DISAMBIGUATION OF THE E-SET FOR CONNECTED-ALPHADIGIT RECOGNITION

Alain J. Vigier* and Harvey F. Silverman**

* Thomson CSF 66, Rue du Fosse Blanc 92231 Cernievlliers Cedex France
** Brown University, Division of Engineering, LEMS, Providence RI 02912 USA

ABSTRACT
A real-time, talker-dependent, connected-speech recognizer has been operational at the Laboratory for Engineering Man-Machine Systems (LEMS) for 3 years. This recognizer analyses strings of connected digits or alphadigits, using dynamic programming (DP) techniques and an expert for final decision. DP often misclassifies within difficult subgroups of the vocabulary such as the E-set (letters e, p, t, b, d, g, v, z, c). In this paper, we present a feedback mechanism for disambiguation of words classified to be in the E-set by the first pass of the recognizer. This mechanism reanalyses the speech data within an appropriate context and uses the analysis as input to a hidden Markov model (HMM) which uses acoustic, phonetic and linguistic knowledge about the elements of the E-set. Experiments have been run using speech from 5 male talkers. A different model was computed for each talker. Each model was trained with 12 replications of each word and tested with 72 utterances. A recognition rate of 95% was achieved.

1. INTRODUCTION
The current speech recognizer in the Laboratory for Engineering Man-Machine Systems (LEMS) is a real-time, talker-dependent, early-decision recognizer [1]. It analyses strings of connected alphadigits, using DP techniques, and is implemented on a home-built system called the speech-station as described in [2].

Inspection of confusion matrices clearly shows the existence of several highly confusable sets such as the E-set (letters e, p, t, b, d, g, v, z, c) or the A-set (letters a, j, k and the digit 8). In the case of the E-set and for a 24-filter filterbank, the recognizer without feedback achieves 68% recognition within the E-set, while 98% of the inputs that belonged to the E-set were recognized as being in this set.

Template matching is not considered to be the most appropriate way to classify stops or voiced fricatives. Indeed, the duration of the discriminatory stop is much shorter than the duration of the voiced sound that follows the stop. Dynamic programming adds up many small errors from the long vowel, the sum of which is noise; this summed error often offsets the true discrimination information from the shorter stop.

Stop identification is a large stumbling block for accurate speech recognition. Several authors have proposed means for extraction of classification information from stops. Most analyses are feature-based. In [3], 1, 3, 12 LPC spectra were used, the burst was hand-marked and the spectra were taken around it. In [4], spectrograms were used to point-out the stop-discrimination information. It showed a relatively-high stop-recognition performance for numerical measurements (frequencies, bandwidths and amplitude of the poles resulting from an LPC analysis). In [5], some reference templates were built to account for the diffuseness and the general slope of the spectra; a set of rules was developed to decide whether a spectrum matched a template. More recently, in [6], some hidden-Markov models have been used, with maximum mutual information estimation of the parameters. Unfortunately, given all these studies, it is difficult to compare performance. Yet, best reported results are within 90-95%.

We thought we could improve performance of our recognizer by using feedback in context. Therefore, we went back to the original time data and performed a more-sophisticated analysis when a member of the E-set was hypothesized by the LEMS DP recognizer. The segment of the speech that is to be reanalysed is specified by the expert system. Figure 1 shows how the feedback loop has been integrated into the actual connected alphadigit recognizer.

![Integration of Feedback Mechanism](image)

**Fig. 1: Integration of the Feedback Mechanism in the recognizer**

After a fine-time analysis and feature extraction, the data are passed through a specific hidden-Markov model. In our problem, we only considered stops ending with the vowel /ɪ/. Therefore, problems of invariance of the characteristics for the closure across broad phonetic context do not arise. However, the stops are extracted from connected-speech; they are not spoken in isolation and the position of the burst is not hand-marked; this makes the task quite challenging!

In this paper, we focused our attention on finding a "good" HMM model. The initial parameters were obtained from hand-labelled data and were not reestimated. As an alternative to an "ignorance approach", in which a general model is supposed to find the underlying-phonetic structure by itself by means of automatic re-estimation of the parameters, our model tries to make the best of our acoustic, phonetic and linguistic knowledge of the E-set.

The feedback system is described in the next section. Results for the system are presented and discussed in the third section.
II. THE FEEDBACK SYSTEM

If the expert system recognizes a member of the E-set, it is able to give an indication of the temporal location of the sound preceding the word to be recognized as well as the temporal location of the voiced part of the word (\(\bar{f}_0\)). Given these two positions, it is possible to go back to the original, stored time data, currently sampled at 12kHz, and run a more appropriate signal-processing analysis. A block diagram for the whole recognizer was shown in Figure 1.

II.1 Short-Time Analysis

Transitions into a stop, into a burst and the burst itself are very rapid speech phenomena. Each is usually shorter than 50 ms. Therefore, to capture these rapid events in speech, a short-time analysis seemed appropriate. Every 3.3 ms, 40 log-energies in the band [0, 6000 Hz] using an 80-point Winograd Fourier Transform algorithm (WFTA) are computed in real time.

![Fig. 2: Signal processing](image)

A 6.7 ms Hamming window is used, which is small enough to observe closure events. Bursts can be seen on 4 to 10 consecutive frames.

II.2 Acoustic observations

Acoustic observations have to be derived from the 40-component frames, as 40 components are too many. It was decided to extract some specific speech-parameters that observation and speech knowledge imply are important cues for discrimination among stops and fricatives. After several experiments, the following 7 parameters were kept: overall energy, low-frequency parameter, second-formant frequency, spectral peak position and 3 parameters from the parabolic interpolation of the short-time spectrum.

Let |\(Z_i(t)\)|\(^2\) be the power of the DFT component \(i\) (about 150\(^*\) Hz), \(i \in [1,40]\), for the feature vector at time \(t\). The unnormalized overall energy in dB is 10\(\log_{10}\sum\limits_{i=1}^{40}|Z_i(t)|^2\).

(Note that the DC component is appropriately ignored!) This energy was normalized so that the maximum overall energy in the utterance is set to a constant value (45 dB in our experiments). The 40 components in the frame are defined as:

\[
y_i(t) = [10 \log_{10} |Z_i(t)|^2] - \Delta \quad i \in [1,40]
\]

where \(\Delta\) comes from the normalization. Given this notation, the seven parameters are defined as follows.

The overall energy is,

\[
P_1(t) = \sum\limits_{i=1}^{40} y_i(t)
\]

The low-frequency parameter is

\[
P_2(t) = \sum\limits_{i=4}^{10} y_i(t)
\]

which sums the log energies on the band [0, 600 Hz].

The third parameter, \(P_3(t)\) is obtained by means of a formant tracker. The tracker has the peculiarity of looking at the entire hypothesized speech segment - voiced and unvoiced - for detecting the formant. It uses a specifically constrained dynamic programming algorithm. First, it detects the two highest peaks in each of the short-time spectra in the band [750-2400 Hz] where the second formant is virtually always found. The initial detection of these peaks, done independently for each spectrum is shown in Figure 3a. If no peak is found, a single default value of 1500 Hz is assumed.

When searching for a path through the utterance, the algorithm is allowed to skip frames. This is implemented by including some of the peaks from the previous frames with the list for the current frame, as shown in Figure 3b. At this point, the data is sent to a DP algorithm that looks for the best path satisfying the following requirements 1) the final path should be as continuous as possible, 2) the amplitude of the peaks is weighted by a logarithm function. Given these constraints, the DP algorithm is able to find the best suitable path (Fig. 3c). It was found that the tracker performed very well in the context of the experience even for difficult-to-track nasals or liquids among all five talkers.

In order to compute the peak position, \(P_2(t)\) each spectrum is first smoothed by a low-pass cepstral filter and the sequence of peaks (in time) is filtered with a median filter on 5 values.

The three parameters from parabolic interpolation of the whole spectrum (40 points) are obtained by fitting a parabolic curve to each short-time spectrum, using the least-squares method.

**Fig 3:** Second Formant for "we were"
The recognition rate as a function of the number of parameters can be observed in Figure 4. For expediency, these experiments have been done with one male talker only. When feature space was of dimension one, amplitude was the only parameter used. We successively added 1) the low-frequency parameter, 2) the second-formant frequency, 3) the spectral peak position and 4) the parameters from interpolation. It is difficult to evaluate the percentage of information brought by each separate parameter. Figure 4 shows, however, that by increasing their number, we improved the recognition rate. If parameters had been added in a different order, the curve would be different.

The selection of this set was not done in haste. The second-formant is very important for identification of voiced sounds, whether they appear before the elements of the E-set or in the /j/ itself. Also a significant spectral peak in the second-formant locus for diffuse-rising spectra (corresponding to alveolar consonants) has been reported in [5]. The second formant, therefore, contains information for stop identification. The presence of formant resonances in the burst comes from the high coarticulation in speech. It is actually possible to hear the /j/ sound in the burst.

The low-frequency parameter (under 600 Hz) is a good indicator of the amount of voicing. It is used to distinguish the voiced fricative /z/ from the unvoiced fricative /s/. The spectral peak position indicates the location of the most prominent pole in the signal. This information is important for bursts. With the current analysis, the presence of a pole at 2200 Hz for T, 4500 Hz for D... were found.

The parabolic interpolation has been found more useful than a linear interpolation. Not only does it contain the general slope of the spectrum, but it also indicates compactness. This way, some of the information reported in [5] could be used without having to build reference templates and to specify rules for the spectra to fit these templates.

The use of a hybrid set of features requires some weighting. Currently the parameters are scaled by coefficients so that the training data has a zero mean and unity variance for each parameter.

### II.3 Underlying Linguistic and Phonetic Structure

It seemed very appropriate to use left-to-right HMM's for our problem, with initial states corresponding to all possible sounds preceding the actual words to be recognized, and a final state /j/. All 9 possible paths (vocabulary of 9 words) lead to this common final state. The observations associated with the voiced part of the word will be produced by this last state. Only the data preceding the voiced part will be used for discrimination. This circumvent the usual problem of dynamic programming where most of the error is developed from the voiced part. The basic phonetic units that are defined are shorter than phones. These are 1) voiced and 2) unvoiced silence, 3) burst and 4) transitional states from a burst to the voiced sound /j/. The full state diagram is given in Figure 5.

![Fig. 5: The Hidden Model](image)

It was necessary to introduce some transitional states to account for the coarticulation in speech. One of the limitations of the model, as in any Markovian model, is the quasi-stationary hypothesis. It is assumed that any observation is emitted by one phonetic state only. Therefore, the model needs a number of states sufficient to meet this hypothesis, although even with transitional states, it is not easy to take coarticulation into account. The reality looks more like several phonetic units can produce the same observation. Another characteristic of coarticulation is the effect of the burst on the preceding sound. We found, for instance, that the second formant of the /z/ in two drops before a P but not before a T: We tried, however, to carefully choose the phonetic units so that acoustic observations could be assigned as easily as possible to phonetic units.

### II.4 General Structure of the Model

A continuous-variable duration HMM was used, as has been proposed by Levinson [7]. This model includes durational probabilities \( d(t) \) that are the probabilities of remaining in state \( q_i \) for time \( t \). Levinson assumed gamma densities for these
durations. The inspection of history for duration led us choose some trapezoidal distributions described by 4 parameters (Fig. 6). The model is comprised of 39 states (Fig. 5) related to the phonetic entities described above. The output probability distribution for each state q_t is Gaussian.

The model is trained from hand-labelled data, from which it is possible to compute (a) coefficients for scaling, such that each component has a zero mean and unity variance (b) the transition matrix (c) the duration parameters and (d) the Gaussian distributions.

For recognition experiments, the sequence of acoustic observations serves as an input to a Viterbi algorithm that finds the path in the model (sequence of hidden states) that maximizes the joint likelihood of both sequences. The path clearly indicates the recognized E-set alphadigit as it is obvious in Figure 5.

![Fig. 6: Definition of Durational Density Distribution and Parameters [t1 → t4]](image)

### III. EXPERIMENTS

Training and test data were gathered in an ordinary computer room with background and fan noise. The talkers used a standard head-mounted microphone and were asked to utter a particular list of 3 connected alphadigits. The middle letter belonged to the E-set. The first one was chosen to represent a possible sound preceding an alphadigit and the last one was a B. For each talker, 108 utterances (12 replications of each word in the E-set) were used for training and 72 for testing: 5 male talkers were recorded. Three of the talkers were untrained and not all were native American. One talker would stress the first letter, unbalancing the overall energy across the utterance. We also realized that voicing for /z/ can be turned on or off for some talkers. All these observations certainly occur in normal speech but do not help in an automatic recognition task. The training data were hand-labelled using phonetic units corresponding to the different states of the model.

For the testing data, the E-set boundaries to be fed to the model were chosen by hand. Table 1 shows the results of an experiment in which a different model was computed for each talker. Table 2 describes a multi-talker experiment. One unique model has been computed using the training data from the 5 talkers. The average performance of this model is 89%. This model recognizes 72% of the letters from a 6th talker which was not used to train the model.

![Table 1: Results for 5 talkers with Talker-Dependent Training](image)

<table>
<thead>
<tr>
<th>Talker</th>
<th>errors</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>experienced</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hfs</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>ln</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td>gpl</td>
<td>6</td>
<td>92</td>
</tr>
<tr>
<td>inexperienced</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmh</td>
<td>5</td>
<td>93</td>
</tr>
<tr>
<td>km</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>total</td>
<td>18</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 2: Results Using a Single Multi-Talker Training

In these experiments, only the front-end of the training part of the current recognizer was used. This way, the training and the testing set had exactly the same format. The data were then sent to a general purpose computer (SUN 3/260) where the parameters of the model were computed (for the training set) and the Viterbi algorithm was run (for the testing set). The examination of confusion matrices shows for the talker-dependent case some confusion for the letter B that probably come from the fact that the burst in B is of very short duration and does not have a very characteristic distribution. Some Z's have been recognized as C's. We mentioned the fact that some talkers do not voice their Z's depending on the context. For the multi-talker case, errors are more evenly spread across the whole matrix. Although, we have not run any experiment with our full recognizer, these preliminary results - using only the front-end of the recognizer - are very promising.

### IV. CONCLUSION

In this paper we have shown how speech knowledge can be incorporated into Markovian modeling. When using speech parameters as observations, only a small number (7 in our case) is necessary. The phonetic structure of the hidden process was built according to our specific task (E-set recognition). Various models were built by taking the hand-labelled data, which meant that we constrained the models according to what we assumed was the correct correspondence between the acoustic signal and the phonetic symbols. Under such conditions, we achieved a 95% recognition rate in a talker-dependent mode (5 male talkers) and a 89% rate in a multi-talker mode. The next step is to release the constraint of using hand-labelled data by trying different parameter-reestimation techniques and see whether this improves or hurts performance.

### REFERENCES