MODELLING ARTICULATORY INTER-TIMING VARIATION IN A SPEECH RECOGNITION SYSTEM

M. Blomberg

Department of Speech Communication and Music Acoustics, KTH, Stockholm

ABSTRACT

A technique is described that automatically predicts certain cases of pronunciation alternatives. The method utilises the fact that differing realisation of an utterance often depends on variation in the synchrony between two or more simultaneous articulatory gestures. The technique has been implemented in a recognition system based on synthetic generation of reference templates. Varying delay values have been systematically generated by the speech production system. In a pilot experiment, the recogniser behaviour was examined for varying time position of the devoicing of utterance-final vowels.

1. INTRODUCTION

As is well known, the production of speech is a highly complex process that involves the control of several articulatory gestures for realizing the intended sound sequence. Different physiological, psychological and environmental factors contribute in creating variability in the pronunciation of an utterance. It is essential for a recogniser to model this variability in an appropriate way. In this report, we will discuss variability in the time synchronisation between different articulators. We will give an example of this effect, discuss consequences for speech recognition systems and suggest a new method for dealing with this type of variability.

A transition from one phoneme to a following one often involves simultaneous movements of more than one articulator. Details of the acoustic realization depends among other things on timing differences between these articulators.

An example of a phoneme boundary where two separate gestures are active is shown in figure 1. The figure shows spectrograms of the Swedish word 'tre', (English: three) spoken by two male speakers. The phonetic transcription is [tre:]. The end of the phrase-final vowel changes gradually towards a neutral vowel, similarly for both speakers. The point of devoicing is different, though. Speaker 1 keeps a steady voicing throughout the neutralisation gesture, whilst speaker 2 aspirates the last part of the vowel. An attempt to align the aspirated vowel portion of speaker 2 to the last part of the vowel for speaker 1 would result in a large spectral error. The earlier point of devoicing for speaker 2 causes a great spectral distortion, which will cause problems for most recognition systems.

An early opening of the vocal folds in this example shortens the voiced part of the vowel and prolongs the duration of the preaspirative segment. Also, the spectral properties of the aspiration will be changed. The tongue will have moved a shorter distance towards its target at the start of aspiration and the spectral shape immediately after the aspiration onset will be quite different compared to the same point in a boundary with a late opening.

Other examples of overlapping articulatory movements are velar opening during vowels before nasals and change of place-of-articulation between adjacent consonants. In the latter case, it often happens that the release from the first consonant precedes the closure of the second one, which will cause a short vocalic segment to occur. If the release occurs after the closure, there will be no such segment.

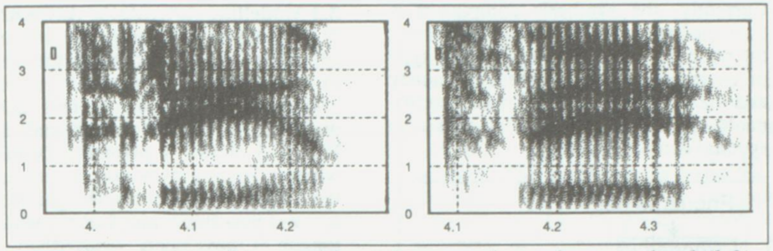

Figure 1. Spectrograms of the word [tre:] for two male speakers, speaker 1 (left) and speaker 2 (right).

2. RECOGNITION APPROACH

The acoustic-phonetic decision part of most existing systems are based on spectral matching without taking into consideration the underlying production parameters. Common techniques like dynamic time-warping and Hidden Markov Modelling [5] are able to compensate for a non-linear tempo variation between two utterances but they do not handle timing asynchrony between the production parameters. Stretching and compression of the time scale of the speech signal implies a uniform time transformation of the underlying articulatory parameters. In these systems, the effect will be reflected by large spectral variation at the phoneme boundaries.

A common way to represent pronunciation alternatives is to use context-sensitive optional rules, formulated by a human phonetic expert, [3] and [4]. The rules operate on the input phoneme string and produce several phonetic output strings. However, they mostly use a qualitative description of the effect of varying delay between the articulators. As discussed above, we also need a quantitative description. This requires a description of the phonetic elements in terms of production parameters.

The optional rules can be modified so that they generate a set of pronunciation alternatives at every phoneme boundary. Within the set, the delay between some of the parameters are varied in a systematic fashion. In this way, a quantitative, as well as a qualitative, description of the articulator asynchrony effect is obtained.

The parameter tracking problem can be avoided by using a synthesis technique for producing reference templates, as mentioned in [1]. In this way, knowledge about the behaviour of different parameters can be utilized, without the need of tracking them from the speech signal. Instead, their predicted values can be used for generating corresponding frequency spectra, and the recognition matching would be performed in the spectral domain.

3. SYSTEM DESCRIPTION

3.1 Recognition System

The recognition system used for this experiment has been described in [1] and [2]. It uses dynamic programming for finding the path through a finite-state network of subphonemic spectra that minimises the spectral distance to a spoken utterance. During the matching of an utterance, an adaptation procedure dynamically normalizes for differences in the voice source excitation function. The subphoneme spectra have not been created by training, as in the majority of current recognition systems, but by a speech production algorithm described below.

3.2 Reference Data Generation

Figure 2 shows a block diagram of the reference template generation component. It is very similar to a speech synthesis system. Its main difference from such a system is that the output consists of spectral sections instead of a speech signal and that the input phonetic description is a network of optional pronunciation alternatives as opposed to a string in the speech synthesis case. The net can describe a single word or the language of a complete recognition task. Currently, the synthesis component is formant-based. In the phoneme library, the phonemes are specified by their type of excitation and by formant frequency and bandwidth values. Certain consonants, like nasals and fricatives also have spectral zeros specified.

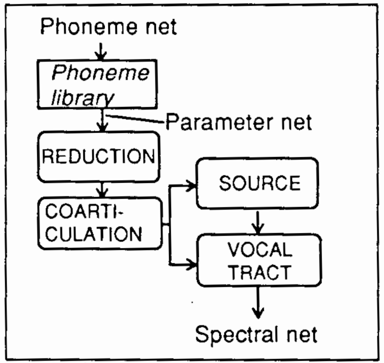

Figure 2. Block diagram of the speech production system used for reference template generation

The reduction and coarticulation components modify and expand the input phonetic network. The reduction part adjusts the targets of vowels depending on their assigned stress and their context. Since there may be more than one left or right neighbouring phoneme, it is necessary to create copies of the vowel node for all the possible contexts before applying context-sensitive formant adjustment rules.

The coarticulation component handles the transient portions at the boundaries between two phonemes. Several subphonemic states are inserted between the steady-state parts. The production parameters in these states are interpolated from the surrounding steady-state values. The number of subphonemic states in a boundary is determined by the spectral distance between the two phonemes.

The final step is to compute prototype spectra from the production parameters at each state. This is done by logarithmic addition of an excitation spectrum and transfer functions of individual formants.

3.3 Modelling Articulator Asynchrony

For ease of illustration, we will in the following example consider the change of only two parameters; the others are assumed to be constant. This can be displayed in a two-dimensional array. Figure 3 shows a phoneme boundary, where a voicing transition occurs during the tongue movement when going from a vowel into an unvoiced consonant. The tongue movement, described by interpolated formant values, and the voicing transition are represented in the horizontal and the vertical axes, respectively. They are quantised into a low number of steps. The upper and lower horizontal lines represent the tongue movement during voicing and aspiration, respectively. Different delays of voicing offset relative to the start of the tongue movement are represented by vertical lines at varying horizontal positions. The duration of the voicing transition is considered to be short compared to the tongue movement, and therefore there is no need for diagonal connections in the lattice.

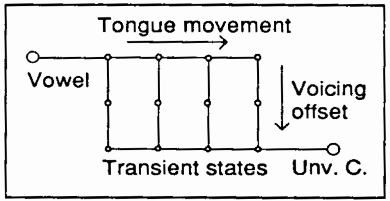

Figure 3. A sub-phoneme lattice representing varying parameter inter-timing in the transition between a vowel and an unvoiced consonant.

4. PILOT EXPERIMENT

Instead of running a complete recognition experiment, we studied the chosen method's ability to align the speech signal to a phonetic transcription of the utterance. The two utterances shown in figure 1 were used for testing the method.

To represent the possibility of devoicing of the final vowel, we implemented a subphoneme lattice similar to figure 3, where the consonant in this case is the phrase-end symbol. This symbol is marked in the phoneme library as unvoiced and having neutral formant targets.

The speech signal was analysed by a computer implemented 16-channel Bark-scale filter bank covering a frequency range from 0.2 to 6 kHz. The frame interval was 10 ms and the integration time was 25.6 ms.

5. RESULT

The paths through the network for the two utterances are shown in figure 4. The predicted, interpolated value of the second formant is displayed for every subphoneme. The path for speaker 1 shows a voicing offset at a later stage of formant transition than that of speaker 2. This conforms well with the spectrogram displays in figure 1.

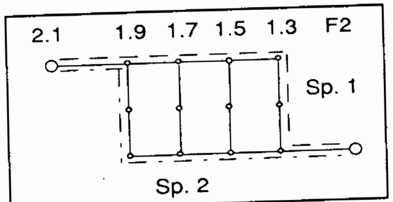

Figure 4. Results of alignment of the last part of the phrase-final [e:] The paths for speakers 1 and 2 are displayed. State values of the second formant are shown.

The accumulated spectral error over the phoneme boundary was also measured. It was compared with the errors using a fixed-delay subphoneme string, having early or late voice offset. The results in table 1 show that the proposed method works well for both speakers, whereas each of the preset delay values gave low error for one speaker only.

Table 1. Accumulated spectral error over the final transition interval of the two vowels in figure 1. Three allowed positions of voicing offset relative to the start of the formant transitions.

Devoicing	Speaker 1	Speaker 2
Early	165	110
Late	133	160
Variable	133	111

6. CONCLUSIONS

The experiment in this report just serves as an illustration of the ability of the presented technique to compensate for articulator asynchrony. Further experiments in a complete recognition task will show the benefit of the proposed method. The technique is expected to increase the robustness of a recognizer, since it is able to predict infrequent manners of speaking that might not occur in a training material.

Much work remains to describe other phoneme boundaries. Our knowledge about their realisation is still incomplete in many ways. Further improvement is dependent on the development of better speech production models. Especially, use of an articulatory model would give a straightforward description of several boundaries, e.g. adjacent consonants. We believe that implementing such a model in the described recognition system would be an important step towards further performance increase.

7. ACKNOWLEDGEMENT

This project was supported by the Swedish Board for Technical Development.

8. REFERENCES

[1] BLOMBERG, M. (1989) "Synthetic phoneme prototypes in a connected-word speech recognition system", ,*Proc. ICASSP 90*, 687-690.

[2] BLOMBERG, M. (1989) "Voice source adaptation of synthetic phoneme spectra in speech recognition", *Eurospeech 89, Vol 2*, 621-624.

[3] COHEN, M., MURVEIT, H., BERNSTEIN, J., PRICE, P., & WEINTRAUB, M. (1990), "The DECIPHER speech recognition system", *Proc. ICASSP 90*, 77-80.

[4] COHEN, P. & MERCER, R. (1975), "The phonological component of an automatic speech-recognition system", *Speech Recognition*, R. Reddy, ed., Academic Press, New York, 275-320.

[5] MARIANI, J. (1989), "Recent advances in speech processing", *Proc. ICASSP 89*, 429 - 439.