LEXICON-BUILDING METHODS FOR AN ACOUSTIC SUB-WORD BASED SPEECH RECOGNIZER

K.K. Paliwal
Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay-400005, India
(Current address: Acoustics Research Dept.,
AT&T Bell Labs., Murray Hill, NJ 07974, USA)

ABSTRACT — Although remarkable performance of word-based speech recognition systems has been demonstrated in small vocabulary tasks, extrapolating it to large vocabulary applications is not straightforward due to training problem. In order to overcome this problem, we use in the present paper the acoustic sub-word units (ASWUs) for speech recognition. We address here the problem of designing word lexicon in terms of ASWUs. Different methods for generating the deterministic-type and the statistical-type of word lexicons are proposed. These methods are evaluated for the recognition of isolated words in a speaker-dependent mode and the results are discussed.

1. INTRODUCTION

Although the research in the area of speech recognition has been pursued for the last 40 years, only the whole word unit (WWU) based speech recognition systems have achieved commercial success so far. The main reason for their success is that they can incorporate explicitly the effects of inter-phonemic context dependence and coarticulation in their word models. Because of this, these systems show remarkable performance for small vocabulary tasks (vocabulary size less than 100 words) [1,2]. But, it is not a straightforward task to extend these systems for the large vocabulary applications (vocabulary size greater than 1000 words). There are several reasons for this; the main one is associated with training. Acoustic characteristics of a word at its boundaries get strongly affected by the preceding and the following words. Therefore, in order to adequately train the speech recognizer, it is necessary to have the training set of speech data where each word appears several times in all the possible phonetic contexts. For small-vocabulary speech recognition systems, it is possible to design such a training set [3]. But, it is not possible to have such a training set for the large vocabulary speech recognition systems because the amount of training data becomes prohibitively large. Therefore, for the large vocabulary speech recognition applications, it becomes necessary to use sub-word units (SWUs) — units smaller than the WWU.

In the SWU-based speech recognition systems, the word models are constructed by concatenating the sub-word models using the word lexicon. This decomposition of words into sub-word sequences removes the dependence of the training data size on the number of words in the vocabulary. The number of SWUs to be used in the speech recognition system is independent of word vocabulary, but it depends on the type of SWU chosen. Usually, this number is quite small. In order to adequately train the speech recognizer, one needs here several examples of each sub-word (in contrast to several examples of each word needed in the WWU-based speech recognition system). Thus, the SWU-based speech recognition systems require smaller amount of training data and can be used for large vocabulary applications. However, since the SWU-based speech recognition systems can not incorporate in their sub-word models the effects of context-dependence and coarticulation as nicely as the WWU-based systems, they are not expected to perform, in principle, as well as the WWU-based systems. Therefore, when one designs an SWU-based speech recognizer, the objective is to approach the performance of the WWU-based speech recognizer. Another disadvantage of the SWU-based speech recognition systems is that they require a word lexicon, while the WWU-based systems do not require the word lexicon.

Traditionally, the SWUs employed in speech recognition have been defined based upon a linguistic description of the language. Typical examples of the linguistic sub-word units (LSWUs) are phonemes, diphones and syllables. The LSWUs (such as phonemes) have the advantage that the word lexicon is available in a ready-made form from a standard dictionary. There is, however, a major problem when it comes to correctly detecting and identifying these units. This is due to the mismatch between the acoustically-based analysis of the actual speech signal and its linguistically-based description in terms of LSWUs. Because of this mismatch, the performance of the LSWU-based speech recognition systems is significantly inferior to that of the WWU-based systems [4].

Recently, acoustically defined SWUs have been used in speech recognition systems [5-11]. These acoustic sub-word units (ASWUs) do not have any one-to-one correspondence with the LSWUs. Segmentation of the speech utterance in terms of ASWUs is done here using a well-defined acoustic criterion. Thus, there is no mismatch problem as encountered with the LSWUs and, hence, the performance of the ASWU-based speech recognition system can be expected to be better than the LSWU-based system. This indeed is the case as shown by Lee et al. [11]. Since there is no one-to-one correspondence between the ASWUs and the LSWUs, the word lexicon is not available here in the ready-made form from a standard dictionary. It has to be designed. How to design the word lexicon in terms of the ASWUs is a major problem. In the present paper, we propose some methods for building the word lexicon in terms of the ASWUs and study their performance on an ASWU-based speech recognizer.

The word lexicon that can be used with the ASWU-based speech recognition system can be either of deterministic type or of statistical type. In the deterministic-type of word lexicon, each word is represented in terms of its few possible pronunciations. Major issues with this type of lexicon are how to choose these pronunciations and how many pronunciations to be used. In the statistical-type of word lexicon, a statistical model (such as the Markov model, the hidden Markov model or the pronunciation network) is used to characterize each word in the vocabulary.

The ASWU-based speech recognition system used in the present paper to study the lexicon-building methods has been described in our earlier paper [10]. In the present paper, we use this system for the recognition of isolated words, but the system is applicable as well for the recognition of continuous speech. In the training phase, our system requires the following operations: 1) Preprocessing of the input speech utterance to convert it to a sequence of LP parameter vectors, 2) Segmentation of LP vector sequence into acoustic segments, 3) Clustering of acoustic segments into N clusters where each cluster corresponds to one ASWU, 4) Generation of hidden Markov model (HMM) for each ASWU using the acoustic segments in its cluster, 5) Generation of the word lexicon. In the recognition phase, the sequence of the LP parameter vectors obtained by preprocessing the input speech utterance is compared with the models of different words in the vocabulary. Here, the model for each word is constructed as a sequence of ASWU HMMs using the word lexicon. Recognition is performed using the maximum likelihood decision rule.

The lexicon-building methods proposed in this paper are evaluated by using the ASWU-based speech recognizer in a speaker-dependent mode to recognize isolated words from the following two vocabularies: 1) Vocabulary V1 containing 9 Norwegian e-set alphabets ('B', 'C',

CH2847-2/90/0000-0729 $1.00 © 1990 IEEE

'D', 'E', 'G', 'J', 'P', 'T' and 'V'), and 2) Vocabulary V2 containing 42 Norwegian alpha-digits (29 alphabets + 10 digits + 3 control words "start", "stopp" and "gjenta"). 120 repetitions of these vocabulary words are recorded over a period of 5 weeks. Two male speakers are used for recording. This speech data base is divided into two sets: 1) the training set containing the first 70 repetitions, and 2) the test set containing the remaining 50 repetitions.

The paper is organized as follows. The ASWU-based speech recognition system used in the present study is described in Section 2. Section 3 describes different methods of generating the word lexicon proposed in the present paper and reports their recognition results. These results are discussed in Section 4. Conclusions are reported in Section 5.

2. THE ASWU-BASED SPEECH RECOGNITION SYSTEM

In this section, we describe the ASWU-based speech recognition system used in the present study. Since this system is already described in our earlier paper [10], we give here only a brief description of this system.

2.1. Preprocessing

Speech utterances of spoken isolated words are lowpass filtered at 3.5 kHz and digitized at 8 kHz sampling rate. Endpoints of the spoken words are detected automatically using an energy criterion with some human supervision [10]. The speech signal is preemphasized using a filter $H(z) = 1 - 0.95z^{-1}$. A 10-th order LP analysis is performed every 15 ms over a 45 ms Hamming window using the autocorrelation method.

2.2. Segmentation

Segmentation is the most crucial step in the training operation of the ASWU-based speech recognizer. The criterion used for segmentation defines the type of ASWUs used in the recognizer. In the present paper, we use piece-wise stationarity in the speech signal as the acoustic criterion for segmentation. According to this criterion, the speech utterance consists of a number of stationary segments where each segment can be represented by its centroid. The maximum likelihood (ML) algorithm proposed by Svendsen and Soong [12] uses this criterion for segmentation. In this algorithm, the speech utterance is segmented into a fixed number of segments (say, M) by minimizing the average intra-segment distortion over all possible segment boundaries. The intra-segment distortion for a given frame is defined here as the distortion between the given frame and the centroid of the segment to which the given frame belongs to. The likelihood distortion measure is used to define this distortion. It might be noted that the average intra-segment distortion for a speech utterance decreases with an increase in M. In our implementation of the ML segmentation algorithm, M is steadily increased until the average intra-segment distortion is less than a predefined threshold which is set here to 0.08.

2.3. Segment Clustering

The ML segmentation algorithm described in the preceding subsection produces a large number of acoustic segments for the data in the training set. These acoustic segments span the speech segment space. Our aim here is to divide this space into N clusters where each cluster corresponds to one ASWU. The clustering is performed by, first, representing the each of the acoustic segments by its centroid and, then, applying the k-means algorithm [13] on these segment-centroid vectors. In the k-means algorithm, the likelihood ratio distortion measure is used to define the distortion between two LP vectors. The clustering procedure generates a codebook having N entries where each entry defines one ASWU cluster.

2.4. Generation of HMMs for ASWUs

Here, the acoustic segments belonging to each of the N ASWU clusters are modeled by a first order HMM. The HMM has three states and is a left-to-right model where single skips between the states are allowed. Single mixture multivariate Gaussian functions are used to characterize the probability density functions of different states. Since some of the ASWU clusters can have a very small number of acoustic segments, it is difficult to estimate reliably all the components of the covariance matrices used in the Gaussian probability density functions. Therefore, we use in the present study only diagonal covariance matrices with the Gaussian probability density functions. For each ASWU, the parameters of the HMM are computed from the acoustic segments contained in its cluster using the Viterbi algorithm [14]. Speech frames are represented here in terms of 10 cepstral coefficients.

2.5. Generation of Word Lexicon

As mentioned earlier, it is necessary to design the word lexicon in terms of ASWUs as it is not available in a ready-made form from a standard dictionary. Since this is the main topic of research in the present paper, it is considered in detail in Section 3.

2.6. Recognition

Here, the speech parameters (10 cepstral coefficients) are computed for each frame of the input speech utterance through 10-th order LP analysis. The parameterized speech utterance is compared with the models of all the words in the vocabulary using the Viterbi decoding algorithm and the recognition is done by applying the maximum likelihood decision rule [14]. In order to generate the model for a given word, the sequence of ASWUs for that word is taken from the word lexicon. The word models are generated by concatenating the corresponding ASWU HMMs.

3. LEXICON BUILDING METHODS AND RESULTS

In this section, we propose some methods for building the word lexicon. These methods are evaluated on the two vocabularies V1 and V2 (described in Section 1) using the ASWU-based speech recognition system and their recognition results are presented.

As mentioned earlier, the word lexicon that can be used with the ASWU-based speech recognizer can be either of deterministic type or of statistical type. In the deterministic-type of word lexicon, a pronunciation in terms of ASWUs is assigned to each word. Since there is considerable variability in the speech signal due to allophonic variations, speaker differences and speaking rates, one pronunciation per word may not be adequate. It may be necessary to assign more than one pronunciation for each word. In the statistical-type of word lexicon, a statistical model is used to describe each word in the vocabulary. These two types of lexicon-building methods are described below.

3.1. Deterministic-Type of Lexicon-Building Methods

In the deterministic-type of word lexicon, each word in the vocabulary is represented by its few pronunciations in terms of ASWUs. In order to determine these pronunciations for a given word, all of its training utterances (70 in the present case) are transcribed in terms of ASWUs. A few of these pronunciations, representative of these training utterances, can be chosen to define the possible pronunciations of the given word. In principle, it is possible to use transcriptions of all the 70 training utterances to define different possible pronunciations of the given word. However, it will be expensive in terms of memory and computation requirements. Also, it is unnecessary to use so many pronunciations per word in the word lexicon as it does not improve the recognition performance. This situation is similar to what one encounters in the WWU-based speech recognition systems where one does not use all the training utterances to define templates for the given word. Instead, a few templates are selected from these training utterances using a clustering procedure (such as the modified k-means algorithm [15]). Thus, the generation of the deterministic-type of word lexicon involves the following two steps: 1) Transcription of training utterances in terms of ASWUs, and 2) Clustering of training of utterances to select a few pronunciations per word. We propose here three different methods for generating the deterministic-type of word lexicon. These methods differ in terms of transcription and clustering procedures. These methods

are described below.

3.1.1. *Method* 1 — Here, the transcription of a training utterance of a given word in terms of ASWUs is done as follows. The training utterance is, first, partitioned into acoustic segments using the ML segmentation algorithm and, then, the acoustic segments are labeled as ASWUs using the N codebook entries (as derived in Subsection 2.3). Clustering is done here by applying the modified k-means algorithm on distances between different training utterances of the given word computed through the dynamic time warping (DTW) technique [15].

Word lexicon generated through this procedure is used with the ASWU-based speech recognizer. Results using 1 pronunciation per word are shown in Table 1 for the V1 vocabulary (with N=64 ASWUs) and for the V2 vocabulary (with N=128 ASWUs) using training data sizes of 50 and 70 repetitions per word. The number, N, of ASWUs used in the system determines how finely the acoustic segment space is sampled; i.e., higher is the value of N, better is the acoustic resolution. In order to see the effect of acoustic resolution, we study the recognition performance as a function of N and the results are shown in Table 2. It can be seen here that the recognition accuracy increases with N, but it saturates very fast; i.e., increasing the acoustic resolution beyond N=64 does not improve the recognition performance much. For small number of ASWUs (i.e., N=16), the recognition results are very poor.

So far we have used in the word lexicon only one pronunciation per word. Now we study the effect of using more than one pronunciation per word on the recognition performance of the ASWU-based speech recognition system. The results are shown in Fig. 1. From this figure, the advantage of using more than one pronunciation per word is not very clear. Therefore, we use hereafter only one pronunciation per word for the deterministic-type of lexicon-building methods.

3.1.2. *Method* 2 — In the recognition phase, the ASWU-based speech recognizer uses the Viterbi decoding algorithm to compute the likelihood of a given word from its model which is constructed as the concatenation of the HMMs of ASWUs defined by its pronunciation. In method 1, this pronunciation is obtained from the transcriptions of word utterances where the ML segmentation algorithm is used for segmentation. This segmentation is inconsistent with the segmentation generated by the Viterbi algorithm during the recognition phase. Therefore, we use here the Viterbi algorithm for transcribing the training utterances of each of the vocabulary words. The Viterbi algorithm uses the HMMs of all the N ASWUs and performs segmentation and labeling in one step. For clustering, we use the same procedure as used in method 1.

This method is applied for generating the word lexicon and the recognition results using one pronunciation per word are shown in Table 3 for the V1 and V2 vocabularies. By comparing this table with Table 1, we can see that this method of lexicon generation gives better recognition results than the method 1.

3.1.3. *Method* 3 — In methods 1 and 2, clustering has been done by applying the modified k-means algorithm on distances between different training utterances of each of the vocabulary words computed through the DTW technique. These distances computed through the DTW technique do not fit well with the likelihoods computed through the hidden Markov modeling approach used in the ASWU-based speech recognizer. Therefore, for clustering the training utterances of a given word, we apply the modified k-means algorithm on likelihoods between different training utterances of this word. For computing the likelihoods for the given word, the Viterbi algorithm is applied, first, on a training utterance of this word to get its transcription and, then, it is applied under the constraint of this transcription on all other training utterances of the given word to get its likelihoods from these training utterances. This procedure is repeated for each of the training utterances of the given word to compute likelihoods which are then used to perform clustering for this word. It might be noted here that the transcription of training utterances which has been done separately from clustering in methods 1 and 2 is a part of clustering operation in this method.

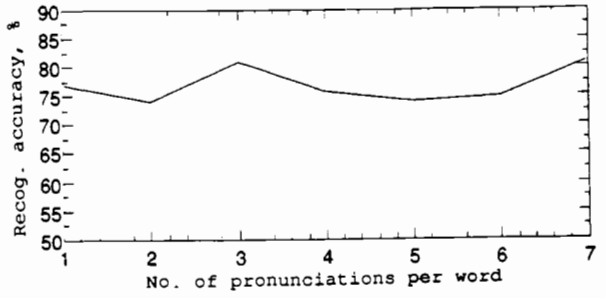

Fig. 1. Recognition accuracy as a function of number of pronunciations per word for the V1 vocabulary. N=64 and number of repetitions used for training is 70.

Table 1: Recognition results from the deterministic-type of lexicon building method 1 (using 1 pronunciation per word) with N=64 for V1 vocabulary and N=128 for V2 vocabulary.

Vocabulary	Recog. accuracy (in %) using training data from	
	50 repetitions	70 repetitions
V1	76.9	76.7
V2	88.4	88.7

Table 2: Recognition results from the deterministic-type of lexicon building method 1 (using 1 pronunciation per word) for the V1 vocabulary using different numbers of ASWUs.

Number of	Recog. accuracy (in %) using training data from	
ASWUs, N	50 repetitions	70 repetitions
16	51.1	51.3
32	66.2	74.7
64	76.9	76.7
128	76.9	79.6

Table 3: Recognition results from the deterministic-type of lexicon building method 2 (using one pronunciation per word) with N=64 for V1 vocabulary and N=128 for V2 vocabulary.

Vocabulary	Recog. accuracy (in %) using training data from	
	50 repetitions	70 repetitions
V1	80.0	84.4
V2	90.0	92.6

Table 4: Recognition results from the deterministic-type of lexicon building method 3 (using one pronunciation per word) with N=64 for V1 vocabulary and N=128 for V2 vocabulary.

Vocabulary	Recog. accuracy (in %) using training data from	
	50 repetitions	70 repetitions
V1	86.2	87.6
V2	91.3	94.0

The word lexicon obtained using this method is applied for recognizing isolated words from the V1 and V2 vocabularies and results for one pronunciation per word are shown in Table 4. By comparing this table with Tables 1 and 3, we can see that this method results in better recognition performance than the methods 1 and 2.

3.2. Statistical-Type of Lexicon-Building Methods

In the statistical-type of lexicon-building methods, a statistical model is assumed and parameters of this model are estimated for each word in the vocabulary from the data in the training set. In the present paper, we assume a first order Markov model for this purpose (though other types of models can also be used as described in the next section). Here, we propose two methods for generating the statistical-type of lexicon. Both these methods use the same first order Markov model to characterize each word in the vocabulary, but they differ in terms of the ways of estimating the model parameters. These methods are described below.

3.2.1. *Method* 1 — Here, each word in the vocabulary is described by a first order Markov model. The number of states in this model is same as the number of ASWUs used in recognition system (i.e., there are N states in the model). Each state corresponds to one ASWU. The model is ergodic; i.e., there is no constraint on the transitions between the states — all the transitions are allowed. The model is

Table 5: Recognition results from the statistical-type of lexicon building method 1 with N=64 for V1 vocabulary and N=128 for V2 vocabulary.

Vocabulary	Recog. accuracy (in %) using training data from	
	50 repetitions	70 repetitions
V1	92.2	95.1
V2	93.8	94.8

Table 6: Recognition results from the statistical-type of lexicon building method 2 with N=64 for V1 vocabulary and N=128 for V2 vocabulary.

Vocabulary	Recog. accuracy (in %) using training data from	
	50 repetitions	70 repetitions
V1	92.9	94.9
V2	94.5	96.1

completely characterized in terms of transition probabilities between different states. This model is similar to the bigram model used to describe the language for the recognition of words [16] The transition probabilities between different states (or, ASWUs) are estimated for each word from the transcriptions of its different training utterances obtained by using the Viterbi algorithm.

This method is used to generate word lexicons for the V1 and V2 vocabularies and the recognition results are shown in Table 5. By comparing this table with Tables 1, 3 and 5, we can see that the statistical-type of word lexicon gives better recognition results than the deterministic-type of word lexicon.

3.2.2. Method 2 — Here, we use the same first order Markov model as used in method 1 for characterizing each word in the vocabulary. However, a different procedure is used to compute the transition probabilities between the states (or, ASWUs). Let A_{ij} be the transition probability from i-th ASWU to j-th ASWU. Also, let a_{ijk} be the transition probability from j-th state to k-th state in the 3-state HMM of the i-the ASWU (as described in Subsection 2.4). In method 1, a_{ijk} and A_{ij} are considered to be independent; i.e., $a_{i33} = 1$ and $\sum_{j=1}^{N} A_{ij} = 1$ for $i = 1, 2, \ldots, N$. In the present method, the transition probabilities a_{ijk} and A_{ij} are considered to be dependent as follows: $a_{i33} + \sum_{j=1}^{N} A_{ij} = 1$ for $i = 1, 2, \ldots, N$.

Word lexicon generated using this method is used to recognize isolated words from the V1 and V2 vocabularies. Results are shown in Table 6. By comparing this table with Table 5, we can see that this method results in better performance than the method 1. Also, comparison of this table with tables 1, 3 and 4 shows the superiority of the statistical-type of word lexicon over the deterministic-type of word lexicon.

4. DISCUSSION OF RESULTS

We have seen in the preceding section that we can improve the recognition performance of the ASWU-based speech recognition system by designing the word lexicon better. The best results are obtained by using the statistical-type of lexicon-building method which uses a first order ergodic Markov model to represent each word in the vocabulary.

As mentioned in Section 1, the objective in the design of a ASWU-based speech recognizer should be to approach the performance of the WWU-based speech recognizer. In order to see to what extent we have succeeded in this objective, we have implemented a WWU-based speech recognizer using the DTW approach. Recognition performance of this recognizer on V2 vocabulary is found to be 97.0% and 97.4% using training data from the 50 and 70 repetitions, respectively. The corresponding results from the ASWU-based speech recognizer studied in the present paper are 94.5% and 96.1%. From these results, we can see that we have been able to take the ASWU-based recognizer quite near to the WWU-based recognizer in terms of its recognition performance. But, there is still a scope for improvement. This improvement can come from better designs of word lexicon. Some directions for improving the word lexicon are listed below. 1) We have used here first order Markov model for characterizing a word in the lexicon. Use of higher order Markov models may improve the recognition performance (in the same fashion as the trigram model does over the bigram model [16]). 2) We have used here ergodic Markov model. It has been reported in the literature [15] that the left-to-right HMM leads to better recognition results than the ergodic HMM for isolated word recognition as it provides a more meaningful temporal constraint. We can use left-to-right constraint either on the Markov model or on the HMM to get better speech recognition performance. The left-to-right first order Markov model can be represented as a statistical pronunciation network where the nodes correspond to the ASWUs and the connection between the nodes are characterized by the transition probabilities. We are currently investigating these models and the results will be reported in future.

5. CONCLUSIONS

In this paper, an ASWU-based speech recognition system is used for the recognition of isolated words. Some methods are proposed for generating the deterministic and the statistical types of word lexicon. It is shown that the use of modified k-means algorithm on the likelihoods derived through the Viterbi algorithm provides the best deterministic-type of word lexicon. However, the ASWU-based speech recognizer leads to better performance with the statistical-type of word lexicon than with the deterministic-type. By improving the design of the word lexicon, the gap in the recognition performances of the WWU-based and the ASWU-based speech recognizers has been narrowed down considerably in the present paper. Further improvements are expected by designing the word lexicon better.

REFERENCES

[1] L.R. Rabiner, J.G. Wilpon and B.H. Juang, "A model-based connected digit recognition system using either hidden Markov models or templates", Computer, Speech and Language, Vol. 1, pp. 167-197. Dec. 1986.

[2] L.R. Rabiner, J.G. Wilpon and F.K. Soong, "High performance connected digit recognition using hidden Markov models", IEEE Trans. ASSP-37, pp. 1214-1225, Aug. 1989.

[3] R.P. Mikkilineni, J.G. Wilpon and L.R. Rabiner, "A procedure to generate training sequences for a connected word recognizer using the segmental k-means training algorithm", Proc. ICASSP, pp. 433-436, Apr. 1988.

[4] L.R. Bahl et al., "Recognition results with several experimental acoustic processors", Proc. ICASSP, pp. 249-251, Apr. 1979.

[5] K.K. Paliwal and A.M. Kulkarni, "Segmentation and labeling using vector quantization and its application in isolated word recognition", Journ. Acoust. Soc. India, Vol. 15, pp. 102-110, Jan. 1987.

[6] J.G. Wilpon, B.H. Juang and L.R. Rabiner, "An investigation on the use of acoustic sub-word units for automatic speech recognition", Proc. ICASSP, pp. 821-824, Apr. 1987.

[7] C.H. Lee, F.K. Soong and B.H. Juang, "A segment model based approach to speech recognition", Proc. ICASSP, pp. 501-504, Apr. 1988.

[8] L.R. Bahl et al., "Acoustic Markov models used in the Tangora speech recognition system", Proc. ICASSP, pp. 497-500, Apr. 1988.

9] V.R. Algazi and K.L. Brown, "Automatic speech recognition using acoustic sub-words and no time alignment", Proc. ICASSP, pp. 465-468, Apr. 1988.

[10] T. Svendsen, K.K. Paliwal, E. Harborg and P.O. Husoy, "An improved sub-word based speech recognizer", Proc. ICASSP, pp. 108-111, May 1989.

[11] C.H. Lee, B.H. Juang, F.K. Soong and L.R. Rabiner, "Word recognition using whole word and subword models", Proc. ICASSP, pp. 683-686, May 1989.

[12] T. Svendsen and F.K. Soong, "On the automatic segmentation of speech signals", Proc. ICASSP, pp. 77-80, Apr. 1987.

[13] Y. Linde, A. Buzo and R.M. Gray, "An algorithm for vector quantization", IEEE Trans. COM-28, pp. 84-95, Jan. 1980.

[14] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proc. IEEE, Vol. 77, pp. 257-286, Feb. 1989.

[15] J.G. Wilpon and L.R. Rabiner, "A modified k-means clustering algorithm for use in speaker independent isolated word recognition", IEEE Trans. ASSP-33, pp. 587-594, June 1985.

[16] L.R. Bahl, F. Jelinek and R.L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Trans. PAMI-5, pp. 179-190, Mar. 1983.