Proceedings of the Institute of Acoustics

UNIT INVENTORY AND MODEL STRUCTURE FOR VOCABULARY-INDEPENDENT HMM RECOGNITION

Wendy J. Holmes, Lynn C. Wood and David J. B. Pearce

**GEC-Marconi Limited, Hirst Research Centre,
East Lane, Wembley, Middlesex, HA9 7PP, U.K.**

ABSTRACT

This paper describes the use of allophonic sub-word units with an allophone-dependent model structure, to improve the performance of sub-word HMM recognition using vocabulary-independent training. The new system is an extension of an approach based on sub-triphone units called *phonicles*. The original system [1, 2] modelled major phonetic context effects by splitting each phone into a left-context dependent phonicle followed by a right-context dependent phonicle. It did not however take account of context effects wider than one immediately adjacent phone or the differences in duration and spectral complexity which exist between different types of phoneme. The recognition system has therefore been extended so that phoneme transcriptions are first converted to allophone transcriptions. Each allophone is then transformed to a sequence of one or more *allophonicles*, where different allophonicles can have different numbers of states and one allophonicle can be shared across allophones. Using a Mel Cepstrum front end, isolated-word speaker-dependent recognition experiments on six example application vocabularies have shown a reduction in the average error rate from 4.9% to 0.3% by using allophonicle models. The paper discusses the results of this and other experiments in more detail.

1. INTRODUCTION

One of the main reasons for using sub-word units in speech recognition of medium-sized vocabularies is the ability to easily add new words to an existing vocabulary or to generate a new application vocabulary, without collecting more training data. Ideally, a general database is used to train, once and for all, a set of *vocabulary-independent* sub-word models, which can then be used for recognition of any application vocabulary. This approach eliminates the time-consuming and expensive process of repeated task-specific training, and allows rapid configuration of speech recognisers for new vocabularies. In order to obtain good performance from sub-word systems, they must adequately represent all the important phonetic context effects which are inherently incorporated in whole-word models. A variety of different context-specific sub-word units have been proposed, with the most popular being some form of *triphone*, which models every phone in the context of its left and right neighbours. There are many possible triphones and so, in vocabulary-independent systems, many of the triphones in the test vocabulary will not have occurred in the training vocabulary. With a conventional triphone system there is however no obvious method for generalising to contexts not observed in the training data. One solution to this problem is simply to collect more training data until good coverage of the possible context effects is achieved [3]. In many cases, however, it may not be practical to collect such large quantities of data. In addition, with an appropriate choice of context-dependent units, such a large data collection task should not be necessary. By exploiting the fact that some groups of phones have similar effects on neighbouring phones, it should be possible to achieve similar levels of recognition performance by making the most effective use of a smaller amount of training data and a more manageable number of models. One approach which has been applied within the triphone framework is top-down decision-tree clustering [4]. For contexts in the test data which were unseen in the training data, the tree can be traversed to find the most detailed model

available. An alternative to the basic triphone approach has been taken here, in which triphone models are built from an appropriate sequence of smaller units. The sub-triphone unit, developed by Wood et al. [1, 2], is called a *phonicle* (*phon*etic par*ticle*) and is similar to the *semiphone* suggested by Paul [5]. By sharing phonicle models across different triphones, similarities between parts of different models are explicitly incorporated and the number of possible context-dependent units is much smaller than the number of possible triphones. Therefore the memory requirements are lower and undertraining is less of a problem, as most contexts will occur at least a few times in a database of manageable size. It is also possible to create new triphone models as long as suitable sub-triphone units are available.

The original phonicle system [1, 2] aimed to model the major effects of phonetic context, by simply splitting each phone into two two-state phonicles, where the first was left context-dependent and the second right context-dependent. Triphone models were built from appropriate pairs of phonicles. This system enabled major context effects to be modelled, but it seemed likely that performance could be improved by making it more flexible, to incorporate the following:

1. **Wider context effects:** There are some cases where a phonicle will be significantly affected by both left and right context. An example is distinguishing between aspirated and unaspirated voiceless stops, where the presence or absence of aspiration is influenced by the preceding sound. There may also be cases where it is important to include effects from more than immediate context. These effects are possible for certain consonant clusters, such as the /r/ in /str/ which tends to be different from the /r/ in /tr/. In addition, the realisations of a phoneme can depend on factors other than phonetic context, such as utterance position and stress.
2. **Allophone-dependent model structure:** Better modelling of utterances could be achieved by allocating the number of phonicles and the number of states per phonicle according to the duration and spectral complexity of each phone. For example, it is reasonable to assume that a clear /l/ would need less states than a stressed diphthong, which is much longer in duration. Similarly, a voiced stop contains only two spectrally distinct regions, for the closure and the burst, and so should require less states than an aspirated voiceless stop, which also has an aspiration portion.

The aim of the work described in this paper was to extend the phonicle system to be allophone-based, with the most appropriate model structure assigned to each unit. The approach has been to model the additional context effects included in triphone-type systems when they are important, but at the same time retain the advantages of the phonicle approach in providing a manageable number of trainable models and enabling the building of appropriate models for unseen contexts.

2. THE ALLOPHONICLE SYSTEM

In the new system, a two-stage process is used to obtain an utterance transcription in terms of sub-triphone units:

1. The phoneme transcription is converted to an allophone sequence, by applying context-sensitive **allophone rules.**
2. Each allophone in the transcription is transformed to be a sequence of one or more *allophonicles*, using a set of **phonicle rules.** These are simple expansion rules, specifying a phonicle sequence with which to replace each allophone symbol.

The new allophonicle system is very flexible:
- Different allophones can be represented by different numbers of phonicles.
- Different phonicles can have different numbers of states.
- Each phonicle can be left context-dependent, right context-dependent or context-independent.
- Any one phonicle can be shared between allophones where appropriate.

For example, there are aspirated and unaspirated allophones of the voiceless stops. The unaspirated allophone is modelled by two phonicles (closure and burst), whereas there are three for the aspirated version (closure, burst and aspiration). The closure phonicle model is the same for both allophones.

The allophone and phonicle rules and the phonicle characteristics are all specified in text files, so it is very simple to experiment with different inventories and model structures. The current inventory of allophonicles was developed using the following sources of information:

1. Initially, allophones were added to the unit inventory based on general "phonetic knowledge" about the types of context effect likely to be important for incorporation into models for recognition.
2. Useful information about consonant cluster modelling in particular was obtained by performing "confusable model tests" on groups of words chosen from the training database. These tests involved selecting a particular sequence of phonemes and identifying all other likely sequences with which it might be confused. Then, for each word containing the phoneme sequence of interest, models were constructed for the set of confusable "words" and a recognition test performed for the single database word compared against all these models. This approach enabled investigations to be performed into the modelling of fine phonetic distinctions, based only on the training data but without requiring highly confusable words to be in the database.
3. The development of the allophone inventory was helped by identifying and analysing the causes of recognition errors, which involved studying segmentations, alignments and distance scores.

3. ALLOPHONE TRAINING PROCEDURE

The recognition system [1, 2] is based on hidden Markov modelling of sub-word units using a single diagonal covariance multivariate Gaussian probability density function, pooled over all states and all models. The models use a simple left to right topology with self transitions and no skip transitions. A special single-state silence model is used. The sub-word units are trained by performing several iterations of an "embedded" Baum-Welch re-estimation procedure, using only the orthographic transcriptions, a pronunciation dictionary, and unmarked training data. No hand-labelling of the data is performed at either the word or sub-word level.

The basic principle used to train a set of context-dependent allophonicle models is to apply a number of stages, with each stage being progressively more context-specific. The process begins by initialising a set of context-independent models to identical values computed from the centroid and variance of the entire training data. Five iterations of the Baum-Welch re-estimation procedure are then applied to train these models. As long as there are enough examples of each context-independent unit in a range of contexts, well-trained models should be obtained. The following stages then gradually become more context-specific, at each stage initialising the models from those trained to the previous level and performing two training iterations. This approach should provide good initial estimates for the models to be trained at these later stages, for which the number of examples will often be quite small. When adding new allophones to the inventory, there is a trade-off between improvements in the phonetic modelling and potential problems from under-training at both the context-independent and context-dependent levels, thus:

- For some allophones there may be only a small number of training examples and so even the context-independent models will not be well-trained. This is particularly problematic as the initial estimates of the context-independent models are identical for all phonicles and therefore do not distinguish any of the characteristics of the different sounds.
- The use of context-dependent phonicles based on allophones means that, the more allophones there are, the greater the number of possible context-dependent phonicles. Thus there is a greater danger of undertraining the context-dependent units than with the original phonicle-based system.

To minimise these potential problems, the following five-stage training procedure is used:

1. **Reduced allophonicles:** The first stage is to apply five iterations of the re-estimation procedure to train a set of context-independent models for a reduced allophone inventory. At this stage, the inventory is chosen so that there are many examples of each unit (generally at least 30). The idea is to include some of the major allophonic distinctions, but only those for which there is an adequate amount of training data. For example, different models for aspirated and unaspirated voiceless stops are included at this stage. However, the first stage does not include the more specific models, such as a special allophone of /t/ for the /str/ context, for which there are only a few examples.
2. **Context-independent allophonicles:** The set of well-trained models from the first stage is then used to initialise the models for the full allophone inventory. This provides quite good initial estimates for the allophonicles, so that two further iterations of context-independent training are sufficient.
3. **Reduced biphonicles:** The first level of context-dependent models groups allophone contexts together into very general categories, which group together those sounds having a similar effect on an adjacent sound. These models depend on the 'broad class' of the adjacent allophonicle (either the left or right according to the context-dependency of the model). Ten broad classes are used, defined by place of articulation.
4. **Phoneme-dependent biphonicles:** The next stage is to train biphonicle models dependent on the identity of the adjacent phoneme within each broad class. This stage forms an intermediate step before the final allophone-dependent stage, and can be very useful when generating models for recognition of allophone contexts unseen in the training data, as explained further in Section 4.
5. **Allophone-dependent biphonicles:** This is the final, most context-specific, stage.

At each context-dependent stage, models are only trained for a given context if there are more than three examples for that context. The requirement for three examples was chosen as this should be enough to produce considerably more representative models than is possible with only one example, while not being such a high threshold that very few allophone-dependent models would be trained.

4. ALLOPHONE RECOGNITION PROCEDURE

To perform recognition for a specified test vocabulary, word models are formed by concatenating the appropriate sequence of sub-word models, using a pronunciation dictionary and statistics of occurrence of the context-dependent units in the training data. Connected recognition is then performed using a one-pass dynamic programming algorithm with a beam search to determine the sequence of word models. When recognising isolated words, the connected algorithm is used with a syntax specifying that the utterance must be a single word, preceded and followed by silence.

In addition to its advantages for producing robust, well-trained models, the hierarchical training procedure described in Section 3 has useful consequences for the recognition of allophones in contexts which were unseen in the training data. By retaining the allophonicle models for all levels of training, it is possible to build models for unseen triphones from the most context-specific models which have been trained. When forming word models for use in recognition, the allophone-dependent biphonicles are used where available. If not, either the relevant phoneme-dependent biphonicle or the reduced biphonicle can generally be used, and it will only rarely be necessary to resort to the context-independent allophonicle model. It is for this use in recognition that the phoneme-dependent stage was found to be particularly useful, as there were quite often cases where the required allophone context had not been trained but the phoneme context was available. Thus the existence of both levels helps to avoid a potential disadvantage of adding allophones: with a larger number of allophones, the coverage of possible allophone-dependent contexts will be worse and hence it will more often be necessary to use more general models. By having a phoneme-dependent stage, the models used will never be more general than they were with the original phoneme-based system.

5. EXPERIMENTS

5.1. Databases

Experiments have so far been performed on speaker-dependent isolated-word recognition for a single speaker, using the following databases:

Training: A 1242-word training vocabulary was used, where the words were not linked to a specific application, but were chosen to provide examples of every phoneme in a range of phonetic contexts. This vocabulary was available as a set of isolated words with one occurrence of each word, and also in the form of 200 sentences. These sentences were based on the same 1242-word vocabulary as the isolated-word database. Several of the words did however occur more than once, and the function words occurred many times. This database was used for comparisons between isolated-word and connected-word training for isolated-word recognition. A further set of 287 isolated words was also used; these were chosen to provide at least three examples of every valid phoneme in both utterance-initial and utterance-final position. These words, which will be referred to as IFC words, were a supplement to the connected training data, which is inherently limited in its coverage of Initial and Final Contexts.

Test: Six different example application vocabularies were chosen, and two occurrences of each word were recorded. These vocabularies provided a wide variety of words, using only a small overlap with the training vocabulary. They provide a good basis for testing the modelling of unseen triphones, as 36% of the triphones occurring in the test databases do not occur in the training data. In addition to the application vocabularies, five confusable vocabularies were also recorded. These were sets of minimally distinct words, chosen to test the ability of the sub-word models to make fine phonetic distinctions, such as discriminating between different stop/semivowel clusters. Vocabularies were chosen to contain groups of valid words that differed in only one consonant or consonant cluster.

5.2. Speech Analysis

The utterances in the databases were recorded with the speech digitised at a sampling rate of 20 kHz, and the data was digitally down-sampled to 8 kHz. Experiments have been carried out using an FFT-based Mel cepstrum at a frame rate of 62.5 frames/second. The front-end parameters were augmented with their time derivatives, which were computed for any one frame as the difference between the parameters for the following and preceding frames.

5.3. Results using isolated-word training

Example application vocabularies: Using the training database of 1242 isolated words, a set of baseline phonicle models was trained using the original system, with each phoneme modelled by two two-state phonicles. The recognition results on the six example application vocabularies are tabulated in Figure 1. The average error rate was 4.9%, although there was a considerable degree of variation between the different vocabularies. A study of the types of error revealed that they occurred mainly for long polysyllabic words with a number of unstressed syllables and words involving sequences of consonants. By studying the alignment of the models with the utterances for words of this type, it became obvious that in some cases the minimum duration imposed by the model was much longer that the duration of the actual utterance, so a bad match for some frames was inevitable. Therefore, a second set of phoneme-based models was trained in exactly the same way as before, but modelling each phoneme with a one-state left-context-dependent phonicle followed by a two-state right-context-dependent phonicle. More states were assigned to the second phonicle to correspond with the idea that English is mainly an anticipatory co-articulation language. The results with the three-state phonicle-based models are also included in Figure 1, and it can be seen that the error rate has reduced considerably to an average of only 1.2%. Most of the problems with misrecognising very long words have disappeared.

A set of allophone models was trained using the chosen allophonicle inventory, and an extremely low average error rate of 0.3% was achieved. In fact the only remaining errors are the misrecognition of "oh" as "no" in the bank set, and two non-semantic confusions in the ATC set (recognising "eh" as "eight" and "leff" as "left"). Overall, the results show that recognition performance is improved by using the allophone models, and it has now reached the best that can be expected for these vocabularies.

Confusable test vocabularies: Recognition experiments with the same three sets of models were also carried out on the five confusable test sets. The aim was to evaluate performance on much more difficult test vocabularies, involving fine phonetic distinctions but also including unseen contexts. Obviously performance levels are much lower for these vocabularies, and even a human listener would be very unlikely to achieve complete accuracy. The average error rate was 14.8% for the four-state phoneme models, 17.5% for the three-state phoneme models and 10.8% for the allophone models. In these tests, the four-state phoneme models have performed surprisingly well, being generally better than the three-state phoneme models. This difference is probably mainly due to the fact that many of the training words and most of the words in the confusable test sets are stressed monosyllables. For stressed monosyllabic words, four states per phoneme is not excessive and allows for more detailed modelling of the spectral changes than is possible with the three-state models.

5.4. Results using connected-word training

For isolated-word recognition, it is obviously most appropriate to also train using isolated words. However, by using connected sentences, it is much easier to obtain a large quantity of data and therefore good coverage of phonetic contexts. The purpose of the experiments described in this section was to compare isolated-word recognition performance for sentence training with that for isolated-word training, based on the same training vocabulary. The comparisons were performed for the three-state phoneme models and the allophone models. The minimum duration imposed by the four-state phoneme models was too long for them to be appropriate when training on connected data.

It was to be expected that models trained on connected data would not perform so well for isolated-word recognition as those trained on isolated words, due to a number of factors:

- Effects of inter-word co-articulation on the realisation of phonemes, and in some cases on the identity of the phonemes.
- Less careful articulation for some words in sentences, and generally different speaking style for connected utterances.
- Greatly reduced coverage of phonemes in utterance-initial and utterance-final position.
- Pauses, breaths etc. may occur in sentences, but cannot be predicted reliably.
- In the case of allophone models, the distinction between stressed and unstressed allophones will be more difficult to make clearly due to effects of sentential stress.

To investigate the effect of these factors on isolated-word recognition performance, the databases described in Section 5.1 were used to make comparisons between models trained on the following:

- The 1242 isolated words used in the experiments described in Section 5.3 (*iso*).
- The 200 connected sentences taken from the same 1242-word vocabulary (*con*).
- The same 200 sentences, supplemented by the 287 IFC words (*con_ifc*).
- The 1242 isolated-word set, also supplemented by the 287 IFC words, to correspond with the previous condition in terms of the within-word contexts which were included (*iso_ifc*).
- The combined group of the two isolated-word sets and the 200 sentences, which offered more training examples than any of the other conditions (*comb*).

It was expected that using the connected-word training data would tend to impair isolated-word recognition performance, but that to some extent the disadvantages could be offset by the fact that the

sentences included more than one example of some of the training words. The main findings are described below for the two vocabulary types in turn.

Application vocabularies: From Figure 2, it can be seen that the average error-rates were generally low for all training conditions, and were consistently lower for the allophone models than for the three-state phoneme models. Training on connected sentences only (*con*) has resulted in worse recognition performance than training on the isolated word set. However, the difference is not very great for these vocabularies where the words are quite easily distinguishable from each other. The addition of isolated words (*con_ifc* and *comb*) improved recognition performance for the connected training condition, which is presumably mainly due to the fact that the isolated words will have provided more examples of the context-dependent phonicles, particularly for utterance-initial and utterance-final positions. The presence of isolated words may also have improved the segmentation of the sentences, particularly in the regions of word boundaries.

Confusable vocabularies: The training conditions had a greater effect on performance for these vocabularies, as can be seen from Figure 2. For both the phoneme and allophone models, overall performance is much worse for the connected-training condition than for isolated-word training and adding the IFC words to the connected training data (*con_ifc*) substantially improved recognition performance. These findings imply that using connected-sentence training data does degrade the ability to make certain fine phonetic distinctions. The *comb* condition using all available training data gave rise to slightly better average performance than just using all the isolated words, although relative performance was different for the different vocabularies. The pattern of errors suggested that, for phonetic contexts that were fairly well-represented in the isolated words, adding sentences tended to make performance worse. However, for contexts which were poorly covered in the isolated-word data, the advantage of additional training material outweighed the disadvantages of the connected nature of the utterances.

It is interesting to note that the advantage of using the allophone inventory rather than the standard phoneme set is greater when training only on isolated words than when connected sentences are included. This finding implies that, for an isolated-word recogniser, the most effective use is made of the allophone inventory when using isolated words for training. This is probably at least partly because any errors in the phonemic transcription of a sentence tend to have a more detrimental effect on the allophonicle models than the phonicle models, as there are more allophones than there are phonemes and therefore less examples to train many of the models.

6. CONCLUSIONS

Improvements to vocabulary-independent isolated-word recognition for a single speaker have been demonstrated, by using a set of allophonicle HMMs in which different model structures are chosen to suit different classes of allophonicle. It has been found that care must be taken to avoid assigning more states to any allophone than is appropriate for its typical duration, specified as a number of analysis frames. When developing an allophone inventory, caution is also necessary before adding theoretically desirable allophones if there are not many examples of those allophones in the available training data. When training allophonicle models at the context-dependent levels, more robust models were obtained by specifying a minimum of three examples rather than only one. This procedure reduces the danger of training models only on examples which are not typical of the relevant context.

The allophonicle approach is now being extended to speaker-independent and to connected-word recognition. Some additions to the allophone inventory will probably be needed when developing a system suitable for a range of speakers, although the general allophone model structure and dependence of allophones on context should apply to all speakers.

7. REFERENCES

[1] Wood L C and Pearce D J B "Sub-word HMM Recognition: An Investigation of Phone Context Modelling and Improved Discrimination", Proc. IOA Conf. on Speech and Hearing, Windermere, pp. 181-188, 1990.

[2] Wood L C, Pearce D J B and Novello F "Improved Vocabulary-Independent Sub-Word HMM Modelling", Proc. IEEE ICASSP, Toronto, pp. 181-184, 1991.

[3] Hon H W and Lee K F "On Vocabulary-Independent Speech Modeling", Proc. IEEE ICASSP, Albuquerque, pp.725-728, 1990.

[4] Lee K F, Hayamizu S, Hon H W, Huang C, Swartz J and Weide R "Allophone Clustering for Continuous Speech Recognition", Proc. IEEE ICASSP, Albuquerque, pp. 749-752, 1990.

[5] Paul D B "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer", Proc. IEEE ICASSP, Toronto, pp. 329-332, 1991.

database	vocabulary size	error rate (%) 4-state phonemes	3-state phonemes	allophones
Applications:				
ATC	98	2.6	2.0	1.0
bank	71	7.8	1.4	0.7
objects	64	1.6	1.6	0
cities	65	3.9	0.8	0
shares	68	8.8	0	0
nations	140	4.6	1.1	0
Average	84	4.9	1.2	0.3
Confusable words:				
aist	64	7.3	12.5	3.6
aits	68	12.7	14.2	6.9
ain	84	14.3	12.7	6.7
ue	41	32.5	35.8	28.5
misc	59	7.3	12.4	8.5
Average	63	14.8	17.5	10.8

Examples of words in confusable vocabularies:

aist: traced, taste, Tate, trade
aits: traits, trades, trays, strays
ain: strain, stain, strains, stained
ue: strew, stew, true, too
misc: ran, rank, rips, ripped

Figure 1: Speaker-dependent isolated-word recognition results for each test vocabulary (isolated-word training)

Training condition	error rate (%) application databases 3-state phonemes	application databases allophones	confusable databases 3-state phonemes	confusable databases allophones
iso	1.2	0.3	17.5	10.8
iso_ifc	1.1	0.3	15.7	9.9
con	1.6	1.2	27.7	26.6
con_ifc	1.0	0.8	20.8	19.0
comb	0.8	0.7	14.9	9.2

Figure 2: Summary of speaker-dependent isolated-word recognition results for each training condition