EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

CREATING UNSEEN TRIPHONES BY PHONE CONCATENATION IN THE SPECTRAL, CEPSTRAL AND FORMANT DOMAINS

Mats Blomberg
Dept. of Speech, Music and Hearing, KTH, Stockholm
E-mail: mats@speech.kth.se

ABSTRACT

A technique for predicting triphones by concatenation of diphone or monophone models is studied. The models are connected using linear interpolation between end-points of piece-wise linear parameter trajectories. Three types of spectral representation are compared: formants, filter amplitudes and cepstrum coefficients. The proposed technique lowers the spectral distortion of the phones for all three representations when different speakers are used for training and evaluation. The average error of the created triphones is lower in the filter and cepstrum domains than for formants. This is explained to be caused by limitations in the Analysis-by-Synthesis formant tracking algorithm. A small improvement with the proposed technique is achieved for all representations in the task of reordering N-best sentence recognition candidate lists.

10.21437/Eurospeech.1997-20

1. INTRODUCTION

Large vocabulary recognition requires accurate modelling of the acoustic properties of the phoneme inventory. Triphones are commonly used for this purpose, since they account for 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 shorter units, diphones and monophones, that occur more frequently in a training corpus. A problem with these units is, however, that at least one boundary is context-independent and, thus, they have larger variation and accordingly lower phonetic discrimination. To improve recognition accuracy it is important to model non-trained and non-frequent triphones more accurately.

We have previously presented a technique for predicting unseen triphones by concatenating phone models with shorter context, such as diphones and monophones [1]. In this technique the advantages of better trained di- and monophones and the higher phonetic discrimination of the triphone models are combined. A three-segment poly-line approximates the parameter trajectories during each phone. The corner points are individual in time for each parameter. Concatenation of two diphones with the same mid-phone identity, a diphone pair, into a triphone is performed by picking the first two line end points from the left-dependent diphone and the last two from the right-dependent one. The mid line segment is estimated using linear interpolation. The new line representation is then converted to subphone spectral states for recognition.

This is compared to a baseline technique, which selects in order a state diphone pair, a diphone or a monophone if the requested triphone has too few occurrences in the training data. The spectral states of the state diphone pair are copied from the first states of the left-dependent diphone, and from the last states of the right-dependent one. Approximately the same number of states is picked from the two diphones, which may differ in their number of states.

In previous work [1], we have 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 better suited for linear interpolation than other types of spectral representation. The results showed that interpolation worked quite well. An algorithm for modelling the coarticulation effect between the context at the opposite sides of a phone improved the performance further. As suggested [2], that work is extended in the current report to test this assumption. We include two other sets of acoustic representation: logarithmic amplitudes of a Bark-scaled 16 channel filterbank and 16 cepstrum coefficients derived from that filterbank. The filterbank covers the frequency range 200 - 6000 Hz.

The proposed technique can be applied in a more straight-forward way for filter amplitudes and cepstra since formant tracking is not performed. It is, however, uncertain if the time evolution of these parameters can be appropriately approximated with the chosen line representation. For example, higher order cepstrum coefficients tend to change less smoothly and lose some detail in the linear approximation [3].

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 system. An overview of segmental HMMs is given in [4]. A comparison of static and linear segmental HMMs is given in [3].

1.1 Biased diphone concatenation

A problem with concatenating two diphones into a triphone is the large phonetic and acoustic variation at the non-specified side of each of the diphones. It is likely that a properly selected subset of the training data for the diphone model is better in predicting the requested triphone than using all observations. We apply this idea by computing a *biased diphone*, where the individual observations are weighted according to the similarity of

the non-matching context side to the requested context. Currently, the weighting factor is based on spectral similarity between the monophone models of the non-matching and the requested phones.

2. EXPERIMENTS

The WAXHOLM speech data base [5] is used for evaluating the proposed technique. Currently, around two hours of spoken dialogues from 66 speakers; 49 male and 17 female, have been collected. Of these, 56 subjects were selected for training. The test corpus consists of 327 sentences (1672 words) spoken by 6 male and 4 female speakers, not in the training group.

The performance of the different approaches has been measured by using three triphone libraries. The first two are trained on separate halves of the training corpus. They contain the same speakers but with different utterances. One library, the training library, is used for creating triphones from shorter units. The second library is used for cross-validation. The third library is trained on the recognition test data.

2.1 Evaluation techniques

The different ways of modelling triphones are currently investigated in the following ways:

2.1.1 Acoustic representation

The three acoustic representations are compared in their precision in line approximation of an input utterance and in their ability to predict unseen triphones. For every triphone identity in the library for evaluation (cross-validation or test), we measure its spectral distortion against trained and created triphones in the training library. The distortion metric is a squared Euclidean distance between average values of time-normalised sequences of filterbank sections, into which the cepstral and formant representations are transformed.

2.1.2 Observation frequency dependence

The created triphones will be used when the number of natural observations during training is low. Hence, it is especially interesting to study their accuracy in these cases. In order to find a good threshold below which to use concatenated units, we have studied the spectral errors as a function of triphone frequency of occurrence.

2.1.3 Recognition accuracy

Experiments have been performed by rescoring N-best sentence candidate lists produced by the recogniser used within the WAXHOLM project [6].The used N-best lists contained 10 candidates on average and enabled an overall word accuracy between 49% and 87%. The average accuracy for the top candidate was 77.1%. After these N-best lists were generated, higher accuracy results have been produced in the ongoing development work

It is not certain that, out of a number of incorrectly recognised sentence candidates, the one with the fewest word errors also is the most correct with respect to its phoneme sequence. To avoid random fluctuations in the performance caused by this effect when the correct word sequence is missing in the N-best list, the correct identity was added as an extra candidate to the list. Out-of-vocabulary words were included in the lexicon for the same reason. These modifications improve the results above the performance in a normal test situation, but make it more likely that a better configuration is reflected in higher performance.

The back-off models in the concatenated triphone case are used in the following order: natural triphones, line concatenated diphone pairs, diphones and monophones. Biased diphones are not included due to the high computational requirements in their estimation. In order to be of practical use they should be computed off-line in advance of the recognition experiment.

The recogniser used in the experiments is described in [7]. The recognition algorithm performs a Viterbi search in the filterbank domain into which the formant and cepstral representations are transformed. Dynamic source adaptation is performed during the search in order to compensate for deviant voices. The system can model duration by a logarithmic Gaussian distribution or by an exponential function. To keep computational time reasonable, we have used the latter model in this report.

For the rescoring purpose, the candidates are merged into a lexical net, in which a new top candidate is searched. Non-unique context at either or both sides of a phone due to branching are made unique by splitting the phone into as many duplicate copies as given by multiplying the numbers of different preceding and following phone identities. This technique enables triphone modelling of every phone in the lexical net, including word boundaries.

3. RESULTS

3.1 Acoustic representation

As shown in Figure 1, modelling the trajectories of filter amplitude and cepstral coefficients by line segments with two corners per phone approximate an utterance with somewhat higher precision than do state average spectra with an average of 3.0 states per phone.

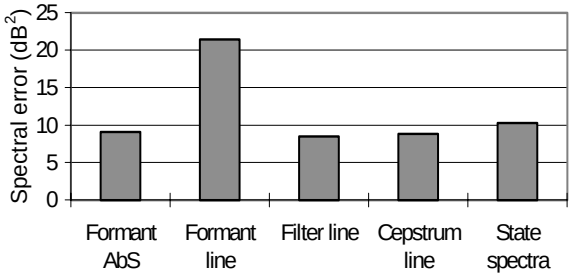

Figure 1. The average distortion of an input utterance in the training corpus for 2-corners-per-phone line approximation in the different representations. The left-most bar shows the average frame-wise AbS error.

Line approximation in the formant domain results in a considerably higher error. The difference is interpreted mainly to be due to the residual spectral error in the AbS algorithm rather than in the line approximation procedure. This error constitutes a large part of the

formant line approximation error. The distortion increase in the line approximation procedure is more close to the line errors of filters and cepstra. The frame-wise modelling error for the two latter types of representation is zero. The cepstral modelling error would be non-zero if fewer coefficients were used. This would also occur in the filter domain if the frequency resolution during evaluation were higher than that of the training library.

When comparing trained phone models there is no 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. Lines are slightly better than five states per phone.

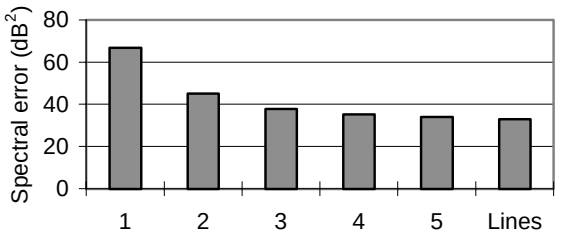

Figure 2. Average distortion between triphones of the same identity in two different libraries, as a function of number of states per phone. The right-most bar shows the distortion when a line representation is used. The results are based on filter domain approximation.

As shown in Figure 3, the concatenation/interpolation technique works well in the filter and the cepstral domains. In the cross validation data, these representations yield lower spectral distortion than formants. In the test data, the spectral error increases substantially for all representations. This effect seems to mask the differences between the representations and formants perform almost as well as filters and cepstra. In average, both diphone pairs predict the test data better than triphones. Line concatenation is better than state concatenation for all three types of representation forms.

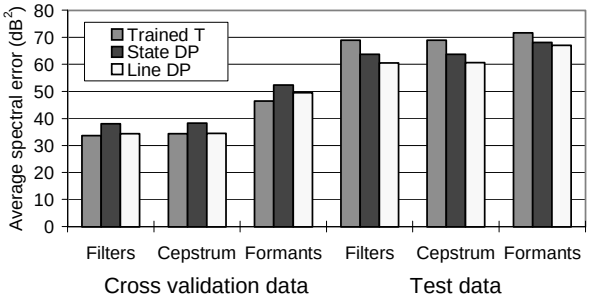

Figure 3. Observation frequency-weighted average spectral errors for concatenated diphone pairs and trained triphones. Results are shown for three types of spectral representation in the cross validation (same speakers) and the test data (different speakers). Labels: T - triphones, DP - diphone pairs.

3.2 Observation frequency dependence

The results in Figure 4 show that the concatenation technique is especially valuable when the observation frequency of a triphone is low. This is believed to be the reason for the superiority of diphone pairs to triphones in Figure 3. When the triphone is missing completely in the training corpus, the spectral errors are 174.3 and 169.6 dB^2 for state and line concatenated diphone pairs, respectively. The errors of biased diphone pairs are not shown, but are close to the two types of diphone pair. The trained triphone errors seem to be lower than the other techniques for observation frequencies above 10. This number is set as back-off threshold during the recognition experiments.

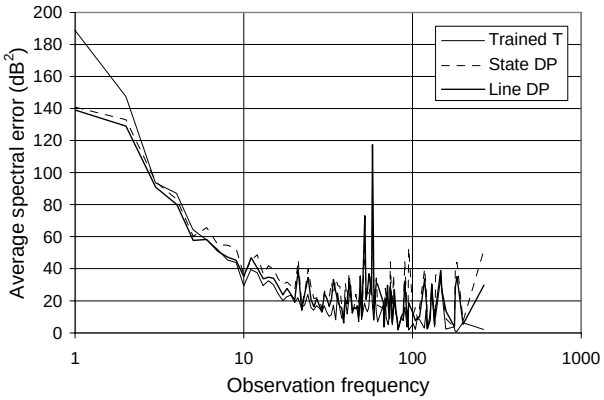

Figure 4. The average filter domain spectral error in the cross-validation library as a function of number of observations of the triphone in the training data.

3.3 Recognition accuracy

Table 1 shows results of the performed recognition tests. Splitting phones that branch in the lexical net results in increased error rate if trained phone units are used. This indicates that the back-off threshold of 10 observations is lower than optimal. In this case, the use of low-frequent triphone models instead of diphones lowers the performance. If created triphones replace the low-frequent triphones, the errors decrease to a lower value than the original configuration for all three types of representation.

Table 1. N-best reordering performance under different acoustic representation, split / no split of branching phones and trained or line concatenated triphones.

Acoustic representation	Split / no split	Trained / concatenated	Word errors (%)
Cepstrum	no split	trained	13.2
	split	trained	13.4
	split	concatenated	12.4
Formants	no split	trained	13.6
	split	trained	14.6
	split	concatenated	12.7
Filters	no split	trained	11.9
	split	trained	12.4
	split	concatenated	11.7

4. DISCUSSION

The proposed technique for creating unseen triphones work somewhat better in the filter and the cepstral representations than in the formant domain regarding spectral errors as well as recognition accuracy. Still, the difference in recognition performance is not very large. The results make the technique suitable for incorporation in practical recognition systems since the problems of formant estimation can be avoided.

The most important improvement in order to raise the performance of the formant parameters is to lower the modelling error of the Analysis-by-Synthesis algorithm. This requires a more accurate speech production model and more effective procedures to search for optimal trajectories.

Creating triphones by concatenation and line interpolation of diphones and monophones lowers the spectral errors compared to trained triphones, especially for those identities that have few observations in the training data. The errors are also somewhat lower than state-connected diphone pairs. There is still room for improvements of the procedure since the difference to the average errors of high-frequent triphones is only reduced to a small degree. The recognition performance is improved in all three types of acoustic representation, although only marginally in the filter domain.

The results suggest that there is little performance improvement to gain by mere line representation of trajectories in order to avoid the stationarity assumption in conventional HMM. The average reduction of the distortion from 3 states/phone is around 15% and from 5 states/phone there is only a 1-2% decrease. Increasing the number of states per phone could be an alternative, since the resulting distortion is almost as low and implementation is easier. However, a remaining drawback compared to segmental HMM will be the independence assumption between the states.

The metric used for comparing phone models does not account for differences in the shape of the probability density functions. In future work, simple distance between phone model averages should be replaced by measures that better accounts for these differences, such as relative entropy or mutual information.

The average spectral error on the cross validation data is substantially lower than the error on the recognition test data. This may be explained by the fact that the training data and the cross validation data are very similar. They contain the same speakers and every second utterance of each dialogue session, which also makes it likely that their word occurrences are similar. The relative inferiority of the formant representation compared to the filter and the cepstral domains is reduced in the test corpus. This indicates that sources of larger variation are involved when the conditions during training and test are different and that an acoustic representation that is better in reducing the influence of these perhaps should be preferred, even if its performance under similar training and test conditions is lower. Such aspects may still motivate the use of formants or other representations closer to the human speech production process. The high potential for speaker adaptation and normalisation of the formant representation has not been exploited in this report but should be considered in future work.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] Blomberg, M., Elenius, K.: “Creating unseen triphones by concatenation of diphones and monophones using a speech production approach,“ Proceedings of ICSLP 96, pp 2316-2319, 1996.

[2] Bridle, J.: Personal communication.

[3] Holmes, W.: “Modelling variability between and within speech segments for automatic speech recognition,” *Speech, Hearing and Language: work in progress,* Vol 9, pp 73-97, 1996.

[4] Ostendorf, M., Digalakis, V.V., Kimball, O.A. "From HMM´s to Segment Models: A Unified View of Stochastic Modelling for Speech Recognition," IEEE Transactions of Speech and Audio Processing, Vol 4, No. 5, pp 360-378, 1996.

[5] Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., Neovius, L., Nord, L. (1993): "An experimental dialogue system: Waxholm", STL-QPSR 2-3/1993, pp. 15-20.

[6] Ström, N. “Continuous speech recognition in the WAXHOLM dialogue system,” *TMH-QPSR* 4/1996, Speech, Music and Hearing, KTH, pp 67-95, 1996.

[7] Blomberg M. “Synthetic phoneme prototypes and dynamic voice source adaptation in speech recognition,” *STL-QPSR* 4/1993, Dept. of Speech Communication and Music Acoustics, KTH, pp 97-140, 1993.