MULTI-PHONE STRINGS AS SUBWORD UNITS FOR SPEECH RECOGNITION

P. O'Neill S. Vaseghi B. Doherty W.H. Tan P. McCourt

The Queen's University of Belfast, N. Ireland.

ABSTRACT

The choice of speech unit affects the accuracy, complexity, expandability and ease of adaptation of ASRs to speaker and environmental variations. This paper explores a method of subword modelling based on the concept of multi-phone strings. The motivation in using the longer duration multi-phone strings is to reduce the loss of contextual information, cross-phone correlation, and transitions. Multi-phone strings are an alternative to context-dependent phones and they include many of the syllables. An advantage of muti-phone units is the existence of more than one valid multi-phone transcription for each monophone sequence, this can be used to improve ASR accuracy. A particular case of multi-phone strings namely phone-pairs is investigated in detail. Experimental Evaluation on TIMIT and WSJCAM0 are presented.

1. INTRODUCTION

In speaker independent large vocabulary continuous speech recognition (LVCSR) the choice of speech unit has a significant impact on the accuracy, complexity, expandability, and ease of adaptation to speaker or environmental variations [1]. Longer duration units are better at modelling within unit context and transitions but they lack generality. However, words units are impractical for LVCSR due to the large amount of training data needed. Furthermore expansion of the vocabulary of a word-based system requires considerable effort in collection and processing of training examples of the new words. A further difficulty is the high level of computational power required in adapting a large number of words to speaker or environmental variations.

Subword units such as syllables or phonemes offer a solution as words can be composed from combinations of a manageable number of subword units. There are two broad classes of subword linguistic units; syllables and phonemes. Syllables are smaller than words, both in number and duration, but there is still too many of them and like words they lack generality. There are several variants of syllables such as demi-syllables or diphones, but these have not been adopted widely partly because they lack the 'elegance' and the accessibility of the phone units.

Phoneme units, or their acoustic realisations phones, are the basic elementary linguistic units. There are about 40-60 phonetic units in various dialects of the English language. The current popularity of phonetic units has several reasons; they are few in numbers, the phonetic linguistic theory is well developed and accessible, there are a number of large vocabulary word-to-phone transcription dictionaries, and there are phonetic transcription of large vocabulary speech data base such as TIMIT, WSJ and WSJCAM0 [5].

Phonetic units do not occur in isolation, they are realised within the context of a word or across the word boundaries. Furthermore the boundaries between neighbouring phones are fuzzy in that there are inter-phonetic transition regions where the phones overlap and these regions carry important information for classification of speech sounds. This observation led to the development of context dependent phones such as bi-phones, where the influence of the left or right contexts are included, and tri-phones where the influence of both the left and right context are modelled [2-4]. Context-dependent models have the same structure as context-independent models but are trained on the training data with the relevant context. In order to equip an LVCSR system with sufficient resolution to differentiate between various triphones a large number of triphone models of the order of several thousands need to be used. The actual number used defines the phonetic resolution of the system.

Phoneme (or phone) length units are currently the most widely used speech units. However, part of the current research efforts for the development of high performance speaker-independent speech recognition is concerned with developing a better speech unit. In this

paper we investigate multi-phones as an alternative subword unit to biphones, triphones and syllables.

2. MULTI-PHONE STRINGS

In general longer duration units are better at encoding correlation, transitions, and intra-unit contextual information. Multi-phones combine the simplicity of phones with the accuracy that a longer duration model provides. In comparison with the monophone units, multi-phone units offer the following advantages : (a) the loss of contextual information is reduced, (b) the loss of inter-phonetic correlation and transitions is reduced, (c) multi-phones are longer and hence the problem of some phonetic units being too short is alleviated, and (d) many syllables are two or three phones long and are contained in the multiphone sets. In the followings we investigate phone-pair strings as speech units. The main difference between the phone-pair and the context-dependent biphone is that a phone-pair simultaneously includes the influence of both the left and the right phones. The usual method for the simultaneous inclusion of the left and the right context is to use the triphone models, the phone-pairs with their two transcription sequence per phone, offer an alternative biphones and to triphones.

2.1 Transcription of Monophone Sequences To Phone-Pair Sequences

As an example of the conversion of a monophone sequence to a phone-pair sequence consider the following monophone transcription of “Even if She” taken from the 39-phone set TIMIT database

0	945625	sil
1871250	2775000	iy
2775000	3175000	v
3175000	3916250	ih
3916250	4647500	n
4647500	6475000	eh
6475000	7711250	n
9325000	9825000	ih
9825000	10025625	f
10025625	10950000	sh
10950000	11437500	ih
11437500	11837500	sil .

The above monophone transcription sequence can be converted into two different phone-pair sequences as shown below :

(a)

0	945625	sil
1871250	3175000	iy_v
3175000	4647500	ih_n
4647500	7711250	eh_n
9325000	10025625	ih_f
10025625	11437500	sh_ih
11437500	11837500	sil

An alternative phone-pair sequence transcription is

(b)

0	2775000	sil_iy
2775000	3916250	v_ih
3916250	6475000	n_eh
6475000	9825000	n_ih
9825000	10950000	f_sh
10950000	11837500	ih_sil.

The existence of two valid transcriptions implies that there are two different possible solutions at the output of the decoder. The transcription of a phone-pair sequence to back a monophone sequence, if required, is a simple one to one mapping process.

2.2 Comparison of Phone-Pairs and Biphones

In this section we explore the similarities and the differences between the context-dependent biphones and the proposed phone-pairs. For illustration consider the following sequence of phonetic symbol

sil a b c d e f g sil

This sequence can be transcribed in phone-pairs as

sil a_b c_d e_f g_sil

Now compares the phone-pair transcription with a left-context biphone transcription

sil sil-a a-b b-c c-d d-e e-f f-g sil

and right-context biphone transcription

sil a+b b+c c+d d+e e+f f+g g+sil sil

In the above notation the minus sign (-) is used to denote a left context, the plus sign (+) a right context, and the underscore (_) a phone-pair. A biphone model is usually a 3 state HMM. For a phone-pair a six state left-right HMM is normally used. A main question is the extent to which a phone pair a_b can be constructed from combination of a right biphone a+b and a left biphone a-b. In the embedded mode of training HMMs, the accumulators for estimation of the respective state observation models of the phone-pair and two the biphones should contain similar quantities. However,

the transition matrix of a phone pair is quite different from that constructed from a left and a right biphone. In the recognition phase, the main difference between the phone-pair and the biphone dictionary transcription is that a phone-pair simultaneously includes the influence of the left and the right phones. However the outer states of a phone-pair HMM are context independent.

2.3 Using Redundancies in Phone-Pair Transcription

Corresponding to the two alternative phone-pair transcriptions for each monophone sequence, discussed above, there are potentially two valid phone-pair sequences at the output of the phone-pair based speech decoder. As shown in the evaluation part of this paper, these two sequences can be used to improve recognition. The simplest method is two have two phone-pair transcriptions for every word in the dictionary. This method would double the size of the dictionary, however the increase in the computational load can be reduced by using a narrower search beam during the speech decoding process. In evaluations, the use of double transcription has resulted in improvement in accuracy.

2.4 Training Data Scarcity and Unseen Phone-Pairs Problems

Depending on the phone set used there are between 1600 to 3600 phone pairs. In practice some phone-pairs do not occur or happen infrequently. In the TIMIT data base using a 39-phone set we observed 1092 different phone-pairs in the training data. There were 28 phone-pairs in the test data which were unseen in the training data. In WSJCAM0 there were about 1500 phone-pairs. The distribution of the number of examples of the phone-pairs in the training database were non-uniform with more than 12 phone-pairs having no more than one example and some 200 having less than 5 examples in the training data.

The conventional practice for dealing with training data scarcity is to cluster acoustically similar states of phonetic models so that training data are shared between these states. With phone-pairs we have also explored constructing phone-pair models, for unseen phones or for phones for which there is too little training data available, by simply concatenating the corresponding monophone models.

2.5 Extension to Phone-Triples

The idea of using phone-pairs as subword units can be extended to include phone-triples. But in practice there will be too many phone-triples and insufficient data to estimate them all. However data clustering can be used to alleviate insufficient training data problem. Further more one could consider a system that selectively combines phone-pairs and phone-triples to get the best results.

3. EVALUATIONS

The evaluations were performed on the TIMT and WSJCAM0 data bases using the hidden Markov model (HTK) tool kit [1]. The speech sampling rate for both databases is 16kHz. Speech was segmented into frames of 25 ms length, with a frame overlap of 10 ms. Each frame was converted to 13 MFCCs (including the zeroth's coefficient) and augmented with the delta and acceleration coefficients.

The phone pairs were modelled with 6 state left-right HMMs. The distribution of each the feature space in each state is modelled with a mixture Gaussian probability.

Evaluation on the TIMIT Database

Table 1 tabulates the recognition and accuracy of the phone-pair units for the TIMIT continuous speech data base as a function of the increasing number of mixture components of the state Gaussian observation model. In table two insertion penalty and a phone-bigram model are used to reduce the number of insertions and to narrow the large gap between the recognition and the accuracy for the 15 Gaussian mixture/state HMM.

Table 1

TIMIT database, Phonepair experiment,

	Recognition	Accuracy
1mix	71.02	47.11
3mix	75.05	54.72
5mix	76.57	58.15
8mix	77.54	59.94
10mix	77.77	60.55
12mix	78.11	60.98
15mix	78.08	60.93

Table 2

The effects of a matrix bigram (**MB**) and insertion penalty (**IP**) on monophone results'

15 Mixtures	Recognition	Accuracy
Original Results	77.74	61.18
MB + IP = -30	72.49	65.08
MB + IP = -40	71.11	65.27
MB + IP = -50	69.97	65.12
MB + IP = -60	68.34	64.13

Evaluation of word recognition on WSJCAM0 5K Vocabulary

The phone pairs were evaluated for word recognition on the WSJCAM0 UK English data base. In these evaluations the 45 phone set BEEP transcription dictionary was used for word to phone transcription, the dictionary was easily adapted to word to phone-pair transcription. The language model used is the 5 k bigram model from the US WSJ database.

Table 3 and 4 present the word recognition and accuracy obtained from the context dependent biphone units and the phone-pair units respectively. The phone-pairs compete well with biphones. Table 4 shows the evaluation results with phone-pairs for the two cases of single entry per word dictionary and two phone-pair transcription per word dictionary. The use of two phone-pair transcription per word improves the recognition and accuracy considerably.

The last row of table 4 shows the result of using a dictionary in which three transcription per word were used. The transcriptions included two alternative phone-pairs and one monophone transcription. This method provides significant improvement in performance.

Table 3 - Left-Context Biphones

12 Mixture Results	**Word**	
	Recognition	Accuracy
1551 biphone set, t=250.0, s=5.0, p=-15.0	70.39	66.97

Table 4 - Phone-Pairs

15 Mixture Results	**WORD**	
	Rec	Acc
t=250.0, s=5.0, p=-15.0, SED,	70.89	67.69
t=250.0, s=5.0, p=-15.0, DED	76.85	73.17
t=250.0, s=5.0, p=-15.0, TED	79.34	76.63

NOTE:
SED = Single Entry Dictionary (One form of phonepair transcription)
DED = Double Entry Dictionary (Both forms of phonepair transcriptions)
TED = Triple Entry Dictionary (Double Entry Dictionary + Monophone Dictionary)

4. CONCLUSION

The choice of speech unit can have a significant impact on the performance of speaker independent LVCSR. The search for an optimal subword speech unit, one that preserves all the discriminative acoustic information, remains an important goal in LVCSR. Multi-phones extend the phone unit concept and also include many of the syllables. Longer than phone duration units such as multi-phones are generally better at modelling context and the within unit transitions, but cross unit context are difficult to model because with longer duration units there are too many context dependent models. The experimental focus in this paper was on the use of phone-pairs. Phone-pairs are two phone duration long. Like biphones they require much less training data compared to triphones and can be estimated robustly. Evaluations of phone-pairs on the TIMIT and WSJCAM0 demonstrated that they compete well with biphones.

REFRENCES

[1] S. Young (1996), "A Review of Large-Vocabulary Continuous-Speech Recognition", IEEE Signal Processing Magazine, Vol. 13, No. 5, pp. 45-57.

[2] Lee K-F. (1990) "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition", IEEE Trans. ASSP, Vol. 38, No. 4, Pages 599-609.

[3] Jershaw D. J., Phonetic Context-Dependency in a Hybrid ANN/HMM Speech recognition System, PhD Thesis Cambridge University 1997.

[4] Odell J. J. The Use of Context in Large Vocabulary Speech Recognition, PhD Thesis, Cambridge University 1995.

[5]Linguistic Data Consortium, http://www.ldc.upenn.edu