AUTOMATIC DETECTION OF THE PHONEME BOUNDARIES IN AN UTTERANCE GIVEN ITS PHONETIC TRANSCRIPTION

Mats Blomberg
Dept. of Speech Communication and Music Acoustics, KTH

INTRODUCTION

Large quantities of speech data are needed to improve the performance of speech recognition and speech synthesis systems as well as for purposes of basic phonetic research. To be useful, the speech material has to be segmented and labelled at the phoneme level. For small amounts of data, this is often done manually by listening to the speech and by visual inspection of spectrograms and displays of the speech signal. Since the positioning of each label requires considerable attention by a skilled phonetician, it is very time consuming for larger amounts, and it is then feasible to replace by an automatic procedure. We have previously worked with methods for time alignment of an utterance to its phonetic transcription (Blomberg & Elenius, 1985). This paper describes further work on such a system. Its performance is evaluated by comparing its labels to labels set manually on a speech corpus consisting of 138 sentences spoken by one male speaker.

PHONETIC TRANSCRIPTION GENERATION

In the KTH speech data base system (Carlson et al, 1989), there are two methods for segmentation and labelling of recorded speech data. One is manual, where the operator uses listening and visual inspection of the speech signal and of spectrograms. The second method is the one described in this paper. Both methods use the grapheme-to-phoneme part of a text-to-speech system (Carlson et al, 1982) to generate an initial phonetic transcription of a given utterance. The transcription sometimes differs from the actual pronunciation, and so it is important to verify its correctness before trying to align it with the utterance. If necessary, the output string can be corrected manually. Another possibility is to generate all possible pronunciations of an utterance by producing several optional forms in the phonetic transcriptions and then let the computer select the phoneme sequence that gives the best correspondence to the utterance. The pronunciation alternatives can be given manually or they could be generated from a base form by a set of rules. In the present grapheme-to-phoneme rule package, only the base form is produced, and the alternatives are inserted by hand, but it would be possible to develop rules that also generates optional forms.

PRINCIPLE OF ALIGNMENT

The time alignment system described in this paper is based on a phonetic speech recognition program, described in more detail by Blomberg (1989). Its most important distinguishing feature is the use of synthetic phoneme reference spectra. The transcription that is to be aligned is transformed to a sequence of frequency spectra in the same manner as in a speech synthesis program, although the actual speech signal is not generated. The argument for using synthetic references is that the training session for building reference phoneme prototypes can be avoided. To account for varying
voices, the spectra are dynamically adapted to the speaker's voice during the alignment process. This has been shown to be very helpful for speaker-independent recognition (See Blomberg, 1989). Adaptation of the phoneme prototypes to an individual speaker is also possible but mostly not necessary.

The boundary positions of the phonemes in the utterance are determined using a dynamic programming procedure, which computes the optimal time alignment between the natural utterance and a synthetic version of the same identity. If optional pronunciation forms are given, it automatically selects those alternatives that give the lowest overall spectral error. The output is the selected phoneme transcription and the time positions of the segment boundaries.

GENERATION OF A REFERENCE UTTERANCE

In the system, the phonemes of a particular language are specified by their synthesis parameter values in a separate data file. It is possible to have separate data files for different languages and regional accents and also for individual speakers. For vowels and several voiced consonants, these parameters are formant frequencies and bandwidths. Unvoiced fricatives and nasal consonants have both spectral poles and zeroes specified. The values of the parameters can be adapted to new speakers by a training procedure, starting from default values.

The conversion from a phoneme string to synthesis control parameter values is at present not done by the text-to-speech system, but by a much less sophisticated algorithm. The produced 'speech' of this system will not reach as high quality as the text-to-speech system. However, the purposes are different. We have considered it important to handle optional pronunciation, which is not dealt with in the synthesis system. Another reason is that for some phonemes, the spectral correspondence between natural and synthetic speech is not very good and we have tried modifications of the synthesis model for these phonemes. Future versions of the two systems may remove these limitations and enable a more complete use of the rule-synthesis power.

When a phoneme string is entered, the context-independent phoneme parameter values are modified to take account for reduction and coarticulation effects. Reduction rules move the vowel targets towards neutral values with an amount dependent on the assigned stress values. Coarticulation effects are modelled by splitting the phonemes into one or several initial parts, one steady state part and one or several final parts. The steady-state values are modified by the adjacent phonemes for unstressed, short vowels. The formant frequencies of the transition segments are then interpolated from the target values of the surrounding phonemes.

The spectral transfer function of each subphoneme is then derived from the pole-zero specification. A voiced, unvoiced or a mixed excitation function is added to give a reference frequency spectrum. This is transformed to a 16-channel Bark filter bank representation with the same characteristics as the one which is used for the analysis of the speech signal. In this way, a sequence of filter bank spectra representing a synthetic utterance of the same identity as the natural speech is generated.

SIGNAL PROCESSING

The digitized natural speech signal is analysed by an FFT procedure and transformed into a 16 channel Bark-scale filter bank representation, covering a frequency band from 200 to 6000 Hz. Simple models of lateral inhibition and forward masking are implemented in the spectral output. For the experiment, the analysis window of the FFT was 12 ms and the frame interval was 5 ms.
EXPERIMENT

In an evaluation experiment, recordings from one male speaker with a moderate Stockholm accent have been labelled both manually and by the automatic method. The material consisted of 138 sentences, stored as separate files. The sentence length varied between 4 and 13 words with an average of about 8 words per sentence. The phonetic transcription was first generated from text using the grapheme-to-phoneme part of the text-to-speech system. It was then manually corrected by comparisons to the recorded speech. A phonetically skilled person determined the time positions of the phoneme boundaries by looking at displays of spectrograms and the speech signal.

The same transcriptions were used by the automatic method. The possibility of allowing several pronunciation alternatives was not used in this experiment, since it would make the evaluation procedure more difficult. The phoneme prototypes were adjusted to average values of Stockholm accent for Swedish male speakers. No speaker adaptation was performed.

RESULTS

Errors made by the automatic alignment technique may be classified into two categories. If the misalignment is less than half the phoneme duration, the error can be interpreted as a result of lack in precision. The estimated phoneme midpoint will still be positioned inside the phoneme, which is important when collecting spectral measurements of phonemes. Larger errors are often caused by a confusion of the searched phoneme boundary with a similar transition at another place in the utterance. For example, a typical error is that a vowel-consonant boundary is moved one syllable forward or backward.

A histogram and cumulative distribution of the deviation between the manual and the automatic method is shown in figure 1.

![Figure 1](image)

**Figure 1.** Histogram (a) and cumulative distribution (b) of time alignment error.

In more than 90% of the boundaries, the deviation is 50 ms or less. The average deviation in the material was 17 ms. It is encouraging that only around 2% of the labels are misplaced by more than 100 ms. Evidently, errors due to boundary confusion is much less frequent than those caused by low precision. Phonemes frequently involved
in longer misalignments are /r/ and /l/. One reason for this is obviously that they exhibit high acoustic variability among different contexts. Their behaviour is regular, though, and could be handled during the generation of the reference utterance. /r/ also contains speaker dependent allophonic differences, which may show up among a larger group of speakers. Multiple allophone templates and speaker adaptation would take care of this variation. One extreme realisation of /r/ is when it is not pronounced as a separate segment at all, but rather as a retroflex colouring of the surrounding phonemes.

The transition between two vowels and between vowels and certain consonants, like /r/ and /l/, are often slow, which makes it difficult even for manual labelling to make a consistent positioning. Differences of this kind between the automatic and the manual outputs should be given a low weight compared to more distinct boundaries.

SUMMARY

More speakers have to be used to get a reliable measure of the technique. Anyway, the results with one male speaker show that it can be an efficient way of labelling a large speech corpus. Different applications have different demands on precision and it depends on the purpose of the data if the accuracy is sufficient. Analysis of spectral properties in steady state regions is quite insensitive to slight deviations of phoneme boundaries. On the other hand, measurement on segment duration needs higher precision than what is achieved in this report. The automatic alignment process could therefore be followed by manual correction. There would still be a significant reduction of time compared to completely manual alignment.

Further improvements could be achieved by modelling spectral and durational distribution, by employing speaker adaptation, and by developing better rules for generating the synthetic prototype spectra.

ACKNOWLEDGEMENT

This work has been financed by the Swedish Board of Technical Development.

REFERENCES


