Creating unseen triphones by phone concatenation in the spectral, cepstral and formant domains

Mats Blomberg
Dept. of Speech, Music and Hearing, KTH, Stockholm

Abstract

A technique for predicting triphones by concatenation of diphones or monophones is studied. The models are connected using linear interpolation of parameter trajectories. Previous work on formant parameters is extended to filter channels and cepstrum coefficients. Preliminary results indicate that the proposed technique works well also in these domains. In both cases, the approximation error is lower than that of the formant domain. This may to a great extent be explained by formant tracking errors.

Introduction

Large vocabulary recognition requires accurate modeling of the acoustic properties of the phoneme inventory. Triphones are commonly used for this purpose, since they account for the coarticulation between adjacent phones. However, the number of triphones in a language is high and large speech corpora are required for the training of their acoustic properties. A common back-off strategy for missing or infrequent triphones is to use the shorter units, diphones and monophones, that are fewer and occur more frequently in a training corpus. A problem with these units is, however, that one or both boundaries is context-independent and, thus, have larger variation and accordingly lower phonetic discrimination.

We have previously presented a technique for predicting unseen triphones by concatenating phone models with shorter context, such as diphones and monophones (Blomberg, 1996). In this technique we combine the advantages of better trained di- and monophones with the higher phonetic discrimination of the triphone models. By using a poly-line representation of the trajectories of the used parameters, we can model the transition phase between the concatenated units with linear interpolation.

In our previous work, we studied the technique in the formant domain, using an Analysis-by-Synthesis (AbS) technique for formant tracking. The choice of this representation was made from the assumption that formant envelopes are more linear in time and therefore more suitable for linear interpolation than other spectral representation forms. The results showed that interpolation worked quite well. The average resynthesis spectral error between a training library and a cross validation library was lowered when using concatenated triphone models compared to the baseline back-off technique using a diphone or a monophone for non-frequent triphones. An algorithm for modelling the coarticulation effect between the context at the opposite sides improved the performance further. In this paper we have extended the original study to include logarithmic amplitudes of a Bark-scaled 16 channel filterbank and 16 cepstrum coefficients derived from that filterbank.

Another possibility with the line segment approach, not yet implemented, is to avoid the limitation given by the stationarity assumption in a conventional HMM recognition paradigm. In this way, frame-wise update of the reference data is possible compared to a stationary description for a whole segment.
Experiments

The performance of the different approaches has been measured by using two triphone libraries. Both have been trained using speech data collected within the Swedish dialogue project WAXHOLM (Blomberg et al., 1993). They contain the same speakers but with separate recordings. One library, the training library, is used for creating triphones from shorter units. The other library is used for cross-validation. For every triphone identity that has been observed in sufficient number in both libraries, we measure the spectral distortion of trained and created triphones in the training library against the corresponding triphone in the cross-validation library. In addition to comparison between the three parameter representations and between naturally trained triphones and concatenated units, the following aspects are currently investigated:

State vs line distortion

For the same number of segments, a line approximation representation should give lower distortion than a sequence of stationary segments. This difference should decrease with increasing number of stationary segments. A poly-line consists of two points per phone.

Observation frequency dependence

The trained triphone is the best choice if its observation frequency in the training data is high. If the frequency is too low, though, then a unit that is based on a larger number of occurrences should be used. In order to find a good threshold under which to use concatenated units, we have studied the spectral errors as a function of triphone frequency of occurrence.

Recognition accuracy

Experiments are being performed by sorting N-best recognition sentences within the WAXHOLM project. Previous work (Blomberg and Elenius, 1996) showed that the use of triphones that had been created using the proposed technique in the formant domain raised the performance somewhat. Work is currently being pursued for the filter and cepstral domains but will not be presented in this report.

Results

State vs line distortion

As shown in Figure 1, two-corners-per-phoneme approximation of filter amplitudes and cepstral coefficients approximate an utterance with higher precision than do average spectra. The average number of states in the trained library is 3.0.

![Figure 1](image-url)

Figure 1. The average distortion of an input utterance in the trained library, for different representations.
When comparing trained phone models there is no longer access to the input utterances. In this case, we use as reference the filter line representation, which according to Figure 1 has the lowest error to the real speech signal. Figure 2 shows the effect of varying number of states per phone. Line trajectories are slightly better than five states per phone.

![Figure 2](image)

Figure 2. Average distortion between triphones of the same identity in two different libraries, when the number of states per phone is varied from 1 through 5, and compared with line segments.

**Observation frequency dependence**

The results in Figure 3 show that the concatenation technique is especially valuable when the observation frequency of a triphone is low. If a natural triphone is observed in less than around 10 samples in the training data, then it is better to use a concatenated diphone pair.

![Figure 3](image)

Figure 3. The average spectral error as a function of number of observations of the triphone in the training data. Solid line: triphones concatenated from diphones. Dashed line: trained triphones. Concatenation has been performed on filter amplitudes.

**Parameter representation**

As shown in Figure 4, the concatenation/interpolation technique works well also in the spectral and the cepstral domains. Line approximated filter amplitudes and cepstral envelopes yield lower spectral distortion than formants resynthesised into spectra. The difference is judged mainly to be due to the residual spectral error in the AbS algorithm (leftmost column in Figure 1) rather than in the line approximation procedure. This error is
zero for filters and cepstra. (If a lower number of cepstral coefficients had been used we would have had a non-zero spectrum modelling error.)

Figure 4. Errors of concatenated diphones and trained triphones for the three types of spectral representation.

The low concatenation error in the filter and cepstrum domains makes the technique suitable for incorporation in practical recognition systems since the problems of formant estimation can be avoided.

Discussion
The line segment approximation of the parameter trajectories lower, as expected, the distortion compared to subphonic average spectra. Line approximation using filter amplitudes and cepstrum coefficients result in lower distortion than for formants. The reasons for this can to a large extent be attributed to formant tracking errors. Better production model and higher precision in the formant tracking algorithm should reduce the differences between the representation forms. Further work also includes recognition experiments using concatenation in the filter and cepstrum domains. The proposed technique for estimating unseen triphones should be especially valuable for small training or adaptation data sets when few natural triphones can be trained. An alternative use would be to raise the triphone coverage in an existing training speech corpus.

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

References