SUB-WORD HMM RECOGNITION: AN INVESTIGATION OF PHONE CONTEXT MODELLING AND IMPROVED DISCRIMINATION.

Lynn C Wood, David J B Pearce

GEC-Marconi Limited, Hirst Research Centre, East lane, Wembley,
Middlesex, HA9 7PP

ABSTRACT

In this paper two techniques are presented for improving the performance of sub-word recognition with open vocabularies. The first technique uses a new sub-triphone unit, called a phonicle, to allow triphone models which have not been encountered in the training data to be built from contexts which have been sufficiently trained. The second uses linear discriminant analysis (LDA) to improve the discrimination between the sound classes. The two techniques have been evaluated for speaker dependent training on an open vocabulary task. The recogniser is based on hidden Markov modelling (HMM) techniques with continuous distributions.

1. INTRODUCTION

One of the main motivations for using sub-word units rather than word size units is that the recogniser may be used with new application vocabularies without the need to record a further application specific database. In order to obtain a performance that is comparable to whole-word modelling, sub-word units must adequately model the phonetic contexts that occur in the new vocabularies. Most attempts to realise this objective have shown significant falls in performance when moving from the closed vocabulary (application specific) case to open vocabularies (new application vocabulary) [1]. One of the reasons for this difference in performance is the mismatch in the contexts seen in the training set and those required in the new vocabulary.

This paper describes two techniques used to improve the performance of sub-word recognition with open vocabularies. The methods also improve the performance in the closed vocabulary case. The recogniser is based on hidden Markov modelling (HMM) techniques with continuous distributions.

The first technique is to use a unit smaller than the phone to allow triphones models, to be built from those sub-triphone particles which have been sufficiently trained. We call these new units *phonicles* (PHONetic partICLE). A top-down supervised clustering approach has been employed to obtain an inventory of context-dependent phonicles and a comparison between the phonicle and triphone approaches is presented for an open vocabulary task.

The second technique uses linear discriminant analysis (LDA) applied to speech [2] to improve the robustness of the models to both the intra-speaker and inter-speaker variability that occurs in the speech data. The IMELDA representation developed by Hunt [2] combines LDA with a mel-scale representation and has been found to significantly improve the robustness of word based recognition systems [2][3].

In this paper, we first describe the training and testing databases for use in the sub-word evaluations. In section 3, a description of the sub-word training and recognition procedures is presented. Section 4 details the training and evaluation of the context dependent phonicle inventory for the open vocabulary tasks. In section 5, the implementation of the linear discriminant analysis in the sub-word domain is described and its recognition performance assessed. Finally some conclusions are given in section 6.

2. DATABASES

The database used for sub-word training was a set of 200 "phonetically rich" sentences spoken by four British English male speakers (the sentences also form part of the acoustic-phonetic corpus in SCRIBE [10]). At present only one speaker (gsw) has been used for training. The sentences were designed to give adequate coverage of the biphone contexts in English. However coverage of triphones is poor (of the 3011 triphones that actually exist in these sentences, 1784 occur only once). The vocabulary of 1242 words is not based on any specific application domain. Two databases were used for recognition evaluation, both based on a 98 word vocabulary for an air traffic control application. The vocabulary included the international communications alphabet, digits and function words (e.g "to", "at" etc.). The two evaluation databases consisted of (i) two repetitions of the 98 words spoken in isolation and (ii) 100 connected sentences. The overlap between the training and test sets was just 13 words and 33% of the triphones occurring in the test data were unseen in the training data.

The recordings of the databases were performed in a sound-proof booth using a Shure FM-10 microphone. The speech signal was digitised at a sampling rate of 20 kHz and the data was digitally down-sampled to 10 kHz. An FFT Mel-based cepstral analysis [11] was applied to obtain 12 cepstral coefficients plus an energy term spanning the range 0-5kHz at a rate of 100 frames/second. The front-end parameters were augmented with their time derivatives, computed as the difference between two frames spaced 40 ms apart [4]. For the isolated test data, word endpoints were obtained by employing an automatic energy thresholding technique. No hand labelling of the data at the word level or sub-word level was used in the training procedures.

A dictionary containing a single pronunciation of each word in the training and testing databases was used to obtain phonetic transcriptions of the data from the orthographic transcriptions, each word was separated by an optional silence. The size of the monophone inventory was 46.

The recognition performance on the connected task was obtained with no grammar, therefore the perplexity of the task was 98. A dynamic programming algorithm was used to align the actual word transcription of the test data with the words output from the recogniser to obtain the number of substitution N_s, insertion N_i and deletion N_d errors. Typically recognition results are presented in terms of %word correct (wc) and % word accuracy (wa) which are computed as follows [8]:

$$wc = (N - N_s - N_d) * 100 / N \qquad (1)$$

$$wa = (N - N_s - N_d - N_i) * 100 / N \qquad (2)$$

where N is number of words in the test data.

In [8], Hunt points out that the DP scoring introduced bias on the total error measure by over estimating substitution errors and underestimating insertion and deletions. A weighted word accuracy measure (wwa) was introduced in an attempt to reduced this bias. The weighted measure is computed as follows:

$$wwa = (N - N_s - 1/2N_d - 1/2N_i) * 100 / N \qquad (3)$$

The three measures for recognition performance will be used in this paper.

3. TRAINING AND RECOGNITION PROCEDURES

The recognition experiments on the open vocabulary tasks were performed using an in-house sub-word recognition simulation. Its development was strongly influenced by work on triphone modelling conducted under the DARPA programme [4][5]. In particular, the simulation encorporates similar features to those used in the MIT Lincoln Laboratory recogniser [5] and the SPHINX system at CMU [4]. These include the use of "grand" variance and embedded training of sub-word models.

The sub-word recogniser is based on hidden Markov modelling (HMM) of sub-word units using continuous probability density functions. Each sub-word unit was modelled using a 4 state HMM (multivariate Gaussian) employing a diagonal covariance matrix pooled over all states and all models; the silence model was a single state HMM. A simple left-right topology was used with self transitions and no skip transitions.

Training of the sub-word units was accomplished by performing several iterations of a "sentence level" Baum-Welch reestimation procedure. Sentence HMM models were obtained by concatenating the appropriate sub-word models using the word pronunciation dictionary, and the Baum-Welch algorithm was employed to provide a mapping of the training data to the sentence HMM's. On application of all the training data, an updated set of HMM parameters was obtained for each sub-word model. Initial estimates were set by assigning the mean and variance vectors at each state in each HMM with identical values [5] computed from the centroid and variance of the entire training data. Five iterations of the reestimation procedure were found to be sufficient to obtain good estimates of the context-independent models. Two further iterations were performed to obtain context-dependent models "bootstrapped" from the context-independent models.

During recognition, word HMM models were formed by concatenation of the appropriate sequence of sub-word context models using the pronunciation dictionary. A one-pass dynamic programming algorithm employing a beam search was used to perform connected recognition.

4. CONTEXT DEPENDENT PHONICLES FOR OPEN VOCABULARIES

The acoustic realizations of phonemes vary greatly depending on the phonetic context in which they occur. The most important sources of co-articulation can be modelled by a phone model which takes into account its left and right neighbouring phones i.e a triphone. Consequently triphone based recognition systems have demonstrated better performance than any other sub-word approach [4][5]. Most of the work on triphone modelling, however, has used training and evaluation databases with the same application domain where there is a large overlap of words occurring in both data sets. The good results achieved with closed vocabularies are not maintained in open vocabulary tests (i.e when the test vocabulary is not in the training set) [1]. One of the main reasons for this difference in performance is that the test data contains some contexts that do not occur in the training data; for these contexts it is necessary to use context independent models which will degrade performance. Current work on open vocabulary recognition has included (i) recording larger training databases to ensure improved triphone coverage [1] (in many cases, however, it is not practical to collect such large amounts of data). (ii) Obtaining more robust context models by clustering those contexts which are judged to be similar.[5][6] (iii) Use of non-application specific training vocabularies [7].

A new approach to the problem of building models for unseen context is to use a sub-triphone element called the phonicle. The basic premise of the approach is that there are sub-components of a triphone model that share the same type of context dependency which can be made use of in the training of triphones and in the construction of unseen triphones. For example, one might expect that the states at the left-hand end of a phone model would be more dependent on the left phone context than on the right phone context and that in the ideal case, the distributions would be independent of the right-hand context. Similarly the states at the right-hand end of a model will be less dependent on their left-hand phone contexts than on their right. In the ideal case it would be possible, for each phone to tie the distributions of the left-hand states of each triphone with the left context and similarly tie the right-hand states with the same right context together. Having done this and trained the models for the phonicles, it is then possible to construct an unseen triphone model from the component parts with appropriate context. The main problem is how best to partition the component parts with a HMM phone model into phonicle units. For the present the phonicle unit is assigned two consecutive states and the phone model is constrained to have an even number of states.

In general, it is expected that the context dependencies will deviate from the ideal suggested above. To accommodate this and to make the approach more general, the phonicle model is combined with a top-down supervised clustering algorithm in order to train phonicle models with more detailed context dependency. The objective is for each unseen triphone to use the most specific context-dependent factors which can be adequately trained. The method is best illustrated by way of an example. Consider the phone /I/ with left context /s/ and right context /ng/ (as in the word "sing"). It consists of three phonicle units I1, I2, I3 where I1 and I3 model the transitions between the vowel /I/ and the adjacent phones, and I2 models the steady state of the phone. To build the triphone s-I-ng we first try to concatenate the phonicle models with the same triphone context ie s-I1-ng, s-I2-ng, s-I3-ng. If either of these context units do not occur in the training data, more generalised contexts must be found which do occur and can be modelled accurately. In the current implementation, five levels of context generalization are considered. These are shown below for the phonicle unit I1.

L1 monophone context	XX-I1-XX	
L2 reduced biphone	ALVEOLAR-I1-XX	
L3 biphone context	s-I1-XX	(XX = don't care)
L4 reduced triphone	s-I1-VELAR	
L5 triphone context	s-I1-ng	

Level 5 is the full triphone context. For the reduced triphone context in level four (L4), the assumption is that the left most phonicle unit in a phone is more dependent on the left phone context than the right; therefore the right context has been reduced to a broad class grouping. Six groups are currently used : Silence, vowels and 4 consonant categories grouped by place of articulation [9]. In level 2 and 3, the phonicle is made independent of right context. Level 1 is the context-independent case which is used only if contexts in levels 2 to 5 do not occur.

When the triphone context occurs in the training data, phonicle models with triphone context are used to build the triphone model; this is equivalent to the conventional triphone modelling. When the triphone is unseen, it will be interpolated from the most specific context factors available rather than simply defaulting to the context-independent models. For this reason this procedure is expected to be more robust for the open vocabulary task. Improved optimization of the context models is also achieved at each level by using the parameters of the previous level as the initial estimates for the training instead of the context-independent models.

Table 1 summarises a set of speaker-dependent word recognition results (speaker gsw) obtained on the open vocabulary task for the isolated and connected databases for the phonicle context modelling described above. The results for the connected task were obtained without a grammar and therefore represent the lower bound of performance for this task. The results show the incremental improvement in recognition performance as the modelling of context is made more specific (tests 1 to 5). The best performance was obtained when all five levels of context were used (test 5) with 95.5% and 93.6% words correct for the isolated and connected tasks respectively. The main source of errors for the connected task was the insertion of the function words in the recogniser output. Analysis of the training sentences showed that a set of 42 function words (e.g "to", "at", "the" etc) accounted for about 40-45% of the training words when weighted by frequency. It is expected that explicit modelling of the function words [4] should improve the recognition performance.

The phonicle context modelling described above was compared experimentally with the conventional approach where the unseen contexts were assigned the parameters of the context-independent model. Recognition results are shown in Table 1 (tests 6,7,8). In test 6 the triphone contexts were trained only if the triphone occurred 3 or more times. The poor performance (85.5% and 80.2%) is due to the poor coverage (only 18% triphones occur 3 or more times in the training set and 66% triphones are unseen in the test set). When the rarely occurring contexts are also reestimated (test 7) the recognition performance is even worse (75.5% and 75.0%) Although the coverage is improved, the contexts which occur very rarely will be inadequately trained. In test 8, the triphone contexts were formed from the left only and right only context dependent phonicles. The significant improvement in performance (93.6% and 91.3%) reflects the good coverage of biphone contexts provided by the training data.

The open vocabulary performance was compared with a closed vocabulary experiment where word HMM models were obtained for each of the CAA test words using five repetitions of each word spoken in isolation. The performance was 100% and 75% words correct for the isolated and connected tasks respectively (test 11). The best open vocabulary performance was 95.5% and 93.6% on the same tasks. For the isolated task the poor performance for open vocabulary training is due to the absence of utterance initial and final contexts since the sub-word models were trained on connected data. For the connected task a significant drop in performance was observed for the closed vocabulary training since the HMM word models were trained on isolated data. In order to obtain comparable performance with the open vocabulary training, a larger training database based on connected utterances would be required which provided adequate coverage of each word in the test vocabulary.

5. LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis (LDA), when applied to speech signals, is used to derive a linear transformation that will convert the original acoustic front-end parameters to a reduced representation, which will preserve as much information about the sounds which have been spoken, as well as improving the discrimination between sound classes. One of the advantages of LDA is its ability to combine heterogeneous sets of parameters into a set of discriminant functions [2] e.g static and dynamic coefficients. Hunt [2] has developed the IMELDA representation which combines LDA with a MEL-scale front-end representation; IMELDA has been shown to out-perform many other representations, to be robust for a wide range of distortions and to be computationally simple. IMELDA finds the directions in parameter space that maximize the ratio of the between-class to the within-class statistics where the individual classes are not known. The resulting transformation has two major benefits (i) the average within-class covariance matrix is the identity matrix providing better statistical modelling and allowing Euclidean distances to be used and (ii) only the transform vectors whose directions maximize the discrimination between sounds are retrained. This reduces the effect of statistical noise and there are fewer variables to be trained. As a result less training data is required to obtain the same performance.

The IMELDA transform was computed for the 200 training sentences using the context-independent phonicle models to obtain the within and between class statistics. The least important 14 dimensions were eliminated resulting in 12 element transformed feature vectors. Table 1 summarizes the experiments performed to evaluate the performance of the IMELDA representation for the context independent model (L1 - Test 8) and the reduced biphone context (L2 - Test 9). Comparison of the performance with and without the IMELDA transformation (tests 1& 2) show a significant improvement in recognition performance when IMELDA is applied.

6. CONCLUSIONS

In this paper we have described two techniques for improving the performance of sub-word recognition with open vocabularies : a new sub-triphone unit called a phonicle and an implementation of the IMELDA representation in the sub-word domain. The results achieved for phonicle context modelling were very encouraging compared to the conventional triphone approach; the best performance achieved was 95.5% and 93.5% on the isolated and connected tasks without a grammar (perplexity=98). Improved performance was also obtained when using an IMELDA transformation. Future work will be involved in extensions and enhancements to the current sub-word recognition system. Recognition evaluations will also be performed on a larger number of speakers.

ACKNOWLEDGEMENTS

This work was conducted for Marconi Speech and Information Systems as part of the Alvey Integrated Speech Technology Demonstrator (ISTD) Project with CSTR and HUSAT. The databases were recorded at the Centre for Speech Technology Research (CSTR), Edinburgh.

REFERENCES

[1] H W Hon & K-F Lee, "On Vocabulary-Independent Speech Modelling", ICASSP 90, Albuquerque, 1990.

[2] M J Hunt & C Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", ICASSP 89, Glasgow, Scotland, 1989.

[3] L C Wood & D J B Pearce, "A Comparison of the Performance of Whole Word HMM Recognisers", elsewhere in these proceedings.

[4] K-F Lee, "Large Vocabulary Speaker Independent Continuous Speech Recognition", Phd Thesis, Carnegie Mellon University, 1988.

[5] D B Paul, "The Lincoln Laboratories Robust Continuous Speech Recogniser", ICASSP 89, Glasgow, Scotland, 1989.

[6] K-F Lee et al, "Allophone Clustering for Continuous Speech Recognition", ICASSP , Albuquerque, 1990.

[7] F R Mc Innes, D Mckelvie & S M Hiller, "The Structure, Strategy and Performance of a Modular Continuous Speech Recognition System", In these proceedings.

[8] M J Hunt, "Figures of Merit for Assessing Connected Word Recognisers", Proceedings of the ESCA Tutorial Day and Workshop, Netherlands, September 1989.

[9] M J Russell et al, "The ARM Continuous Speech Recogniser", ICASSP 90, Albuquerque, 1990.

[10] "SCRIBE: Spoken Corpus Recordings in British English", IED/SERC Pilot Project, August 1989.

[11] S B Davis & P Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. ASSP, August 1980.

TABLE 1 RECOGNITION EXPERIMENTS IN PHONICLE CONTEXT MODELLING

TEST	CONTEXT UNIT	INITIAL ESTIMATES	COVERAGE %		ISOLATED TEST	CONNECTED TEST PERPLEXITY = 98		
			(a)	(b)	%	%WC	%WA	%WWA
T1	L1 MONOPHONE	-	100	100	84.7	78.0	50.4	64.2
T2	L2 REDUCED BIPHONE	L1	88.0	98.6	89.1	89.1	67.7	78.5
T3	L3 BIPHONE	L2	48.0	88.5	94.6	92.0	69.5	80.9
T4	L4 REDUCED TRIPHONE	L3	30.0	64.2	95.2	93.1	71.0	81.5
T5	L5 TRIPHONE	L4	18.4	33.0	95.5	93.6	72.0	82.0
T6	L5 TRIPHONE	L1	18.4	33.0	85.5	80.2	54.3	67.3
T7	L5 TRIPHONE	L1	100	66.0	76.7	75.0	51.0	63.0
T8	L3 BIPHONE	L1	48.0	88.5	93.6	91.3	68.6	79.0
T9	L1 + IMELDA	-	100	100	87.2	79.3	51.9	65.7
T10	L2 + IMELDA	L1	88.0	98.6	91.5	90.1	68.0	75.3
T11	CLOSED VOCABULARY	-	100	100	100	75.0	64.7	73.3

Notes:
(a) : Percentage of contexts occurring three or more times in the training data
(b) : Percentage test set coverage of contexts (a)
WC = % Words Correct, WA = % Word Accuracy, WWA = % Weighted Word Accuracy