S12.22

WORD RECOGNITION USING WHOLE WORD AND SUBWORD MODELS

Chin-Hui Lee, Biing-Hwang Juang, Frank K. Soong and L. R. Rabiner

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

ABSTRACT One of the key issues in designing a speech recognition system is the selection of the fundamental unit for recognition. The choice of the fundamental unit for a recognition task generally depends on the size of the vocabulary to be recognized and the availability of sufficient training data for creating effective reference models. In this paper, we address the problem of how to select and construct a set of fundamental unit statistical models suitable for speech recognition. We discuss a unified framework which can be used to accomplish the goal of creating effective basic models of speech. We also compare the use of three types of fundamental units, namely whole word, phoneme-like and acoustic segment units in an 1109-word vocabulary speech recognition task. We point out the relative advantages of each type of speech unit based on the results of a series of recognition experiments.

1. INTRODUCTION

Speech recognition deals with the problem of mapping acoustic signals to (a sequence of) linguistic codes. The difficulty of speech recognition lies in the fact that such a mapping is generally ill-defined, especially in the case of continuous speech. When examined at various levels, such as the sentence, phrase, word, morpheme, and phoneme, the mapping, at each level, involves a different amount of acoustic variability and ambiguity which affects both the efficiency and effectiveness of any automatic speech recognition mechanism which attempts to perform the mapping.

It is generally acknowledged that linguistic ambiguities in specifying words are minimal. Therefore many traditional recognizers are built upon the choice of whole word units (WWU) implemented either as statistical models or templates. Although remarkable performance of word-based recognition systems has been demonstrated in small vocabulary tasks such as spoken digit recognition, extrapolating the requisite techniques to large vocabulary applications is not straightforward. There are several reasons for this. The main one is the problem associated with training. In order to adequately train a set of whole word models, each word in the vocabulary must appear several times in each context of interest. For large vocabularies, this implies a prohibitively large training set. In some task configurations, the recognition vocabulary consists of words which had not appeared in the training phase. As a result, some form of word model composition technique is required to generate models for those words not seen sufficiently of them during training. As a consequence of the above problems, the need for subword models thus becomes obvious.

A natural alternative to a word-based approach is to use *linguistically defined* subword unit such as phoneme, diphone, demi-syllable or syllable. Such an approach to speech recognition relies on the assumption that a word model can be constructed based on existing linguistic knowledge. Typically, a *lexicon* in the form of a dictionary in terms of the chosen subword units is used to describe the words of the vocabulary. Each word in the vocabulary is usually represented by a concatenation of subword models as specified by the lexical entries associated with that word. While some of the training problems in large vocabulary word-based systems are eliminated, one major problem remaining is the extraction of these linguistically defined units from the acoustic data. This problem is usually handled by manually segmenting a smaller database (e.g. the TIMIT database) according to a linguistic specification of the speech to initialize the set of subword models. Some form of bootstrapping procedure is then used to automatically segment a larger set of training data in order to have sufficient training for the subword models. Since linguistic boundaries are often ill-defined in the acoustic signal, the segmentation and the resulting models are not guaranteed to be consistent with their original linguistic definitions.

Another alternative is to use *phoneme-like units* (PLU's) or *diphone-like units* (DLU's). A phoneme-like (or diphone-like) unit is a unit similar to a phoneme (or a diphone) in the sense that their linguistic roles as a constituent in a word representation are identical. Thus, the same lexicon is applicable in both cases. However, specification of the locations of PLU's (or DLU's), as will be explained in greater details in the next section, capitalizes mainly on acoustic similarities and can be accomplished via automatic procedures. Therefore, these units are defined both acoustically (through the use of an acoustic similarity measure) and linguistically (due to the prescribed lexical constraints).

A third alternative is to rely entirely on a measurable acoustic similarity in defining the units. This results in what is called *acoustic segment units* (ASU's) as reported at ICASSP-88 [1]. In the acoustic segment formulation, we attempt to find a set of acoustic segment models that span the acoustic signal space based on the principle of minimum average distortion. We then use the acoustic segment models as the fundamental units and construct all the models for recognition from the acoustic definition of their signal realizations. As will be discussed, one of the unique requirements of this strictly acoustic approach is that a word lexicon be built based upon the acoustic segment units. In contrast to the phonetically-defined lexicon, such a lexicon is called an *acoustic lexicon*.

In this paper, we address some of the key issues related to the selection of fundamental units for large vocabulary speech recognition. Specifically, we discuss the use of three types of fundamental units, namely whole word units, phoneme-like units and acoustic subword segment units, for isolated and connected word recognition. In the following sections, we show that a unified framework based upon the segmental *k*-means training procedure [2] can be used to create effective basic models of speech. We describe the two major modeling steps required for word recognition using a statistical approach, namely: (1) unit modeling, in which we use statistical modeling techniques to characterize the acoustic properties of the chosen units; and (2) word model composition, in which we construct word models from the unit models (which may be whole word units). We then report recognition results using a speaker-trained, isolated word, speech recognition task with a vocabulary of 1109 basic English words [3]. Based on a database of three male speakers, the average word recognition accuracy using the different speech units ranged from 82% to 95%. A preliminary comparison was made to determine the advantages and disadvantages of the various choices of basic speech units.

2. WORD MODEL COMPOSITION

If units smaller than a word are used for word recognition, then word model composition is required. As mentioned above, the two choices of subword units we consider in this paper are the phoneme-like units and the acoustic segment units respectively. The basic difference between the two, aside from the unit modeling details to be discussed in the next section, lies in the fact that the former uses an *a priori*, linguistically based, word representation in terms of phoneme strings (in a lexicon) and the latter uses an *a posteriori*, acoustically based, word representation constructed from the acoustic segment units, extracted directly from the acoustic signal.

In the case of phoneme-like units, an initial set of phonemes are defined a priori. The definition is based on linguistic notions rather than acoustic ones. Each word in the vocabulary is assigned a pronunciation string or a set of pronunciation strings which can be considered as lexical entries in a phonetically-based lexicon (PL). Each entry in the lexicon provides a *baseform* pronunciation for a speaker to follow when a word is spoken. The lexicon thus implies "tying" constraints on the PLU's through the linguistic definitions of the phonemes for all acoustic data in the training set. The advantage of using such a lexicon is that word models not adequately observed in the training data can still be constructed based on the prescribed baseform. However, this lexicon does not take into account speakers' acoustic variabilities, inconsistencies

CH2673-2/89/0000-0683 $1.00 © 1989 IEEE

in different contexts, and allophonic variations of sounds. As a result, it is likely that signal modeling of the phoneme-like units is not sufficiently accurate, and the resulting ambiguity will cause confusion in word models such that a word model needs not recognize the word properly in all contexts. One way to improve the acoustic discrimination among words is to introduce context dependency (e.g., [4]) into the definition of phonemes so that the phoneme-like models provide adequate acoustic resolution for word composition. Another way is to incorporate data-dependent lexical entries into the lexicon as explained below.

Given a set of phoneme-like unit models for a given speaker, an a training set in which each of the vocabulary words appears at least once, we can also construct a speaker-dependent lexicon by performing optimal decoding on each of the training tokens for each word from the same speaker. The decoded phoneme symbol sequences can serve as additional lexical entries for that word. For simplicity, we call such a lexicon an acoustically-derived, phonetically-based lexicon (APL). The APL can be used to perform word model composition for recognition. The APL can also be combined with the speaker-independent PL to form a lexicon which is both acoustically and linguistically defined.

When word models are derived strictly from acoustic signal realizations, an acoustic lexicon (AL) is required to describe each word based on the set of acoustic segment units. In contrast to the conventional linguistically-defined lexicon, every lexical entry in the acoustic lexicon is represented by sequences of acoustic segment symbols, which can be generated by performing optimal decoding on the set of the training data using the acoustic segment models. Different acoustic manifestations of the same word may have different acoustic lexical representations although they are related to each other through a prescribed acoustic similarity measure. Multiple AL representations of a word can be derived from multiple training tokens of the word and from multiple candidate sequences decoded from the same training token.

3. MODELING OF THE BASIC UNITS

For a given recognition task and a set of training data, our objective is to choose a set of fundamental units so that the training data are utilized efficiently and effectively, leading to a speech recognizer with good performance. If we assume that the basic speech units have already been defined and a set of constraints on the interactions of the units are given, the segmental *k*-means training procedure can be used to accomplish the tasks of basic unit modeling and word model composition simultaneously. We shall first explain how the basic units are defined initially in a statistical modeling framework.

3.1 Defining the Basic Units

One of the key considerations in modeling the basic units is the availability (and correctness) of a priori knowledge about the given training data. We enumerate three possible scenarios: (1) The set of units is externally defined and each token in the training data is associated with a single, specific label according to a set of units. (2) The set of units is again externally defined, but each token in the training set may encompass and thus be labeled with a number of units, with the boundaries between units unspecified. (3) Neither the units nor the training token compositions are defined a priori; only the set of training data itself is given. These scenarios cover essentially all the possibilities we encounter in the design of a recognition system.

Consider the first case in which each token in the training set has been assigned a unique unit label. Typical instances of this type of training set include isolated word data or hand-segmented signals with specific transcriptions of the units. Since the training data are presented as a sequence of discrete, segmented units, be they words or phonemes, no further segmentation is required and we can assume all the training data associated with a given unit label are generated by a single but unknown source. Statistical modeling techniques such as the maximum likelihood approach can then be used directly to obtain the unit models.

In the second case, the units are pre-defined but the training data are only block-labeled. That is, the sequence (string) of units for each training token is given, but no label boundaries are provided. As an example, in the connected digit case, we may define the units to be whole word digits. The training set, however, often contains information only about the digit sequence without detailed segmentations to specify the spoken digit boundaries. Similarly, for phoneme-like unit representation, we define the units to be a set of phonemes and each training utterance is given a set block labels through the phonetic lexicon, without detailed specification of the phoneme boundaries. As will be explained, the segmental *k*-means training procedure provides a solution to the model parameter estimation problem as required for the unit modeling.

In the third case, the units are not pre-defined. The first step in this modeling approach is to find a reasonable set of acoustic units that efficiently characterize the training data. Such an initial set, which replaces the standard linguistically defined set, can be found by applying a minimum average distortion (dissimilarity) criterion in an iterative procedure similar to the generalized Lloyd method widely used in vector quantizer designs. Again, using the segmental *k*-means algorithm, modeling of the defined acoustic units can be accomplished automatically. Using the resulting set of of acoustic units we can construct a word lexicon or other linguistic components techniques explained in the last section.

3.2 Model Parameter Estimation

Throughout this paper, hidden Markov models are used to characterize all three types of fundamental speech units. Once the unit definitions are made, HMM techniques can be applied directly to model the units. As discussed above, for all three types of units, the segmental *k*-means training procedure provides a unified framework for both unit modeling and word model composition. The segmentation part of the algorithm provides a set of labeled segments. The labeled unit sequences associated with a word are used for word model composition when necessary. Finally, the *k*-means part of the algorithm is used to estimate the parameters of the model for each unit. We now briefly describe the procedure for each type of unit.

3.2.1 Whole Word Units

In the case of isolated word training data, no detailed unit segmentation is required when whole word units are used. The well known Baum-Welch algorithm or the segmental *k*-means algorithm can be used to obtain the maximum likelihood estimate of the model parameters. The algorithm iterates until a fixed point solution is reached. Typically, only 5 to 8 iterations are needed to get a good word model.

3.2.2 Subword Units

As described above, two types of subword units, namely phoneme-like units and acoustic segment units, are studied in this paper. The phoneme-like units are defined through linguistic definitions, while the acoustic segment units are defined strictly from acoustic signal realizations.

In the case of phoneme-like unit modeling, we start with a set of phoneme symbols and a lexicon based on this set of units. To train the unit models, we first perform uniform segmentation on each of the training tokens according to the number of phoneme symbols assigned to that word token. We then collect all the speech segments associated with the same phoneme-like symbol and create the phoneme-like unit models using the HMM training procedure described above. In an iterative manner, the unit models are used to improve the segmentation, and the segmented data are used to improved the unit models. This procedure is similar to the one used in model training for connected digit recognition [5], except the digit sequence is now replaced by a PLU symbol sequence as specified by the lexicon. Since phoneme boundaries within a word are usually more ambiguous than word boundaries in a connected word sequence, there is no guarantee that the resulting phoneme-like segments will agree well with their corresponding phonetic definitions.

In the case of modeling acoustic segment units, the model initialization and training procedure is summarized in the block diagram shown in Figure 1. In order to obtain a reasonable acoustic segment model, the speech training material is first segmented into a set of acoustic segments using a maximum likelihood automatic segmentation algorithm. All the segments resulting from initial segmentation are then grouped into initial segment clusters representing individual sound classes. To model each sound cluster, all the acoustic segments in a single cluster are considered to have been generated from a single source. Once the HMM structure (topology) is specified, the standard HMM technique is then applied to estimate the parameters of each of the individual segment models corresponding to each sound class. Again, following the segmental *k*-means procedure, the unit models and the segmentation results (used to construct an acoustic lexicon) are alternatively and iteratively improved until some convergence criterion is met.

4. RECOGNITION EXPERIMENTS

We now describe the experimental setup and present some recognition results using the three types of fundamental units described

above. The task we selected to evaluate and compare the use of the three fundamental units is the recognition of a vocabulary of 1109 words from the basic English vocabulary of Ogden [3]. For this vocabulary, there are 605 monosyllabic words; some of them are homophones, and some of them form confusable word pairs. The database consists of three male talkers, each uttering the entire 1109 words four times, in isolated word mode. The input speech signal was recorded off a standard dial-up telephone line, bandpass-filtered between 100 Hz and 3200 Hz, and digitized at a sampling rate of 6.67 kHz. The speech data had been located by an automatic endpoint detector. It was then preemphasized and blocked into frames of 45 ms with a 15 ms shift. A Hamming window was then applied to each frame of the data; and for each utterance, a sequence of vectors of nine autocorrelations were produced for model training and algorithm testing.

A block diagram of the isolated word recognizer is shown in Figure 2. The feature extractor computes the feature vector to be used by the Viterbi decoder. The word model composition module takes as input the unit models and the associated lexicon, and produces word models for every word in the vocabulary. For recognition using whole word models, no word composition is required. For recognition using PLU models, the lexicon is obtained from a standard pronunciation dictionary with a single pronunciation for each word. For recognition using acoustic segment models, we used the acoustic lexicon generated in the training phase. The Viterbi decoder uses dynamic programming search techniques to find the most likely segmentation of the input utterance with respect to the models of the vocabulary words and computes the likelihood of the input feature vector sequence, given the set of the models and the lexicon. If more than one lexical entry per word was used, the maximum of the likelihoods is selected as the score for that word. The score sorting module chooses the top K candidates that produce the top K scores among all possible words in the vocabulary.

The task performed was a speaker-trained, isolated word recognition test. For one session of testing data, there are three other sessions of training data available for model estimation and lexicon generation. Therefore, in testing, up to three sets of acoustically-derived lexical entries (APL or AL) for each word can be used to construct the word model for that word. The experimental setup using the three types of units are now described in more detail.

4.1 Whole Word Units

For the 1109-word database, we used an 8-state HMM to model each word. Up to 5 Gaussian mixture densities, each with a diagonal covariance matrix, were used to characterize the state observation density. The feature used was a vector of 24 components, consisting of 12 liftered cepstral coefficients and 12 corresponding time derivatives. Since at most 3 tokens for each word could be used to train the whole word model, there were not enough samples in each state to properly estimate the variance. To deal with this problem, we used a variance clipping procedure described in [6], and our results indicated that reasonable, speaker trained, whole word model could be obtained even with only one training token per word.

4.2 Phoneme-like Units

For recognition using phoneme-like units, we need to define the set of phone symbols and a lexicon based on this set of units. For convenience, a set of symbols used in the Bell Laboratories Text-to-Speech Synthesis System was chosen for our experiments. The set consists of 43 phoneme symbols including a silence symbol. The lexicon was obtained by applying the letter-to-sound rules to all the 1109 words in the vocabulary. There were some inconsistencies between the acoustic data and the lexical specifications for some of the vocabulary words. For example, the word "A" was specified as "EY" in the lexicon; however the speakers were instructed to pronounce the word "A" as the schwa "UH" when the data were recorded. No attempt was made to modify the dictionary.

The PLU model used in our experiments is a 2-state HMM for each unit. The feature used was the same as the one in whole word unit modeling. Up to 5 mixtures of Gaussian densities, each with a diagonal covariance matrix, were used to characterize each state observation density. We also used the same variance clipping strategy, because some of the PLU's occur infrequently in the lexicon, and therefore very little acoustic data were used to create models for these units.

Given the set of PLU models for each speaker, we also constructed a speaker-dependent lexicon by performing optimal decoding on each of the training tokens for each word from the same speaker. We compared the derived APL with the prescribed PL for one set of the training data from one speaker. Only 1% of the lexical entries from both lexicons agreed with each other exactly. The results indicated that better acoustic models and lexical models are needed in order to properly characterize PLU-based word models. We feel that better acoustic models can be obtained by exploring context-dependent PLU's. In this study, we try to obtain better lexical models by combining both APL and PL in lexical representations.

4.3 Acoustic Segment Units

For recognition using acoustic segment units, we used a 1-state model to characterize each of the segment units. The numbers of units ranged from 32 to 256. The feature vector consisted of 16 liftered cepstral coefficients. The state observation density was a multivariate Gaussian density with a diagonal covariance matrix; no mixture densities were used. The acoustic lexicon was constructed by performing optimal decoding on all training tokens using the segment models. Up to three lexical entries were obtained for each word in the vocabulary.

4.4 Recognition Results and Discussion

The basic recognition setup was to perform a recognition test on the three sessions (except session 1) of test data from the three male talkers. Since session 1 was always used to estimate the PLU models and the acoustic segment models, no test on session 1 was performed. As a result, a total of nine sessions were tested. The setup was such that for testing on one session, the acoustic lexicons obtained from the other three sessions were used to perform word model composition and Viterbi decoding.

The average word accuracy rates (in per cent) over all the nine testing sessions, are listed in Table I. There were some errors caused by homophones, but no attempt was made to correct the homophone errors. The actual performance for the top candidates was about 0.5% higher than the recognition rates listed if homophone errors were corrected. In the row labeled "1109WWU", we show the recognition results obtained when one whole word model per word was used. The recognizer also used a strategy similar to the system in [5] which includes energy, state duration and word duration information to help eliminate unlikely word candidates. An average word accuracy of 95% was obtained for the top candidate, and an almost perfect performance was achieved when the top five candidates were included for scoring. This is the best performance reported for this database. Almost all the errors were caused by homophones and confusable monosyllabic word pairs.

Model	Top	2	3	4	5
1109WWU	95.1	98.4	99.2	99.4	99.6
256ASU	89.3	95.7	97.3	98.0	98.3
128ASU	88.3	95.4	97.3	98.0	98.3
64ASU	85.5	93.8	96.4	97.5	98.1
43PLU	81.6	89.8	93.1	94.7	95.8
32ASU	79.2	90.7	94.3	96.0	96.9

Table I. Average recognition rates using whole word and subword units

The row labeled "43PLU" corresponds to results obtained when 43 context-independent PLU's and their associated lexicon (PL) were used. Energy, and duration information were also included in recognition scoring. The rows labeled "ASU" correspond to the results obtained when using acoustic segments models and up to three acoustic lexical entries per word for recognition. The number of segment models were 32, 64, 128 and 256 respectively, for the four ASU experiments evaluated. The best performance, among the four choices, was obtained when 256 segment models were used. However, the performance gain decreases as the number of segment models increases. There was only a slight performance improvement going from 128 to 256 ASU models.

For the case using 43 PLU's, there were a total of 86 distinct acoustic states. However the PLU performance was only slightly better than the case when using 32 acoustic states (32 ASU), and was significantly worse than the case using 64 acoustic states (64 ASU). We feel that PLU performance can be greatly improved if context-dependency is properly incorporated. We also feel the performance using acoustic segment models can be improved by incorporating additional features, such as time derivatives of the cepstral coefficients, energy, word duration and state duration, in recognition. All the recognition results listed in Table I are summarized in Figure 3, in which we plot the average

recognition rates versus the number of candidates included in recognition scoring. The figure shows that in all six cases tested, at least 90% word accuracy was achieved using the top 2 candidates, and better than 95% performance was observed when the top five candidates were included.

To examine the effect of the amount of training data on recognition performance, we used only the training data from session 1 of each speaker to create the models and lexicon required for testing. The recognition test was performed on the data in session 4 from the same speaker. All the results for each of the three speakers and the averages over all 3 speakers are listed in Table II. We found that even when trained with only one token per word, the whole word model still gives a reasonable performance. The result obtained from using PLU models stayed about the same level as those shown in Table I, when a PL was used for recognition. To see the effect of using the APL, we also constructed an APL from the available training data. Using the APL alone, the average performance was much worse. However for speaker 2, there was a performance improvement, which was due the the fact that the data from speaker 2 were acoustically more consistent than the data from the other two speakers. When the APL was combined with a PL in testing, there was a significant improvement in average performance. This gain was mainly due to the addition of speaker-dependent lexical representations. For example, the word "A" which was properly recognized in all cases tested except in the case of PLU recognition based on only PL. This error was corrected, when an APL was incorporated into the lexical representations. For testing the acoustic segment models using only one lexical entry per word, we found the performance similar to the one using PLU and an APL. For comparison purposes, we also list the results obtained with an APL and an AL generated from all three sets of training tokens. It is interesting to note that there was virtually no change in performance in the case of using PLU models. However, there was a big performance gain in the case of using acoustic segment model. The results showed that proper lexical modeling is crucial when purely acoustic units are used for word recognition.

	Spkr1	Spkr2	Spkr3	Average
1109WWU	87.7	89.5	86.6	87.9
43PLU(PL)	81.6	84.8	76.7	81.0
43PLU(1APL)	74.2	85.8	72.0	77.3
43PLU(PL+1APL)	83.4	88.6	77.8	83.2
256ASU(1AL)	74.3	84.5	74.1	77.6
43PLU(PL+3APL)	84.2	88.6	77.6	83.4
256ASU(3AL)	86.9	92.4	85.1	88.1

Table II. Recognition performance comparison using various lexicons

5. SUMMARY

In this paper, we have discussed the use of three types of units, namely whole word, phoneme-like and acoustic segment units. A unified framework based on the segmental *k*-means training procedure was used in all three cases to accomplish unit modeling and word model composition simultaneously. In terms of modeling effectiveness, we found that using whole word units maintains the integrity of a word, and by far achieves the best performance among all the three types of units tested. For words that are acoustically more variable, such as short function words which are likely to appear more often in training data, whole word models should be used. However, in term of modeling efficiency, there is no model sharing in either the acoustic or lexical domains. For recognition tasks requiring more training data to properly model acoustic variabilities, such as speaker-independent, large vocabulary recognition, whole word models do not use all the training data efficiently. The need for subword models thus becomes essential.

Two types of subword units, phoneme-like and acoustic segment units, were studied. For modeling using phoneme-like units, we found that the baseform dictionary provides an efficient way to construct word models not seen during training. However, issues related to model ambiguity, context dependency and lexicon consistency need to be studied further. For modeling using acoustic segment units, acoustic consistency is maintained throughout both unit modeling and acoustic lexicon generation. When compared with units derived from phonetic definition such as the PLU's, this segment model approach is acoustically less ambiguous. However, lexical modeling also requires careful investigation so that a more effective way can be used to incorporate different amounts of training data in constructing the acoustic lexicon.

In summary, we have addressed some of the key issues related to the selection of fundamental units for speech recognition. From the results of our experiments, we feel that a hybrid approach based on a combination of both whole word and subword models should be used in order to utilize the training data more efficiently and more effectively.

REFERENCES

[1] C.-H. Lee, F. K. Soong and B.-H. Juang, "A Segment Model Based Approach to Speech Recognition," *Proc. ICASSP-88*, pp. 501-504, New York, April 1988.

[2] L. R. Rabiner, J. G. Wilpon and B.-H. Juang, "A Segmental *k*-means Training Procedure for Connected Word Recognition," *AT&T Tech. Journal*, Vol. 65, No. 3, pp. 21-31, May/June 1986.

[3] C. K. Ogden, *Basic English: International Second Language*, Harcourt, Brace and World Inc., New York, 1968.

[4] K.-F. Lee and H.-W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," *Proc. ICASSP-88*, pp. 123-126, New York, April 1988.

[5] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," *Proc. ICASSP-88*, pp. 119-122, New York, April 1988.

[6] L. R. Rabiner, C.-H. Lee, B.-H. Juang, D. B. Roe and J. G. Wilpon, "Improved Training Procedures for Hidden Markov Models," J. Acous. Soc. Am. Suppl. 1, Vol. 84, S61, Fall 1988.

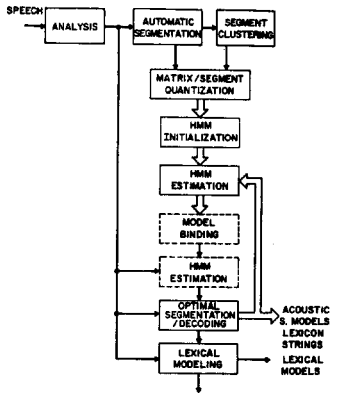

Figure 1. A block diagram of the acoustic segment model approach

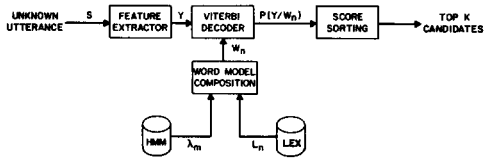

Figure 2. A block diagram of the HMM-based isolated word recognizer

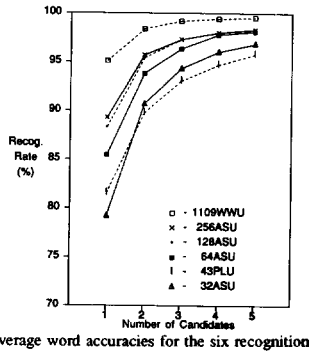

Figure 3. Average word accuracies for the six recognition configurations