SYNTHETIC PHONEME PROTOTYPES IN SPEECH RECOGNITION

Mats Blomberg
Department of Speech Communication and Music Acoustics
Royal Institute of Technology
Stockholm

ABSTRACT

A phonetic recognition system based on synthetic phoneme prototypes is described. The phoneme templates are specified in terms of synthesis parameters. They are transformed to the spectral domain in the form of a 16 channel filter bank section, the same way as the incoming speech is analysed. The vocabulary and grammar is described in a finite state network where each node contains a phoneme and its context-independent frequency spectrum. The recognition process is to find the phoneme sequence in the network that gives the lowest spectral distance to the input utterance. Preliminary results are given for recognition of connected Swedish digits. Future improvements include context-dependent phoneme spectra and speaker voice source adaptation. Another application for the program is automatic time alignment of phonetic transcription to speech data bases.

INTRODUCTION

Speech recognition of large vocabularies and speaker independent recognition are normally implemented using large amounts of training data. There is a wish to decrease the time spent during adaptation of a system to each speaker for speaker adaptive systems. There is also a need to decrease the very large number of speakers required to achieve high accuracy speaker independent recognition. Since it is unreasonable to train a system by repeating each word of a vocabulary consisting of several tens of thousands of words, training is often done on smaller units, like syllables, phonemes or allophones. Still, there is considerable amount of training needed to describe every unit in all its possible contexts. In this paper, the approach is to represent acoustic properties of a phoneme inventory by means of explicit formulation of existing phonetic-acoustic knowledge. In this fashion the need for training can be considerably reduced if not eliminated.

The phonetic-acoustic knowledge can be formulated in different ways. In the bottom-up fashion, rules are written to classify a part of the speech signal if certain conditions are present. Expert systems for speech recognition are normally designed in this way. A problem with this method is that many rules are needed to cover all possible variation of the realisation of a phoneme due to phonetic context, reduction, etc.

Another way to express the knowledge is in the form of a speech production system. This is a top-down approach. A hypothesis is generated at a high level of a recognition system. Rules are then used to generate a synthetic version of the part of the utterance. The hypothesis is verified or rejected at the acoustic level. Each individual rule can describe a local effect. The results of different rules are combined to a single output. If the rules are independent, the number of rules can be lower than in the bottom-up case. The top-down approach is chosen in this report.

A question is if the current knowledge is enough to yield sufficiently high recognition accuracy. The quality of speech generated by synthesis-by-rule systems has reached a point where it is well understood even by naive listeners.

There is still a lot of improvements to be made though, before synthetic speech comes close to the quality of natural speech. It is therefore not expected that the recognition results of recognition-by-synthesis systems will be better than those which use natural speech for reference generation.

RECOGNITION SYSTEM OVERVIEW

In previous reports we have used a synthesis-by-rule system to build word templates for recognition by a dynamic-time-warping based isolated-speech recognition system (2). In this paper, we generate a library of synthetic phoneme spectra instead of whole word templates. The recognition algorithm is also changed. The vocabulary and syntax are compiled into a finite-state network with each node represented by a phoneme and its frequency spectrum. The recognition process uses a dynamic programming technique to find the phoneme sequence in the network that gives the lowest spectral distance to the input utterance. One difference to the work in (2) is that optional pronunciation of each word is allowed, which should improve recognition accuracy. Another difference is that, in this report, the phonemes are treated mainly as stationary segments, disregarding coarticulation effects. This should lower the accuracy. In future work, we will integrate the two features and thus hope to improve the recognition rate.

Since also the time position is computed of each phoneme in the selected phoneme string, it is possible to use the program for labelling of large speech data bases. In that case the phonetic network consists of the correct phoneme string with optional pronunciation alternatives.

ACOUSTIC-PHONETIC REPRESENTATION

The phonemes in the reference library are described in terms of synthesis control parameters. For vowels these are the frequencies and bandwidths of the lowest six formants and the voice source parameters. The frequencies of the lowest four formants were taken from male speaker data in (3) for vowels in stressed position in sentences. Data for Swedish consonants could not be found, so it was computed using an analysis-by-synthesis technique for one male speaker. The same speaker was later used in the recognition experiment described in this report. Nasalized vowels and nasal consonants sounds are specified not only by six formants, but also by two zeroes. Unvoiced fricatives and plosive bursts have two poles and one zero.

There is an important argument for storing synthesis control parameter in the phoneme reference library while doing the matching in the spectral domain. Application of coarticulation and reduction rules is easier on the control parameters than at the acoustic level. Although these rules mainly concern shifts of formant and antiformant frequency values, we don't need to perform an error prone tracking of spectral poles and zeroes on the incoming speech wave. Instead, the modified control parameters are transformed to the spectral domain, where the matching is made.

The main argument against using the control parameters is that the production model used for speech synthesis is a simplification and is not capable of making a perfect copy of natural speech. New and better production models will develop, however. They will be easy to incorporate into the system.

It will also be difficult to make an automatic adaptation of the phoneme library to a new speaker, since it requires a transformation of spectral data to control parameter values. If the identity of the adaptation material is known, though, the search space for the parameters can be considerably reduced, which will minimise the risk of tracking errors.

SPECTRAL ANALYSIS

The speech signal is recorded using a sampling frequency of 16 kHz. The incoming speech is transformed to the spectral domain by an FFT procedure using an analysis window of 25 ms and a frame rate of 10 ms. The frequency range from 200 to 5000 Hz is divided into 16 channels, linearly separated on a Bark scale.

The synthetic spectra are generated by computing a transfer function from the pole-zero specification of the phonemes. For voiced sounds, a voice source is superposed, using the model of Ananthapadmanabha (4). For unvoiced sounds, a white noise source is assumed. The synthetic spectra are transformed to the same representation as that of the incoming speech.

RECOGNITION EXPERIMENT

At this stage, a preliminary recognition experiment has been performed. The task was to recognise utterances consisting of 3 connected Swedish digits. One male speaker recorded 100 tokens of 3-digit strings in a sound-proof booth. The word and string recognition rates were 88 and 63 % respectively. A parametric display of a recognised utterance is shown in figure 1.

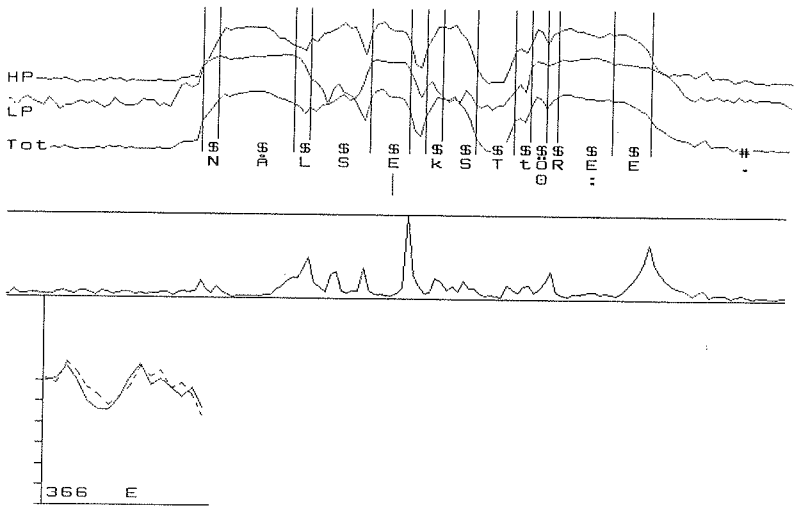

Figure 1. Display of a recognised utterance. The identity is 063 with pseudophonetic transcription /nålsekstre:/. The three intensity parameters are from bottom to top overall intensity, low pass 400 Hz and high pass 500 Hz. The marker below the phonetic symbol /E/ shows the time point of the spectrum section, drawn in solid line. The dashed line is the reference spectrum for /E/. The curve below the intensity parameters is a spectral distance curve, showing the acoustic distance between the prototype spectra of the labeled phonemes and the input spectrum at each time point.

A confusion matrix is given in table 1. A large proportion of the errors were confusion of the digits 3 ('tre') and 5 ('fem') with 7 ('sju').

The error probability seems to be dependent on the position of the digit in the utterance. The error rate for the last digit in the string was twice as high as for the first or second positions. This might be explained by the subglottal pressure drop at the end of the utterance, which causes an amplitude decrease. Also, the glottis opens at the end of the utterance and the voice source spectrum gets a low frequency bias. Furthermore, the duration increase of the final syllable might be responsible for some errors, although the durational constraints for the phonemes are quite wide.

	Recognised digit									
	0	1	2	3	4	5	6	7	8	9
Correct										
0	29	.	.	1	.	.	.	.	.	.
1	.	27	.	.	1	.	1	.	1	.
2	.	.	30	.	.	.	.	.	.	.
3	.	.	3	19	.	.	.	7	.	1
4	3	.	2	.	22	1	.	1	.	1
5	.	.	.	.	.	22	.	8	.	.
6	.	.	.	.	.	1	29	.	.	.
7	.	.	.	.	.	.	.	30	.	.
8	1	.	1	.	.	.	.	.	28	.
9	.	.	.	.	.	.	.	.	.	30

Table 1. Confusion matrix between individual digits.

DISCUSSION

The results presented are encouraging. Many errors could be interpreted in a phonetic manner and it is then possible to correct them by including the appropriate knowledge into the system. Especially, better treatment of voice source dynamics and transitionary parts at phoneme boundaries are believed to increase the accuracy substantially. Future work will also take care of coarticulation effects and speaker adaptation.

REFERENCES

(1) M. Blomberg and K. Elenius: "Time Alignment of Speech to a Phonetic Transcription," STL-QPSR 2/85, KTH,Stockholm 1985.

(2) M. Blomberg, Rolf Carlson, Kjell Elenius, B. Granström: "Speech Recognition Based on a Text-to-Speech Synthesis System", Proc. of the European Conference on Speech Technology, Edinburgh 1987.

(3) U. Stålhammar, I. Karlsson, G. Fant: "Contextual effects on Vowel Nuclei", STL-QPSR 4/1973, KTH, Stockholm, 1973.

(4) T.V. Ananthapadmanabha,: "Acoustic Analysis of Voice Source Dynamics",STL-QPSR 2-3/1984, KTH, Stockholm, 1984.