Training Speech Synthesis Parameters of Allophones for Speech Recognition

Mats Blomberg
Dept of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

ABSTRACT

A technique for training a speech recognition system at a production parametric level is described. The approach offers potential advantages in the form of small training corpora and fast speaker adaptation. Triphones that have not occurred in the training data can be generated by concatenation and parametric interpolation of diphones or context-free phones. The triphones are represented by a piece-wise linear approximation of the production parameters. For recognition, these are converted to subphone spectral state sequences. A 97.6% connected-digit recognition rate has been achieved when training the system on one male speaker and performing recognition on 6 other male speakers. In preliminary experiments with generation of unseen triphones, the performance is still slightly lower compared to using seen diphones and context-free phones. Experiments with fast speaker adaptation is also going on. Resynthesis of speech by concatenating triphones has been used to verify the quality of the triphone library.

INTRODUCTION

Speech recognition is commonly performed using parameters that describe the spectral shape of the speech signal in an effective way, such as cepstral parameters. An alternative approach is to use parameters that are more closely related to the human speech production mechanism. This representation offers several potential advantages. The relations between different phonemes are more straight-forward. As a consequence, less training data would be required for training or adaptation to a new speaker. In previous work (Blomberg, 1993), we have used a system where the phonemes were described in terms of values of control parameters to a cascade formant speech synthesiser. For recognition, the parameters were used to compute corresponding speech spectra. Reference phoneme prototypes were determined using a manual analysis-by-synthesis procedure. Due to the labour-intensiveness of the technique, only context-independent phoneme data for one male speaker was collected. In order to train the much larger number of context-dependent allophones, it is necessary to use automatic methods. This report describes the continuation of that work into automatic training of a context-dependent phone library.

SYNTHESIS MODEL

In this report, some further modifications have been made to the synthesiser in order to enable better spectral match to natural speech. In the original model, there is a separate parameter that controls the amplitude in the nasal branch. We have replaced this amplitude control with a nasalisation parameter that controls the proportion of the voice source energy that goes into the vocal tract and the nasal branch. A single parameter is now sufficient for the variation of the degree of nasalisation of a vowel from no nasalisation to a nasal consonant with no oral output. In the fricative branch, the two lowest poles have the same frequencies as formants number two and three in the vocal branch. This is intended to give formant continuity in fricative consonants in intervocalic position.

CONTEXT-DEPENDENT PHONE REPRESENTATION

The prototype units are context-dependent phones. In this report, they are denoted context-free phones, diphones or triphones if they have specified neighbouring phones at zero, one or two sides, respectively. Diphones are left- or right-dependent with respect to which side of the middle phone that is connected to a specified phone. The phones are described at the parameter level by connected piece-wise linear segments. Each parameter in the phone is described by an initial, a medial and a final line segment. The time points of the edges are separate for each individual parameter. For the purpose of recognition, each phone is divided into a sequence of subphones, each represented by a spectral shape.

TRAINING

The control parameter line segments are tracked using an analysis-by-synthesis technique and a line approximation algorithm. In the analysis-by-synthesis procedure, parameters are extracted for each frame of an utterance without regard to the neighbouring frames. The parameter tracking is guided by a manually corrected, automatically aligned phonetic transcription of the utterance. Predicted values for each time frame are determined from library look-up of context-free values of the aligned phoneme modified by coarticulation and reduction rules (Blomberg, 1993). A range around the predicted value for each parameter is searched to minimise the spectral distance between the generated synthetic and the natural speech spectra. The selection of a tracked parameter vector to a corresponding spectral input section is performed by minimisation of the spectral distance in a 16-channel Bark filterbank, including the spectral first and second frequency derivatives. A three-level iterative search loop is inspired by Carlson and Glass (1991).

In a second step, the parameter curves are approximated by line segments. Line approximation is a compact and efficient way for describing the time function of the parameters in a phone and is an established technique in practical synthesis systems (Carlson and Granström, 1990). We use a dynamic programming technique to find the optimal line endpoints.

Due to difficulties in parameter tracking, the production parameters are not usable directly for speech recognition. We need to transform the line representation of a phone into a sequence of spectral states. Also in this case, dynamic programming is used to find the best segmentation. The number of subphone states of a phone depends on the spectral variation within the phone.

Benefits for training and adaptation

As mentioned above, the chosen parametric representation is particularly appropriate for fast adaptation to a new speaker. Certain parameter values may be regarded as common

to a whole phoneme group or are closely related within the group. One observation of such a parameter value may be generalised to the whole group. This technique may be used for voice adaptation and vowel normalisation. Another possibility is to use the deviation of an observed phoneme from the corresponding reference triphone to adapt all units in the reference library that contain this phoneme. A third technique is to concatenate trained phones or diphones into triphones. This is done using linear interpolation between the parameter line end-points in adjacent units. By doing this, the wide distribution of both sides of the context-independent phones and the non-specified side of the diphones is replaced by a much more narrow distribution of a synthesised triphone with defined phonetic context at both sides. An important requirement for this technique is that the parameters are suitable for linear interpolation. Speech production parameters should perform better in this respect than spectral amplitudes.

RESYNTHESIS

The use of production parameters for describing the acoustic properties of the phonetic inventory offers the possibility to perform resynthesis of speech from a phoneme transcription. The purpose of the resynthesis experiments in this report is mainly to verify the quality of the trained phoneme library and the function of generation of unseen phoneme sequences.

The synthesis model is implemented in the digital signal processing software package "Aladdin" (Neovius and Ternström, 1992). Figure 1 shows a spectrogram of a synthetic utterance based on concatenated diphones and triphones along with a natural utterance of the same identity. Informal listening experiments reveal that the speech produced is intelligible, however far from natural. One major limitation is the lack of fundamental frequency information. We have used a linearly descending pitch throughout the utterance.

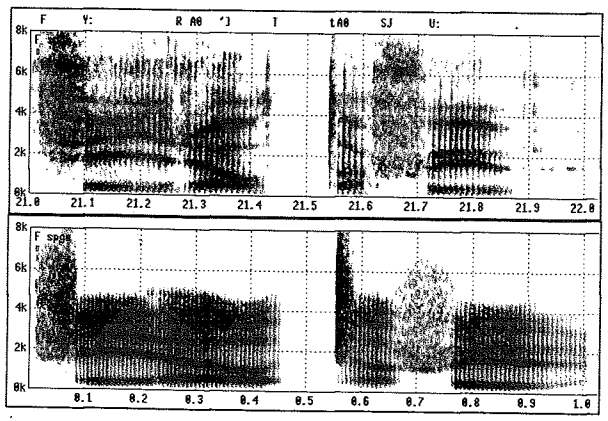

Figure 1. *Spectrogram displays of a natural (top) and a synthesised (bottom) utterance from concatenated context-dependent phones. The utterance identity is "fyra åtta sju" (four eight seven).*

RECOGNITION EXPERIMENTS

The recogniser used in the experiments is described by Blomberg (1993). It features Viterbi search in a finite state network and performs dynamic source adaptation to voice fluctuations. We are using the same corpus of three connected digits spoken by seven male speakers as in that report. One speaker is used for training and recognition is performed on the other 6 speakers. We are comparing the performance with the use of an extended triphone library where word boundary triphones are generated by concatenation and interpolation from diphones and context-free phones. We also try to adapt the triphone library to new speakers using a small number of training digit strings, where each digit is pronounced at least once.

RESULTS AND DISCUSSION

The digit recognition performance in the base-line setup is 97.6%. In preliminary experiments, the error rate drops to 97.2% when concatenated word boundary triphones are used. Possibly, this is caused by distorsion in the transition intervals of these units.

The speaker adaptation experiments are still very preliminary. The vocabulary used is so small that its quite feasible for a user to train all words and the advantage of adaptation is limited. The benefits in this case would be to enable the use of cross-word triphones and a larger number of observations for the estimation of variation.

Several components still have to be improved in the system. Higher spectral resolution and more advanced statistical analysis tools are required. The parameter tracking as well as the line approximation techniques need to be improved. However, the performance we have reached is promising and we hope that the proposed fast adaptation technique will increase the usability of speech recognition in a multi-speaker environment.

ACKNOWLEDGEMENTS

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

REFERENCES

Blomberg, M. (1993): "Synthetic phoneme prototypes and dynamic voice source adaptation in speech recognition," QPSR 4/1993, Dept. of Speech Communication and Music Acoustics, KTH, pp 97-140.

Carlson, R. and Glass, J. (1992): "Vowel classification based on analysis-by-synthesis," Proc. ICSLP 92, Banff, Canada, pp 575-578.

Neovius, L. and Ternström, S. (1992): "An object-oriented approach to real-time signal processing and its application to voice synthesis - a preliminary study," FONETIK'92, Papers from the Sixth Swedish Phonetics Conference (Ed.: Huber, D.), Technical Report no 10, May 1992, Department of Information Theory, Chalmers University of Technology, Göteborg, Sweden, pp 95-98.

Carlson, R., Granström, B. and Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications," in *Speech, Hearing and Language Processing*, Vol 1 (Ed: W. A. Ainsworth), JAI Press Ltd, London, pp 269-296.