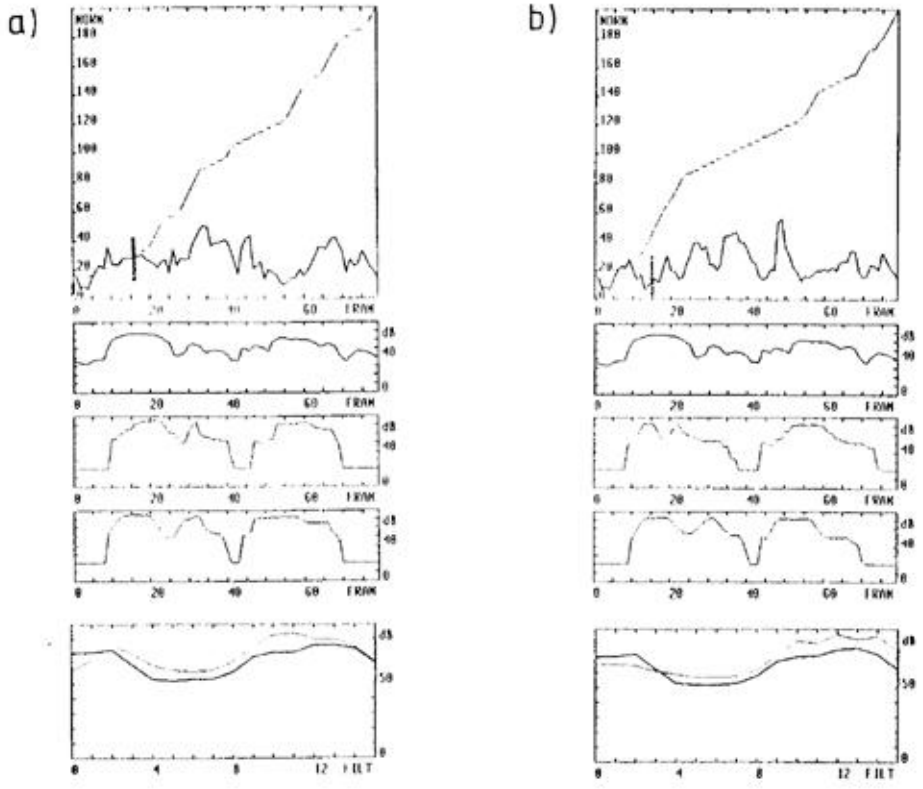

Fig. 8. *Plots showing word matchings in the identification component. In 8a the test word 'äventyr' by speaker GF is matched to the same synthesized reference word and in 8b the same word 'äventyr' is matched against the synthetic word reference word 'ingenting'. See section 4.6 for more information.*

In the last experiments we reached an important goal in our work strategy. We have created an experimental system that gives us control of each separate module of the system. We can easily do a spectral comparison between synthesis and human speech. We can adjust the spectral shapes in order to adapt the synthesis to a specific speaker. This will give us valuable feed-back on both the prediction/synthesis component and the matching algorithm, and some information on how these components should interact when exposed to a variety of speakers.

References

Blomberg, M., Carlson, R., Elenius, K., & Granström, B. (1984): "Auditory models in isolated word recognition", pp. 17.9.1-17.9.4 in *Proc. IEEE-ICASSP*, Vol. 2, San Diego.

Blomberg, M., Carlson, R., Elenius, K., & Granström, B. (1987): "Speech recognition based on a text-to-speech system", pp. 369-372 in (J. Laver & M.A. Jack, eds.): *European Conf. on Speech Technology, Edinburgh, Vol. II,* CEP Consultants Ltd., Edinburgh. .

Blomberg, M., Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1988): "Word recognition using synthesized templates", p. 1171 in *Proc. SPEECH'88, 7th FASE Symposium, Edinburgh.*

Carlson, R. & Granström, B. (1985): "Rule-controlled data base search", STL-QPSR 4/1985, pp. 29-45.

Carlson, R., Granström, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module", pp. 1604-1607 in *Proc. IEEE-ICASSP, Vol. 3,* Paris.

Carlson, R., Granström, B., & Hunnicutt, S. (1986): "A parallel speech analyzing system", STL-QPSR 1/1985, pp. 47-62.

Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1986): "Phonetic properties of the basic vocabulary of five European languages: implications for speech recognition", pp. 2763-2766 in *Proc. IEEE-ICASSP, Vol. 4,* Tokyo.

Chamberlain, R.M. & Bridle, J.S. (1983): "ZIP: A dynamic programming algorithm for time-aligning two indefinitely long utterances", in *Proc. IEEE-ICASSP.*

Elenius, K. & Blomberg, M. (1986): "Voice input for personal computers", pp. 361-372 in (G. Bristow, ed.): *Electronic Speech Recognition,* William Collins Sons & Co. Ltd., London.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A four-parameter model of glottal flow", STL-QPSR 4/1985, pp. 1-13.

Hunnicutt, S. (1986): "Lexical prediction for a text-to-speech system", pp. 253-263 in (E. Hjelmquist & L-G. Nilsson, eds.): *Communication and Handicap: Aspects of Psychological Compensation and Technical Aids,* Elsevier Science Publ.

Hunt, M.J. (1984): "Time alignment of natural speech to synthetic speech", p. 2.5.1 in *Proc. IEEE-ICASSP.*

Höhne, H.D., Coker, C., Levinson, S.E., & Rabiner, L.R. (1983): "On temporal alignment of natural and synthetic speech", p. 807 in IEEE-ASSP, Vol. 31

Woods, W., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., akhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., & Zue, V. (1976): "Speech understanding systems - Final Technical Progress Report", Report No. 3438, BBN, Cambridge, MA, USA.