International Conference on
Information, Communications and Signal Processing
ICICS '97
Singapore, 9-12 September 1997

Connected Phoneme HMMs with Implicit Duration Modelling for Better Speech Recognition

Sitaram Ramachandrula and Sreenivas Thippur

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, India
email: sitaram@protocol.ece.iisc.ernet.in & sreeni@iis.fhg.de

Abstract

The duration of speech units is an important cue in speech recognition. But most of the current speech recognizers, based on HMMs, do not use this durational information satisfactorily. Previously the duration was incorporated into HMM based systems by modifying the HMM state duration modelling ability, but this has some limitations. In this paper we propose a better way of using duration of speech units in HMM based systems. Here, the implicit duration modelling ability of the whole HMM is exploited. Connected phoneme HMMs with implicit duration modelling are proposed as better word models, than those that are trained using whole word speech. In a speaker independent, isolated word recognition experiment, having confusable words in the vocabulary, the word HMMs formed by concatenating phoneme HMMs with duration modelling, have improved the word recognition accuracy by 7-8 %, compared to word HMMs trained using whole word speech.

1 Introduction

It is well known that duration of speech units, like phonemes, provides an important cue in speech recognition [1]. This durational information plays a major role in scenarios where some words in the vocabulary of speech recognizer have confusable pairs, e.g., rider vs. writer, beat vs. bit, back vs. pack, killed vs. kilt, etc. In all these word pairs, the duration of the differentiating phoneme is the only discriminating feature and all other spectral features do not help much, as they are very similar. Therefore any speech recognizer with these sort of confusable words in the vocabulary should use the duration of the differentiating phonemes in recognition.

Presently the highly successful methods of speech recognition are based on hidden Markov models (HMM) [2]. Considering the case of isolated word recognition using HMMs, for each word in the vocabulary a HMM is constructed, either by training the HMM by using repetitions of the whole word speech data or by concatenation of phoneme HMMs which constitute the word. Once all the word HMMs are thus formed by following any one method, the recognition of a test word is done by finding the likelihood of it from each word model, where the maximum of these likelihoods corresponds to the recognized word. This method could also be applied to isolated speech, confusable word recognition, but this will not be successful unless the durational information of the discriminating phonemes in the words is used in word HMMs.

Capturing of durational information (of speech units) into HMMs has been an active research problem for more than a decade. The researchers have mainly concentrated on improving the state duration modelling of HMM, after realizing that the basic HMM state implicitly models a geometric duration distribution, which is not suitable for speech events. Many solutions for this problem have been proposed, like Semi-HMM [3], In-homogeneous-HMM [4], etc, where the states are made capable of achieving arbitrary duration distributions. But in all these solutions there is still only one observation pdf per state, which cannot effectively model the spectral variation of even a phoneme, as phoneme can be non-stationary e.g., stops, diphthongs, (a single observation pdf can model only a stationary segment). A non-stationary phoneme requires more than one observation pdf, i.e., more than one HMM state to characterize the spectral manifestation along its length, then state duration modelling may not help much, as each state will be characterizing the spec-

0-7803-3676-3/97/$10.00

tral and durational variability of only a part of a phoneme. Because of this reason and high computational complexity, speech recognition using these new HMMs with state duration modelling, was not pursued further, though the performance was slightly better than the ordinary HMM [5].

The emphasis, therefore, should be on allocating more HMM states to a phoneme (for capturing non-stationarity) and at the same time capture the total duration spent in them. This is equivalent of saying, allocate a HMM for a phoneme and model duration in it. Following this requirement, in this paper an investigation into the implicit duration modelling ability of a HMM is done, and it was found that the implicit duration distribution of HMM can be suitable for speech events. Later using this property a solution is proposed to the confusable word recognition problem, which is tested experimentally. Here, we discuss the implicit duration modelling of HMM in Sec. 2., and propose a way of using it advantageously for above mentioned confusable word recognition in Sec. 3., followed by experimental evaluation in Sec. 4., and conclusions in Sec. 5.

2 Implicit duration modelling in HMM

Unlike the case of duration in a single HMM state, here the duration spent in the entire HMM, i.e., the summation of durations spent in all of its states is of interest. As a typical case, here we discuss the implicit duration modelling in a strict left-right HMM (LR-HMM), shown in Fig. 1. To discuss the probability of duration spent in a single HMM state i, there is a requirement of an exit-state transition probability $(1 - a_{ii})$, otherwise its duration distribution will become flat until infinite duration [2], (here a_{ii} is self transition probability of state i). Similarly to discuss the duration spent in the HMM, it should have an exit-HMM transition probability. In the LR-HMM considered in Fig. 1, the exit-HMM probability is $(1 - a_{NN})$, as it is only possible to exit LR-HMM from its last state N. In a LR-HMM normally the $a_{NN} = 1$, but here it is assumed to be in the range $(0 < a_{NN} < 1)$, otherwise the exit HMM probability, $(1 - a_{NN})$, will become zero.

As the duration d spent in the LR-HMM, λ, is the summation of durations $d_1, d_2, ...d_N$, spent in each of its states, the HMM duration distribution, $P_\lambda(d)$, will be the convolution of duration distributions of each state $P_1(d), P_2(d), ..., P_N(d)$, [6] (which are all

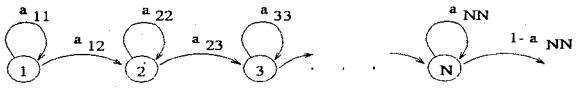

Figure 1: A Left-Right HMM

geometric distributions). Evaluation of this convolution is very tedious, a simpler way of finding this distribution $P_\lambda(d)$, is by the summation of probabilities of all possible state sequences of length d in the given HMM, such that at $d+1$ th instant a transition is made out of last state N of HMM.

$$P_\lambda(d) = \sum_{q_1,q_2,\cdots,q_d} \pi_{q_1} a_{q_1,q_2} a_{q_2,q_3} \cdots a_{q_{d-1},q_d} (1 - a_{q_d,q_d}) \tag{1}$$

such that $q_d = N$.

Here $q_1, q_2, ...q_d$ denote the states occupied at time instants $1, 2, ...d$, and the initial state occupancy probability, $\pi_{q_1} = 1$, only if $q_1 = 1$, as it is a LR-HMM. Equation (1), can be easily evaluated for a given HMM by an algorithm similar to *Forward* algorithm [2], which is given below:

Algorithm:

$$Defining,\ \delta_t(i) = \sum_{q_1,q_2,\cdots,q_t} P(q_1, q_2, ..., q_t = i/\lambda), \tag{2}$$

i.e., the sum of probabilities of all state sequences of length t, such that the state occupied at time t, $q_t = i$ in the given HMM, $\lambda(A, B, \Pi)$

Initialization: For $1 \leq i \leq N$

$$\delta_1(i) = \pi_i \tag{3}$$

Recursion: For $1 \leq t \leq T - 1$ and $1 \leq j \leq N$,

$$\delta_{t+1}(j) = \sum_{i=1}^{N} \delta_t(i) a_{ij} \tag{4}$$

Now the duration distribution of LR-HMM $P_\lambda(t)$, can be found by using δ's of last state N of the LR-HMM, for all time instants t, i.e.,

$$P_\lambda(t) = \delta_t(N)(1 - a_{NN}) \tag{5}$$

Here, $1 - a_{NN}$ gives the transition probability out of LR-HMM i.e., out of last state of LR-HMM. If the HMM considered is not left-right type, then there may be more than one terminating states N_i, then the Eq. 5 is evaluated and added for all these states using the respective exit-HMM probabilities, to get the HMM duration distribution.

Now taking a strict LR-HMM, with equal self transition probabilities ($a_{ii} = 0.9$) for all states, the duration distribution $P_\lambda(d)$ is calculated according to the above algorithm, and plotted in Fig. 2, for different number of states N. It can be seen from Fig.

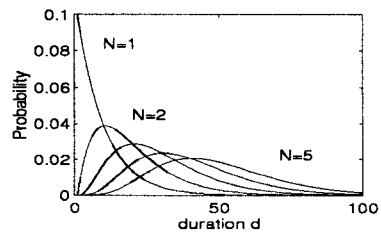

Figure 2: Some duration distributions possible by HMM.

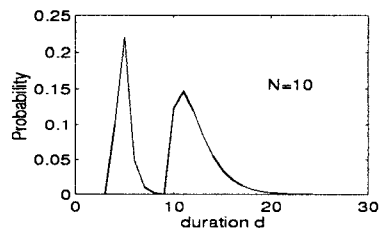

Figure 3: Bi-modal duration distribution

2, the shapes of duration distributions modelled by LR-HMMs for $N \geq 2$, are suitable for speech events. The mean and variance of these distributions could be changed by varying the number of states and the state transition probabilities. It is also possible to have multi-modal durational distributions, (see Fig. 3), by just having a bigger skip transitions in the LR-HMM of Fig. 1, though this sort of duration distributions are not common in speech units. Thus it is seen that an ordinary LR-HMM can implicitly assume arbitrary duration distributions suitable for speech events.

Now the question arises, given a speech data, how to fix the HMM parameters (especially transition probabilities) such that the duration distribution of data is modelled implicitly by the HMM. We already knew one method of estimation of HMM parameters by Baum-Welch algorithm, [2], which maximizes the likelihood of training data, does this yield the HMM which also captures the duration of data ? It is experimentally observed that, by careful choice of number of states and their connectivity in the LR-HMM, based on the speech unit being modelled, the Baum-Welch training takes an initial random HMM with some arbitrary implicit duration distribution, to a final HMM which implicitly models the duration distribution of the training data and at the same time has the maximum likelihood for training data. This is illustrated in Fig. 4, where the duration distributions of actual data and initial random HMM and final trained HMM are given, it can be seen that the initial HMM duration distribution doesn't fit with the actual data, but after Baum-Welch training it fits well. Now equipped with this knowledge we can propose a solution for previously discussed confusable word recognition.

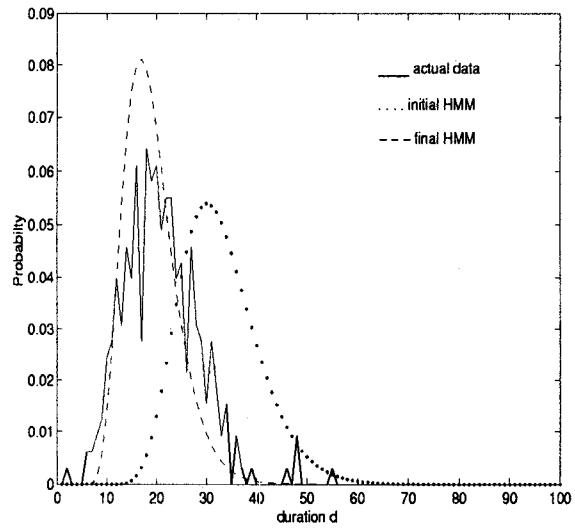

Figure 4: Duration distributions of actual data, initial HMM and final HMM

3 Connected Phoneme HMMs for Confusable Word Recognition

The confusable word recognition using the HMMs is now obvious. Taking the case of discriminating words 'bit' and 'beat' with phonemic spellings /b//ih//t/ and /b//iy//t/ respectively. First LR-HMMs for each of the phonemes in these words are trained such that the HMMs implicitly capture the duration of the phonemes (as was discussed in previous section). In this case phoneme HMMs for /ih/ and /iy/ are only needed to be constructed such that they capture the duration, as the other phonemes are anyway same in both words. But it is better to have all the phonemes in the words to have HMMs

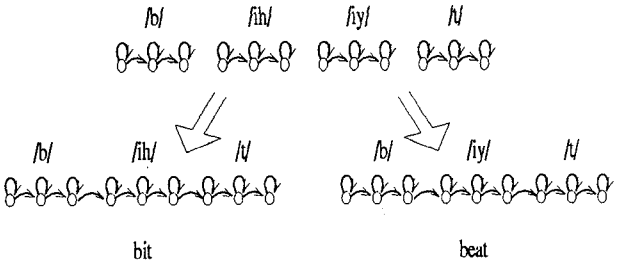

Figure 5: Formation of word HMMs by connecting phoneme HMMs

which also model their durations. Once the phoneme HMMs are available they are concatenated according to phonemic spelling of the words 'bit' and 'beat' into word HMMs, Fig. 5, (the state transition probability connecting two phoneme HMMs was empirically chosen to be 0.1).

The word HMMs thus formed capture the duration of the discriminating phonemes compared to word HMMs which were trained directly on whole word data. Also, the temporal structure within the word is well captured in connected HMMs. Once the word HMMs are thus formed by connecting phoneme HMMs, the recognition can be carried out by standard techniques.

4 Experimental Evaluation

4.1 Task

The idea of connected phoneme HMMs with duration modelling is tested on a speaker independent isolated word recognition task, with the vocabulary containing confusable words. Once the idea is tested on isolated speech, the same solution can be applied to connected speech also. The isolated word vocabulary chosen for this experiment is {rider, writer, beat, bit, cold, gold, pack, back, try, dry, killed, kilt, came, game}. These confusable words are mainly chosen as phoneme duration has to be used in recognition.

4.2 Database

The present idea requires training of phoneme HMMs constituting each word in the vocabulary. The isolated speech database should therefore have phonemic labelling to facilitate the construction of phoneme HMMs separately. As there is no such standard database available for isolated speech, an artificial database is created using TIMIT acoustic-phonetic database. TIMIT contains continuous speech sentences, sampled at 16 KHz and labelled into phonemes. Now each of the 14 words in the task vocabulary are artificially created by concatenating the speech segments of corresponding phonemes from TIMIT according to the phonemic spelling of the words given in Table. 1. While, thus forming a word, always phonemes segments of same speaker are concatenated together. The dialect 2 of TIMIT is used for creating this database. For each of the 14 words, 52 training occurrences are created, each from a different speaker of training set of dialect 2. For testing purpose, 26 occurrences of each word are created from unseen speakers, from the testset of dialect 2, again each occurrence by a different speaker. Though the words thus formed are artificial and may not sound well, they can still be used as speech patterns for recognition as they have similarity within a word class and differences across word classes, which is enough for testing any pattern recognition idea. Also the database thus formed meets our requirements of duration sensitive words with phonemic labelling. Here the words are not exactly spoken in isolation, but are artificially created separately. Actually these artificial word patterns are much difficult to recognize than original speech as they are created from continuous speech having high coarticulation effects, and also each phoneme will have high variability as the contexts from which each occurence of it is taken may be different. Any recognition improvement shown on this difficult database will be more pronounced on natural speech, which has lesser variability as the context of a phoneme with in a word is always fixed. Here on, this artificial database is referred to as isolated word database with phonemic labelling.

4.3 Preprocessing

The entire speech database, created as explained above is analysed in frames of 16 ms with an overlap of 8 ms between frames. From each frame of speech, 18 LPC derived cepstral coefficients are determined after pre-emphasizing with a factor of 0.95. Feature vectors of all the speech sentences are later VQ coded using a VQ codebook having 256 codewords, which is designed using LBG algorithm.

4.4 Construction of word HMMs

First, using the VQ sequences of each phoneme, obtained using the labelling information of speech database, phoneme HMMs are trained. The number of states and connectivity in these HMMs are fixed

Table 1: Phonemic Spellings of the Words used

Word	Phoneme spelling
writer	/r//ay//t//axr/
rider	/r//ay//d//axr/
beat	/b//iy//t/
bit	/b//ih//t/
killed	/k//ih//l//d/
kilt	/k//ih//l//t/
back	/b//ae//k/
pack	/p//ae//k/
cold	/k//ow//l//d/
gold	/g//ow//l//d/
came	/k//ey//m/
game	/g//ey//m/
try	/t//r//ay/
dry	/d//r//ay/

Table 2: Word Recognition Accuracy

Word HMMs used	Train Set	Test Set
Whole word trained	96.84%	59.61%
Connected HMMs	96.02%	67.03%

such that the phoneme duration distribution is implicitly modelled by the HMMs, after Baum-Welch re-estimation. Later the HMMs for each of the 14 words in vocabulary are constructed by concatenating these phoneme HMMs as explained in Sec. 3 (see Fig. 5). For comparison, another set of word HMMs for each of 14 words having the same number of states as in the connected HMMs described above are trained using the whole word speech, without using phoneme labelling information.

4.5 Recognition results

Isolated word recognition results of test and train set, with both sets of word HMMs, are given in Table. 2. It can be seen that the connected phoneme HMMs have improved the word recognition accuracy by 7 to 8% on test set compared to whole word trained HMMs, though the training set performance is similar.

5 Conclusions

In this paper, the idea of linear concatenation of phoneme HMMs with implicit duration modelling, is proposed as better method of word HMM construction than the word HMMs trained using whole word data. This idea has shown an improvement of 7 to 8 % on word recognition accuracy on a confusable word recognition task, where the phoneme duration plays a major role. It can be concluded that connected phoneme HMMs with implicit duration modelling are better for any scenario of speech recognition as the temporal structure is well captured at no increment in computational complexity.

References

[1] Ratnayake, M., Savic, M., and Sorensen, J., " Use of Semi-Markov models for speaker independent phoneme recognition," Proc. ICASSP'92, pp. I-565 - I-568, 1992

[2] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of IEEE, vol. 77, no. 2, pp. 257-285, Feb. 1989

[3] Ferguson, J. D., " Variable duration models for speech " In Proc. Symp. on Applications of HMMs to text and speech, J. D. Ferguson Ed. Princeton, NJ, pp. 143-179, 1980

[4] Ramesh, P., and Wilpon, J. G. "Modelling State Durations in Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, pp.I-381–I-384, 1992.

[5] Russel, M. J., and Cook, A. "Experimental evaluation of duration modelling techniques for automatic speech recognition," Proc. ICASSP. **4**, pp. 2376–2379, 1987.

[6] Papoulis, A., "Probability, Random variables and stochastic processes", McGraw hill, Inc., 1991