An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition

J. G. Wilpon
B. H. Juang
L. R. Rabiner

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

Abstract. An approach to automatic speech recognition is described which attempts to link together ideas from pattern recognition such as dynamic time warping and hidden Markov modeling, with ideas from linguistically motivated approaches. In this approach, the basic sub-word units are defined acoustically, but not necessarily phonetically. An algorithm was developed which automatically decomposed speech into multiple sub-word segments, based solely upon strict acoustic criteria, without any reference to linguistic content. By repeating this procedure on a large corpus of speech data we obtained an extensive pool of unlabeled sub-word speech segments. Then using well defined clustering techniques, a small set of representative acoustic sub-word units (e.g. an inventory of units) was created. This process is fast, easy to use, and required no human intervention.

The interpretation of these sub-word units, in a linguistic sense, in the context of word decoding is an important issue which must be addressed for them to be useful in a large vocabulary system. We have not yet addressed this issue; instead a couple of simple experiments were performed to determine if these acoustic sub-word units had any potential value for speech recognition. For these experiments we used a connected digits database from a single female talker. A 25 sub-word unit codebook of acoustic segments was created from about 1600 segments drawn from 100 connected digit strings. A simple isolated digit recognition system, designed using the statistics of the codewords in the acoustic sub-word unit codebook had a recognition accuracy of 100%. In another experiment a connected digit recognition system was created with representative digit templates created by concatenating the sub-word units in an appropriate manner. The system had a string recognition accuracy of 96%.

I. Introduction

The process of automatic speech recognition (ASR) can be represented as a sequence of quasi-independent procedures as shown in Figure 1 [1-7]. Following a standard feature analysis, in which a spectral characterization of the speech is computed, the first stage of recognition is matching (i.e. spotting or detecting) a set of basic speech units. This matching process, which uses an inventory of reference units that has been created during a training phase is often some type of dynamic programming procedure.

The next stage in the recognition process is lexical decoding. This is the process of converting the stream of recognized units to a stream of word units. The decoder requires a word dictionary (in which words are orthographically represented in terms of the chosen speech recognition units) to generate a list of possible word candidates. There are at least three distinct methods for lexical decoding namely:

1. using statistical models [8];
2. using a deterministic state diagram representation of words [9];
3. using a lexical access retrieval method like redundant hash addressing (RHA) [10].

Each of these methods has a completely different embodiment of the word dictionary.

The next stage in the recognition process is syntactic analysis in which the stream of word units is checked for grammatical correctness. All word sequences which are not syntactically valid are eliminated. Again there are at least three distinct ways of implementing the syntax analysis, including:

1. representing the grammar by a statistical model (e.g. triagram word probabilities), and associating probabilities with each possible sequence of words [8]
2. representing the grammar by a deterministic state diagram, thereby enumerating all possible word sequences [8,11]
3. representing the grammar in a formal manner (e.g. parsing the word string and applying some type of language grammar [4,12].

Almost invariably, the stages of syntax analysis and lexical decoding are combined so that only syntactically valid word sequences are generated. In this manner, the computation of the combined stages is usually minimized.

The final stage in the recognition process is semantic analysis in which a model of the task is used to eliminate sentences which are syntactically correct, but semantically meaningless in the context of the current task. Although semantic analysis will ultimately be an important step in the speech recognition process, its immediate applicability has been of limited utility.

Although a great deal of research has gone into trying to develop a continuous speech recognition system with all the processing shown in Fig. 1, most practical ASR systems have been based on whole word speech units [1-6]. The advantage of using whole word speech units is that the need for a word dictionary and lexical decoding is eliminated, thereby greatly simplifying the entire recognition process. The disadvantage of using whole word speech units is the difficulty in obtaining good whole word reference models from a limited amount of speech training material. Hence the major successes of whole word speech recognition systems have been when the required vocabulary is of relatively small size (e.g. digits [1], airline terms [1,12], etc). When one is interested in large vocabulary speech recognition systems, a full implementation of the recognition system of Fig. 1, based on acoustic sub-word units, is required.

The major problem with using sub-word speech units is that robust and reliable algorithms for automatically determining the *presence* and/or *identity* of such units do not exist. This problem exists because the sub-word units that have been investigated, e.g. phonemes, diphones, demisyllables, etc., have been defined based on a linguistic description of language and not on what can actually be reliably detected acoustically. Hence there is a mismatch between sub-word speech units defined linguistically, and those defined acoustically (according to some well defined criterion). The well defined linguistic sub-word units make lexical decoding an easy task, since a standard dictionary pronunciation will generally provide a simple and straightforward mapping between the chosen linguistic units and the word orthography.

CH2396-0/87/0000-0821 $1.00

However it would be difficult, if not impossible, to automatically build such linguistically defined sub-word unit models from any realistic training set. Alternatively, sub-word units which are derived acoustically and to which no simple linguistic interpretation can be attached are convenient for training purposes, but lead to great difficulties in lexical decoding since no simple and/or straightforward mapping to words would be possible.

In this paper we describe an approach to automatic speech recognition in which the basic sub-word units are defined *acoustically*, without any reference to linguistic content. A procedure has been developed which automatically breaks a speech utterance into segments based upon well defined acoustic criteria. Using well defined clustering procedures, a small set of *acoustic* sub-word units is created from the segments detected over a wide range of training speech. A recognition system which uses these sub-word speech units has been developed for a small vocabulary. We bypass the difficulties of lexical decoding by using one of two simple procedures for describing words in terms of the sub-word units.

In Section 2, we describe how the acoustic speech segments are created. In Section 3 we describe the procedure for clustering segments to form an acoustic unit reference set or codebook. Finally, in Section 4 we describe a series of recognition experiments performed to assess the recognition capability of the acoustic sub-word unit set.

II. Automatic Procedure for Creating Acoustic Segments

There are two ways that a set of sub-word units can be created. The first is by hand labelling speech based on a linguistic interpretation of what was spoken. This is the approach taken in most acoustic-phonetic recognition systems. In fact, in Ref. [3] it was shown that by using acoustic-phonetic information, in addition to standard LPC spectral features, in the recognition process, a significant increase in the accuracy on a connected digits task (using the Texas Instruments connected digit database [14] resulted. However, to obtain the reference acoustic-phonetic models took a great deal of time for hand segmentation of a connected digit database. For vocabularies larger than 10 words, such a laborious and tedious procedure usually is not practical.

The alternative to hand segmentation of speech into acoustic phonetic units is to devise an automatic procedure which can provide a consistent identification of sub-word units in speech signals. We now describe one possible implementation of such a procedure.

2.1 Automatic Segmentation Procedure for Producing Acoustic Sub-Word Units

The automatic segmentation procedure described here is based upon measurements of spectral variation over time. From the spectral variation contour we define acoustic events and perform segmentation on the speech signal. We first describe the calculation of the spectral variation contour.

Given a speech signal, we denote the spectral representation for the ℓ^{th} frame as x_ℓ, where x_ℓ can be a cepstral vector, an LPC vector, or any other appropriate spectral representation. We assume that an appropriate distortion measure between pairs of spectral representations exists, and we denote it as $d(x_i, x_j)$. Typical distortion measures include the Itakura-Saito distance for LPC vectors and the Euclidean distance for cepstral vectors or other direct spectral representations.

The stationarity of a signal at frame i is defined by calculating the spectral variation of the signal over the $2L$ frame interval $[i-L+1, i+L]$. To do this calculation, we first calculate the $(2L-1)$-dimensional distortion vector

$$D^t(i) = [d_{i-L+1}, \ldots, d_i, \ldots, d_{i+L-1}] \tag{1}$$

where

$$d_k = d\left[\frac{1}{L+k-i}\sum_{j=i-L+1}^{k} x_j\ ,\ \frac{1}{L-k+i}\sum_{j=k+1}^{i+L-1} x_j\right], \tag{2}$$

$$k = i-L+1, \ldots, i+L-1$$

If the distortion measure is Euclidean, the averaging in Eq. (2) is done directly on spectral vectors; if the distortion measure is of the Itakura-Saito type, the averaging is done on autocorrelation vectors. The distortion vector is further smoothed by averaging the current distortion vector and a shifted version of the previous smoothed distortion vector:

$$D_s(i) = D_s^t(i-1)H_1 + D^t(i)H_2 \tag{3}$$

where the smoothing matrices are

$$H_1 = \begin{bmatrix} 0 & & & & \\ 0.5 & 0 & & O & \\ 0 & 0.5 & 0 & & \\ & & \ddots & & \\ & & & \ddots & \\ & O & & \ddots & \\ & & & 0.5 & 0 \end{bmatrix}, \text{ and } H_2 = \begin{bmatrix} 0.5 & & & \\ & 0.5 & & O \\ & & \ddots & \\ & & & \ddots \\ & & & \\ & & & 1.0 \end{bmatrix}.$$

Finally, all the elements in $D_s^t(i) = [d_{i-L+1}, \ldots, d_i, \ldots, d_{i+L-1}]$ are averaged to produce a distorted variation, $v(i)$, at time i:

$$v(i) = \frac{1}{2L-1}\sum_{j=i-L+1}^{i+L-1} d_j \tag{4}$$

We call the sequence $\{v(i)\}$ the spectral variation contour. Figure 2 shows an example of the spectral variation contour for a representative speech signal. Figure 2a shows the acoustic waveform for a 1 sec. long utterance. Figure 2b shows the calculated spectral variation contour. It can be seen that regions with high $v(i)$ values are usually associated with transient sounds, while regions with low $v(i)$ values are usually associated with steady state sounds.

2.2 Defining Acoustic Events According to the Spectral Variation Contour

The spectral variation values are instantaneous indications of the degree of variation of the associated speech signals. Unlike the maximum likelihood decoding associated with hidden Markov models, segmentation based upon the spectral variation contour can be done instantaneously without having to wait for the entire observation of the signal. In addition, the segmentation algorithm that we propose is free from the constraint of a prescribed target number of segments; it segments the signal according to a well defined definition of local acoustic events and will give the number of segments that are found.

Our definition of acoustic events is centered around the notion of steady state sounds and transient sounds. By a steady state sound we mean a sequence of spectra that are similar to each other; a steady state sound is considered a stationary random process. When the signal displays rapid characteristic changes, mostly reflected by large spectral variations, we say it is a transient. These definitions, although only qualitative, can be consistently and quantitatively observed from the spectral variation contour.

Since a steady state sound must undergo a transient phase before moving to another steady state, the first step in extracting (or segmenting) the acoustic units is to find the anchor points that correspond to the peaks of spectral changes. These anchor points of the transient segments are obtained by using a "matched filter" and a smoother. An example of the detected transient anchor points is given in Figure 2c for the example of Figure 2b. As can be seen, these anchor points accurately mark the instances where locally (along the time axis) maximal spectral changes occur. It is important to re-emphasize that these anchor points are obtained virtually instantaneously without any global constraint on segment

locations or counts.

Once the points of maximum spectral variation (i.e. the anchor points) are found, transient segments can be located easily by examining the spectral variation contour near these anchor points. A transient segment is extracted in the vicinity of an anchor point, by starting from the frame where the spectral variation exceeds some threshold (transient onset) and ending at the frame where the spectral variation drops below another threshold (which may be different from the onset threshold).

The detection of steady state segments is easier than the detection of transient segments. We define a steady state segment as a consecutive sequence of frames of speech with corresponding spectral variation values below a certain threshold. When the threshold is the same as the transient onset/offset threshold, these segments are simply those frames of speech between adjacent transient segments.

The above procedure for detection of transient and steady state segments includes a procedure for detection of silence. Acoustic segments are defined only when they are not silence.

2.3 Segment Detection Parameters

The above description of our automatic segmentation mechanism is meant to be general. In actual implementations, proper choices of analysis parameters are important.

For the implementation used in the remainder of this paper, we first analyze the speech with an 8th order LPC analysis. Each analysis frame is 15 msec long and consecutive frames correspond to a shift of 5 msec in time.

We use the COSH measure [15] as the distortion measure for the calculation of the spectral variation contours. The transient onset/offset thresholds as well as the steady state thresholds are all set equal to 0.275.

III. Experimental Procedure for Generating Sub-Word Units

A series of experiments was performed using speech from one female talker who was an inexperienced user of ASR systems. Four hundred fluently spoken connected digit strings of varying length (from 1 to 7 digits per string) were recorded over local dialed-up telephone lines. The speech was recorded at a 6.67 kHz rate and was bandpass filtered from 100-3200 Hz. The data was divided into disjoint training and testing sets.

One hundred of the connected digit strings were used in the segment generation phase. For each digit string, the beginning and ending points were determined using an automatic endpoint detector so as to eliminate the silence regions before and after the spoken digit string. The segmentation procedure was then run on the connected digit strings. The output of this process was approximately 1600 unlabeled speech segments (segments which had a length less than 20 msec were eliminated). Using a modified k-means clustering algorithm [16], the set of 1600 segments was clustered into 25 sub-word units. The number of sub-word units was chosen based on a loose phonetic description of the digits vocabulary. A final set of reference sub-word units was created by averaging all segments within each of the 25 clusters. The entire process of creating the sub-word unit reference set took about 1 hour on an Alliant FX/8 computer system.

3.1 Auditory Interpretation of Reference Sub-Word Units

To determine if any physical meaning could be attached to the sub-word units, we listened to each pattern and tried to characterize the sound. Since speech waveforms were not available, an LPC resynthesis (with constant pitch) from the autocorrelation coefficients was performed for each sub-word unit. Of the 25 units, 17 produced distinct *vowel-like* sounds, while the remaining 8 segments generated uncharacterizable sounds. This result was mildly encouraging since it implied that there was some underlying structure to the sub-word segments.

IV. Recognition Experiments

For this very preliminary study, our goal was to determine if these new acoustic sub-word units could automatically be detected and used for speech recognition purposes. To accurately assess the potential value of these sub-word units, the problem of lexical decoding had to be dealt with. Since a dictionary, in terms of sub-word units, did not exist, nor was it clear how we could build such a dictionary, we chose to create such a dictionary through a training procedure using actual speech.

4.1 Probabilistic Approach to Lexical Decoding

The simplest way to create a word dictionary, for a small vocabulary speech recognition task, was to use individual (isolated) word tokens to determine how often each of the 25 acoustic units matched each spoken word. If we pay no attention to the temporal sequence of the matches, the resulting dictionary turns out to be equivalent to a single state hidden Markov model [6,7] in which the output probabilities correspond to the probabilities of a sub-word token occurring for a given word.

This simple recognition scheme was tested on a digits vocabulary for a single talker. To determine the segment probabilities, a set of 70 isolated digits (7 replications of each of the 10 digits) was used. To determine which sequences of sub-word units best matched for each training token, a level building algorithm was used which gave the 10 best sequences of concatenated sub-word units for each training token.

An independent test set of 100 isolated digits (10 replications of each of the 10 digits) was used to test the recognition accuracy of this simple system. *No* recognition errors occurred in this case.

4.2 State Diagram Approach to Lexicon Design

To provide a somewhat more difficult test of the recognition potential of the acoustic sub-word units, we ran a connected digits recognition test. In this case, the word dictionary had to be more sophisticated than the simple probabilistic approach used for isolated digits, i.e. we had to introduce temporal information into the lexical decoding. The simplest way of performing this task was to build word reference templates in terms of sequences of the acoustic sub-word units. We used standard training procedures to create a set of 12 reference patterns per digit (from isolated training tokens), where each reference pattern was an optimal sequence of sub-word units as determined from the level building matches. This reference set was then used to recognize the independent testing set, which consisted of 100 connected digit strings of varying length. The standard level-building algorithm for whole word patterns (obtained from the concatenation of the sub-word units) was used as the recognition system. A string recognition accuracy of 96% was achieved. (The 4 string errors all were second choice candidates). Additionally, this system was used to recognize the 100 isolated digit database described above, and again yielded a 100% digit recognition accuracy.

V. Summary

In this paper we have tried to lay a foundation for the use of sub-word units for automatic speech recognition systems. We have presented an automatic segmentation procedure, based on a spectral variation contour, which was used to determine the spectral variation of the speech signal over time. Using this contour we are able to find acoustic events in the speech stream and segment the acoustic signal. This process is fast, easy to use, and requires no human intervention. We have tested this algorithm in a speaker dependent, small vocabulary, speech recognition framework with great success — 100% isolated digit recognition accuracy, and 96% connected digit string recognition accuracy.

REFERENCES

[1] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition – Theory and Selected Applications," *IEEE Trans. on Comm.*, COM-29, No. 5, pp. 621-659, May 1981.

[2] T. B. Martin, "Practical Applications of Voice Input to Machines," *Proc. IEEE*, Vol. 64, pp. 487-501, Apr. 1976.

[3] N. R .Dixon and T. B. Martin, Eds., *Automatic Speech and Speaker Recognition*, New York: IEEE Press, 1979.

[4] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Inst. Acoust.*, Autumn 1979.

[5] G. R. Doddington and T. B. Schalk, "Speech Recognition: Turning Theory to Practice," *IEEE Spectrum*, Vol. 18, No. 9, pp. 26-32, Sept. 1981.

[6] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP*, Vol. 3, No. 1, pp. 4-16, Jan. 1986.

[7] S. E. Levinson, "Structural Methods in Automatic Speech Recognition,". *Proc. IEEE*, Vol. 73, No. 11, pp. 1625-1650, Nov. 1985.

[8] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. IEEE ASSP*, Vol. 64, pp. 532-556, April 1976.

[9] D. H. Klatt, "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," *Jour. of Phonetics*, Vol. 7, pp. 279-312, 1979.

[10] T. Kohonen, et. al., "A Thousand Word Recognition System Based on Learning Subspace Method and Redundant Hash Addressing," Proc. 5th Int. Conf. on Pattern Recognition, Miami Beach, Fl., pp. 158-65, 1980.

[11] L. R. Bahl, J. K. Baker, P. S. Cohen, A. G. Cole, F. Jelinek, B. L. Lewis, and R. L. Mercer, "Automatic Recognition of Continuously Spoken Sentences from a Finite State Grammar," *Proc. IEEE Int. Conf. on ASSP*, Washington, D.C., pp. 418-421, 1979.

[12] S. E. Levinson, A. E. Rosenberg and J. L. Flanagan, "Evaluation of a Word Recognition System Using Syntax Analysis," *Bell Syst. Tech. J.*, Vol. 57, pp. 1019-1626, May-June 1978.

[13] M. Bush and G. Kopec, "Network-Based Connected Digit Recognition Using Explicit Acoustic-Phonetic Modeling," *Proc. IEEE ICASSP '86*, Vol. 2, pp. 1097-1100, Apr. 1986.

[14] G. Leonard, "A Database for Speaker-Independent Digit Recognition," in *Proc. IEEE*, ICASSP '84, Mar. 1984.

[15] A. H. Gray Jr. and J. D. Markel, "Distance Measures for Signal Processing," *IEEE Trans. on ASSP*, ASSP-24, pp. 380-391, 1975.

[16] J. G. Wilpon and L. R. Rabiner, "A Modified k-means Clustering Algorithm for use in Speaker Independent Isolated Word Recognition," *IEEE Trans. ASSP*, ASSP-33, Vol. 3, pp. 587-594, June 1985.

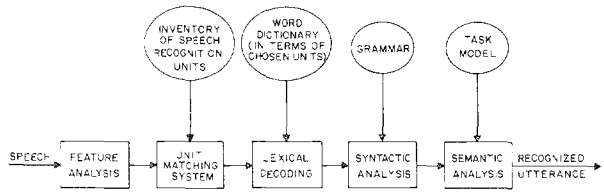

Figure 1 Block diagram of a canonic speech recognition system

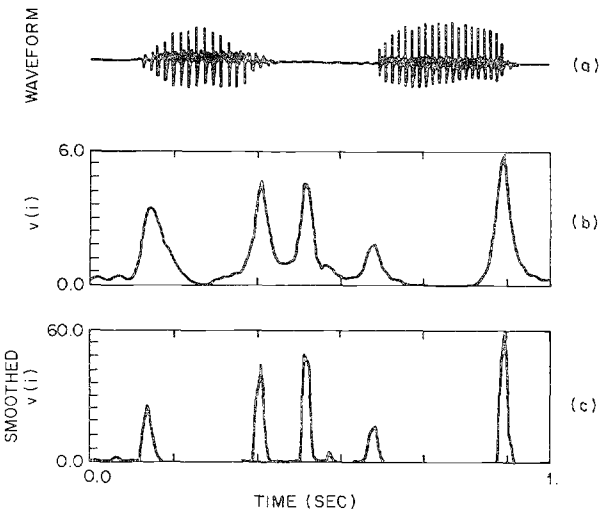

Figure 2 Plot of speech waveform and spectral variation contours