CREATION OF UNSEEN TRIPHONES FROM DIPHONES AND MONOPHONES USING A SPEECH PRODUCTION APPROACH

Mats Blomberg and Kjell Elenius

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

ABSTRACT

With limited training data, infrequent triphone models for speech recognition will not be observed in sufficient number. In this report, a speech production approach is used to predict the characteristics of unseen triphones by concatenating diphones and/or monophones in the parametric representation of a formant speech synthesiser. The parameter trajectories are estimated by interpolation between the endpoints of the original units. The spectral states of the created triphone are generated by the speech synthesiser. Evaluation of the proposed technique has been performed using spectral error measurements and recognition candidate rescoring of N-best lists. In both cases, the created triphones are shown to perform better than the shorter units from which they were constructed.

1. INTRODUCTION

The triphone unit is the basic phone model in many current phonetic speech recognition systems. The reason for this is that triphones capture the coarticulation effect caused by the immediate preceding or following phonetic context. One drawback is that the triphone inventory is quite large. To include all triphones of a language in sufficient number of occurrences in a training corpus for speech recognition is not practically possible. The low-frequent triphones will occur in too small numbers or not at all.

Current training corpora for large vocabulary speaker independent speech recognition systems are very large in order to include as many triphones as possible in sufficient number. Typically, hundreds or thousands of speakers are used and the total duration of the recordings may amount to 100 hours or more.

An extra problem is that the corpora are often application specific. The recognition accuracy of a system trained on such a corpus may drop significantly if it is used in another application than the one for which the corpus was designed. It is therefore a risk that the reusability of collected training corpora will be limited and the development cost has to be shared among a few tasks with similarities in their application types. Low budget applications generally cannot afford to collect and organise the necessary speech data for training large vocabulary speaker independent recognisers.

Different techniques may be used to overcome the problem of unseen triphones. The training corpus can be designed to give higher priority to application independence and to phonetic balance. Another approach is to try to make better use of the existing training data. A common technique to account for limited training data in systems based on Hidden Markov modelling is deleted interpolation [1], which combines different context models of a phone in a probabilistically weighted fashion. Tying of acoustically similar models has also been used to handle the insufficient training data problem. Lee [2] clustered phones into generalised triphones using entropy measures. Decision-tree based algorithms have been used to cluster allophones [3, 4] or subphone states [5, 6]. With these techniques, data is shared between different models, thus reducing the required amount of training data. Decision trees can also be used to predict triphones that have not occurred at all during training.

The underlying hypothesis in this report is that it may be possible to predict unseen triphones by using relations between similar elements in a limited triphone library. We believe that these relations are easier to describe and understand using a speech production oriented representation than a spectral (or cepstral) one. Specifically, the transition region across a phoneme boundary can be predicted, to a first approximation, using linear interpolation in the formant domain. The ideas described in this report explore some of the possibilities enabled by this representation.

2. TRIPHONE REPRESENTATION

In the following, context-dependent phones are denoted monophones (context-free phones), diphones or triphones if they have specified neighbouring phones at none, one or two sides, respectively. Diphones are left- or right-dependent with respect to which side of the phone that is connected to a specified phone. A diphone pair is defined as one left- and one right-dependent diphone that have the same identity of the central phone.

In our work, the context-dependent phones are represented at both a production parameter and spectral level [7]. The production parameters are those of a cascade formant synthesiser. The parameter envelopes are approximated by piece-wise linear segments according to Figure 1. The formant representation is used for triphone training while matching is performed at the spectral level. The formant representation is suitable also for speaker adaptation and different types of transformation. Male-to-female phone library conversion has been performed with good results [8].

During training, formants are tracked using an analysis-by-synthesis technique. The trained line segments of each context-dependent phone are converted into a spectral state sequence by the formant synthesiser and a dynamic programming algorithm.

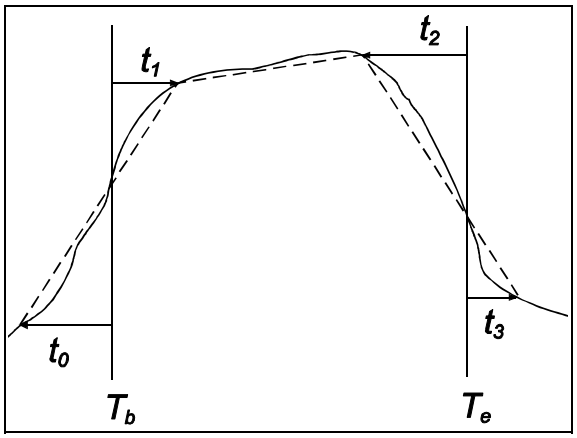

Figure 1: Definition of time positions of connections between successive line segments that approximate a parameter curve. T_b and T_e are phone boundaries. t_0 and t_1 are relative to T_b while t_2 and t_3 are relative to T_e .

3. TECHNIQUES FOR CREATION OF UNSEEN TRIPHONES

Different techniques for predicting unseen triphones using a speech production approach have been suggested in [9]. These can be categorised into concatenation techniques, composition from phonetic feature trigrams and transformation of similar, existing triphones. In this report, we have investigated phone/diphone concatenation combined with linear interpolation of parameter trajectories between line end points in adjacent units. The motivation for this technique is that interpolation in the formant domain supposedly approximates the spectral behaviour during transient intervals better than if it is performed at the spectral level. A first attempt with this technique is described in [7]. It is compared to a baseline technique, which uses a diphone pair, a diphone or a monophone if the requested triphone has too few occurrences in the training data. The spectral states of the 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 each diphone.

3.1. Formant Concatenation of Diphones and Monophones

Two phone units can be concatenated at the formant level and later be converted to spectral states. If both units are diphones, i.e. a diphone pair, the values of the first two edges of the piece-wise linear trajectory of each parameter are taken from the first diphone and the values of the last two edges are picked from the second one. See Figure 1.

One limitation in this type of concatenation is that the two halfs of the new unit are independent and do not contain coarticulation effects from the opposite phonetic context. In order to account for coarticulation, we have tested adjusting the values at the line break points. For each control parameter p and time position t_j, a new value $V'(p,t_j)$ is given by

$$V'(p,t_j) = \sum_{i=0}^{3} w_{ij} * V(p,t_i) \ , \quad \sum_{i=0}^{3} w_{ij} = 1 \ .$$

Each w_{ij} is currently the same for all parameters. It is likely, though, that some parameters are more prone to coarticulation than others and, accordingly, the coefficients should be parameter specific.

If no diphone pair exists for a requested triphone, the triphone can be created from a diphone and a monophone. In this case, three points are taken from the diphone and the remaining point is taken from the monophone. A problem is that one of the diphone values and the monophone value are context-independent. The accuracy of the parameter envelopes in these positions will, therefore, be reduced and their variability will be increased, resulting in lower phonetic discriminative ability.

4. RECOGNISER FRAMEWORK

The recogniser used in the experiments is described in [10]. In the experiments described in this report, it is used for rescoring an N-best list of sentence candidates produced by an A* stack search algorithm [11]. For this purpose, the candidates are merged into a lexical net. Phones with multiple and non-identical left or right context due to branching are duplicated in order to allow triphone modelling of every phone state, including word boundaries. The recognition algorithm performs a Viterbi search in the Bark spectral domain to find the best path through the net. Dynamic source adaptation is performed during search in order to compensate for deviant voices. Duration is alternatively modelled by a logarithmic Gaussian distribution or an exponential function as in conventional HMM systems.

The objective in this report is not to optimise the total system performance but to compare different techniques for expanding a triphone library. We have therefore removed one feature that would inhibit a correct comparison between the techniques. The residual spectral error in the analysis-by-synthesis formant tracking algorithm is normally compensated for in the spectral estimates of the trained triphones. There is currently no estimation of residual spectral error in the created triphones so we have disabled this compensation in all experiments. This feature has been shown to increase system performance and the expected accuracy will therefore be somewhat lower than optimal.

5. EXPERIMENTS

The WAXHOLM speech data base [12] is used for training the triphone library and for the tests. Currently, spoken dialogues from 65 speakers, 48 male and 17 female, have been collected and used in the experiments. Of these, 56 subjects were selected for training. The test corpus consists of 327 sentences (1672 words) spoken by 5 male and 4 female speakers, not in the training group. 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%. Higher

accuracy N-best lists are continually produced in the ongoing development work.

Cross validation experiments were performed by splitting the total training corpus into two equal parts. The duration-normalised average squared spectral error was computed between created triphones from units in the remaining training part against original triphones in the cross validation part. The individual errors were observation frequency weighted. The cross validation triphones were trained using the technique described above. This error measurement was also performed between the test corpus and the full training data. There were 1330 and 1057 different triphones in the cross validation and in the test data, respectively, that had sufficiently high observation frequency in both parts to be used for these measurements. The values of the coarticulation weight coefficients w_{ij} were empirically chosen to lower the cross validation errors of triphones created from diphone pairs.

In the N-best rescoring experiments, triphones created from diphone pairs were tested. Two different conditions for creating a triphone were compared. In the first case, it is created when the requested triphone does not exist but the diphone pair does. In the second case, also existing triphones are replaced by triphones created from diphone pairs. The two strategies were compared to the baseline strategy which uses the unit with longest context dependence range with sufficient number of observations in the training data of triphones, diphone pairs, diphones and monophones. The lower limit of the number of observations was heuristically set to 5. With this setting, the average proportions of unit types in the lexical net were as shown in Table 1.

Triphones	**Diphone pairs**	**Diphones**	**Monophones**
47%	28%	21%	4%

Table 1: Relative use of different unit types in the baseline system.

Although both the spectral average and variances of the phone substates are derived from the formant representation, only the average is used in this report. The variance is set to a constant value.

Reordering of N-best candidate lists may not be the optimal test environment for comparing speech recognition techniques due to the limited number of alternatives and the possible lack of the correct sentence identity among the candidates. On the other hand, it enables testing of computationally expensive techniques. Reordering is an important component of a complete system and it is desirable to find techniques that improve its performance.

6. RESULTS

The results of the spectral error measurement in the cross validation and the test data are shown in Figure 2. For all unit types, using the proposed technique to concatenate these units into triphones reduces the average spectral error. In other words, it is better to concatenate three monophones into the requested triphone than to use the appropriate monophone. The same applies to diphones and diphone pairs.

Incorporating coarticulation weights w_{ij} reduces the error for all unit types in the test data as well as in the cross validation data, from which they were estimated. Evidently, the positive contribution of the adjustment for coarticulation is kept in the independent material. There is even an indication in both sets that the average error of the concatenated diphone pairs is lower than that of the triphones. The explanation to this may be that the triphones have been observed in lower numbers and are not as well trained as the diphones. The interpolation and coarticulation adjustment techniques evidently function well in predicting the parameter trajectories.

Inspection of individual errors of the created triphones shows that the most difficult phoneme categories are inter-word silent intervals and unvoiced plosives.

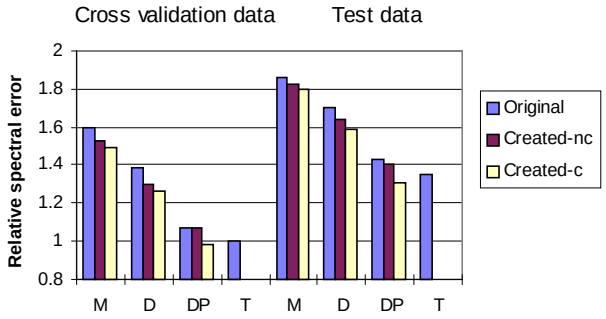

Figure 2: Spectral errors for different unit types in the cross validation and in the test data. The values are normalised to the average error between identical original triphones in the training and in the cross validation set. The "Created" columns in each group represent triphones created from this unit type. Labels: M - Monophone; D - diphone; DP - diphone pair; T - triphone. The suffixes -c and -nc indicate that the coarticulation transformation has (has not) been applied.

We also measured the errors when creating triphones that exist in the cross validation data but not in the training data. This condition corresponds more directly to a recognition situation when the requested triphone does not exist. The frequency-unweighted average errors for triphones created from diphone pairs and for the original diphone pairs were 242 and 255 $(dB)^2$ respectively. This indicates the relative advantage of using the former units in a practical recognition task.

The results of the N-best rescoring tests in Table 2 are consistent with the spectral errors for the different units. Creating unseen triphones from seen diphone pairs exhibit a small improvement compared to using the diphone pairs themselves. The improvement from also replacing the original triphones is positive, but too small to be statistically significant.

	Baseline system	Created triphones replace diphone pairs	Created triphones replace diphone pairs and triphones
Full training	72.8	73.0	72.8
1/4 training	71.8	72.3	73.1
Full training, log Gauss dur	76.9	77.3	78.1

Table 2: Word accuracy when training on all or 1/4 of the training data. The left column shows the result of the baseline system. In the middle column, triphones are created from diphone pairs by formant interpolation if the requested triphone does not exist. In the right column also existing triphones are replaced by created triphones. The bottom row includes logarithmic Gaussian duration distribution.

7. DISCUSSION

The results indicate the potential power of a production-oriented description of speech. The required size of a training corpus for building a triphone library can be reduced essentially if it can be constructed from shorter, more frequent units, *e. g*., diphone pairs.

Within each unit size, the proposed technique reduces the spectral error compared to the original unit. However, the error still exceeds that of the next unit size. The possible exception is for triphones created from diphone pairs. Explanations to this might be that the higher observation frequency for diphones than for triphones make them better estimated and that the coarticulation adjustment is particularly accurate when the phonetic context is known at both sides. The larger average error of original diphone pairs compared to that of original triphones, indicates that the formant interpolation and the coarticulation transformation techniques can be given the main credit for this result. Further improvement in recognition accuracy is expected by incorporating created triphones also in those cases where original diphones or monophones are used.

The manual setting of weight coefficients in the coarticulation algorithm could be improved by automatically optimised weight coefficients, specific for each parameter.

It is likely that more elaborate, context-sensitive methods for parameter trajectory modification will further improve the quite simple concatenation and interpolation techniques used in this report. Future work includes comparison with other techniques for triphone prediction and prediction of the statistical distribution of the created models.

8. REFERENCES

1. Bahl L. R., Jelinek F. and Mercer R. L. “A maximum likelihood approach to continuous speech recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5: 179-190, 1983.
2. Lee K. F. *Automatic Speech Recognition: The Development of the SPHINX system*, Kluwer Academic Publishers, Boston, 1989.
3. Bahl L. R., Bakis R., Bellegarda J., Brown P. F., Burshtein D., Das S. K., de Souza P. V., Gopalakrishnan P. S., Jelinek, F., Kanevsky, D., Mercer, R. L., Nadas, A. J., Nahamoo, D. and Picheny, M. A.. “Large Vocabulary Natural Language Continuous Speech Recognition,” *Proceedings of ICASSP 89*: 465-467, Glasgow, 1989.
4. Högberg J. “A Phonetic Investigation Using binary Decision Trees,” *Papers From the Eighth Swedish Phonetics Conference*, Lund, 76-79, 1994.
5. Hwang M.-Y., Huang X. and Alleva F. “Predicting unseen triphones with senones,” *Proceedings of ICASSP 93*: 311-314, Minneapolis, 1993.
6. Young S. J., Odell J. J. and Woodland P. C. “Tree-Based State Tying for High Accuracy Acoustic Modelling,” *ARPA Workshop on Human Language Technology*: 286-291, Plainsboro, New Jersey, 1994.
7. Blomberg M. “Training production parameters of context-dependent phones for speech recognition,” *STL-QPSR* 1/1994, Dept. of Speech Communication and Music Acoustics, KTH, 59-89, 1994.
8. Blomberg M. “A common phone model representation for speech recognition and synthesis,” *Proceedings of ICSLP 94*: 1875-1878, Yokohama, 1994.
9. Blomberg, M. “Creation of unseen triphones from seen triphones, diphones and phones,” *TMH-QPSR* 2/1996, Dept. of Speech, Music and Hearing, KTH, 113-116, 1996.
10. 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, 97-140, 1993.
11. Blomberg M., Elenius K., Ström N. “Speech recognition in the Waxholm Dialogue system,” *Papers from the Eighth Swedish Phonetics Conference*: 22-23, Lund, 1994.
12. Bertenstam, J., Blomberg M., Carlson R., Elenius K., Granström B., Gustafson J., Hunnicutt S., Högberg J., Lindell R., Neovius L., de Serpa Leitao A., Ström N. “Spoken dialogue data collected in the WAXHOLM project,” *STL-QPSR* 1/1995, Dept. of Speech Communication and Music Acoustics, KTH, 49-74, 1995.