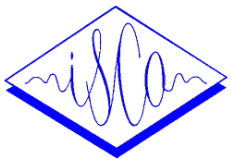

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

HMM modeling in the public telephone network environment : experiments and results

F. Canavesio L. Fissore M. Oreglia P. Ruscitti

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
Tel.: (39) 11 21691

Abstract

A voice-activated inquire system to be used over the telephone network imposes severe constraints upon the speech recognition component: it must be able to accept input speech from untrained speakers, it must be robust against spurious words and extra-linguistic phenomena (environmental noises and telephone line noises) and it must deal with reduced bandwidth. This paper approaches these aspects in the HMM framework.

1 Introduction

In order to achieve the highest degree of success of a complete voice activated service, two major topics have to be considered in the design of the overall system. The former is related to the human-factor aspects [5] of the ASR (automatic speech recognizer) based dialogue; the latter relies upon the bare performance of the ASR itself.

This work focuses on the latter topic, that is on the development of training and recognition algorithms which are effective on the telephone network.
Two significant aspects have been considered. The first one is the presence of additive noise, which affects the signal to be recognized. The large variety of noise sources (telephone line noises, environmental and extra-linguistic noises etc.) makes difficult its characterization. In this work, we have modelled these events in an explicit way, without considering the techniques able to improve the SNR (spectral subtraction, adaptive noise cancelling, etc.).
The second significant aspect is the large variety of speakers, telephone set types and line conditions; it has been approached by collecting a speech data base large enough to be statistically significant. Considering the aspects introduced above, we have focused our interest on the improvement of the performance of the overall system to make it robust enough to the background noise environment.

Each word of the vocabulary is represented in terms of Continuous whole word Hidden Markov Models [4].
The implementation of the overall ASR on commercial hardware (based on a TMS320C30 DSP) ad-hoc programmed is in progress.

Some experiments will be presented, refering to a speech database which has been collected in the application field through both local and long distance calls.
This paper is organized as follows :

- Section 2 presents some details about speech data base collection and handling.
- Section 3 presents the simulated HMM-based recognizer.
- Section 4 is devoted to a review of the obtained results.
- Finally, in Section 5 some conclusions and future directions are drawn.

2 Speech data base collection and handling

The selected vocabulary, consisting of 15 Italian words (digits + 5 command words) has been collected from 1000 speakers, equally partitioned between males and females, with a wide age spread and with different remarkable dialectal intonations.
This database was stored according to the specifications released by the SAM Project (ESPRIT Project no.1541). In particular, the speech sample files and the outputs of the recognition module were stored according to the standards to be able to use the SAM assessment methodology.
The data base is equally partitioned between local and long distance calls. Each speaker, called at home by an operator, uttered the words of the vocabulary through its own telephone handset. The talker was required, by an automatic voice guided procedure, to say each word after a tone, during a temporal window 3 seconds long.
Analog to digital conversion of the signal, band-pass filtered in the bandwidth of 300-3400 Hz, has been performed through a 16 bits A/D converter at 12800 Hz sampling rate. Preemphasis is carried out on the digitalized speech signal by a first order digital network with the transfer function $H(z) = 1 - 0.95z^{-1}$.
Then, an energy-based End-Point Detection (EPD) algorithm finds out the beginning and the end of each token

by a set of thresholds, which are able to adjust their values according to the levels of the signal and of the noise present on the network [3]. The detection of the portion of the signal corresponding to the uttered word is accomplished as a segmented data base is required by the training session (HMM training), at least for the bootstrapping step [9]. Afterwards, a computer-aided check is performed by an operator, through listenings and graphic inspections of speech waveforms to guarantee a good accuracy in the segmentations. Extraneous sounds such as clicks, environmental noises, breath noises, dialling noises have been labeled in order to design an explicit model.
The signal-to-noise ratio (SNR) for the stored utterances has been computed for local and long distance calls as

$$SNR = \frac{E[s^2] - E[n^2]}{E[n^2]}, \qquad (1)$$

where $E[s^2]$ is relevant to the segmented utterance and $E[n^2]$ is relevant to the segments of background noise surrounding the utterance itself. Fig. 1 reports the distribution of the SNR values of the utterances.

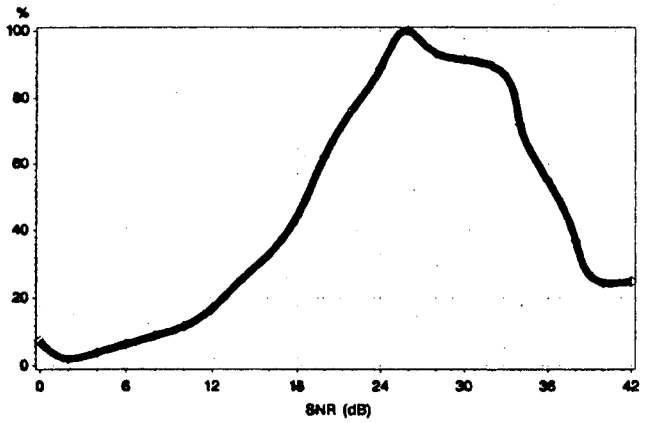

Figure 1: Distribution of the SNR values for the uttered words.

A parametric representation of input speech at each frame of 10 ms is performed by an FFT over the speech weighted by a 256-sample Hamming window with an overlapping of around 50%. The signal bandwidth has been subdivided into 13 Mel-spaced bands by a filter-bank centered on the critical bandwidths of the human auditory model. At each frame, a discrete cosine transform (DCT) is applied, obtaining a vector of 12 cepstral values [2]. The temporal derivative of the cepstral coefficients is computed by a first order orthogonal polynomial :

$$\Delta C_i(l) = \sum_{k=-K}^{K} k \cdot C_i(l-k) \ , \ i = 1, \ldots, N_C, \qquad (2)$$

where l is the examined frame and N_C the number of cepstral coefficients. In the reported experiments $K = 2$ and $N_C = 12$. Moreover, a statistically weighted cepstral distance measure has been inserted, whose advantages in a noisy environment are presented in [8]. Also the differenced energy per frame is considered, providing useful information about the changes in loudness.

3 HMM-based Recognition System

As the Recognition Systems relying upon continuous HMMs have been described extensively in literature [6], only few details will be given in this context.
Each word of the vocabulary is modeled by a left-to-right Hidden Markov Model of first order without skips. A HMM is described by the following parameters :

1. the number of emitting states, N;
2. a matrix of transition probabilities, $\mathbf{A} := \{a_{ij}\}, i = 1, 2, \ldots, N, j = i, i+1$, where a_{ij} is the probability of transition from state i to state j ;
3. matrices $\mathbf{M} := \{m_{ijk}\}$, $\mathbf{\Sigma} := \{\sigma_{ijk}\}$ and $\mathbf{\Lambda} := \{\lambda_{i,k}\}, i = 1, 2, \ldots, N; j = 1, 2, \ldots, 24; k = 1, 2, \ldots, N_G$, whose entries are the means, the variances and the weights of the N_G Gaussian densities that are to approximate the actual densities of the parameters S, according to the expression :

$$p_i(\mathbf{S}) = \sum_{k=1}^{N_G} \lambda_{ik} N(m_{ik}, \sigma_{ik}, \mathbf{S}) \qquad (3)$$

being $N(m, \sigma, \mathbf{S})$ the multivariate Gaussian probability density function with diagonal covariance matrix.

The HMM training is performed by means of the Forward-Backward algorithm after an initial set-up, which is accomplished by a linear segmentation and K-Means clustering [10,7].
The sequence of spectral vectors for each test utterance is matched against the HMMs, which represent the words of the chosen vocabulary, using a Viterbi scoring algorithm to give the optimal alignment state sequence.
A background noise model (made up of a single state) is built by using all the frames in the training database labeled as background noise. The parameters $\mathbf{\Lambda}$, $\mathbf{M}$ and $\mathbf{\Sigma}$ for the noise model are estimated through the K-Means clustering algorithm.
Optional leading and trailing noise state is appended to the states of each word model to account for noise frames possibly included during the segmentation.

In a real application, the recognizer must be ready to correctly detect the presence of extraneous sounds (environmental and telephone line noises). We have approached this aspect with an explicit model trained by using all the frames labeled as spurious sounds. The topology used for this model is characterised by a 12 states left-to-right model

without skip transitions, and the emission probability density distribution of each state is modeled by 8 multivariate Gaussians. Some results are reported in the following section.

Another significant aspect of a real application is the presence of words not belonging to the application vocabulary. The database at our disposal is not well suited to approach this situation, as it consists of isolated words which were collected in a controlled procedure, rather than during actual man-machine interactions. This aspect will be faced when we will test the real prototype operating on the application field.

The porting of the algorithms for feature extraction and Viterbi acoustic decoding on commercial hardware has been planned in two phases.
As first step, we have designed a prototype which relies upon two boards, working in parallel :

- a DSP board (OROS-AU21, based on a TMS320C25) for feature extraction;
- a DSP board (Banshee ASPI, Atlanta Signal Processing, Incorporated) for dynamic programming.

Each of these DSP is connected to a local extension memory. The host system (IBM-PC compatible with AT bus) is the platform to develop the software code for the two boards and to manage the system [1]; it also performs the initialization step and the loading of the knowledges involved in the recognition process (HMM models).
Currently, the implementation of the recognizer on a single DSP board is in progress. This board is part of a Banshee System (ASPI), consisting of integrated hardware/software system based on TMS320C30 DSP at 33 MFLOPS speed. The software environment includes C language, Assembler and SPOX real-time operating system.
The mother board contains 128KW RAM (with memory expansion) and an 8K dual port memory for the communication with the host PC. An I/O Daughter Board such as the AD16 two-channel 16-bit A/D - D/A system can be plugged directly into the Mother Board.

4 Experiments and results

A preliminary assessment of the performance of the modules of the overall system has been carried out on laboratory recorded data collected over the PSTN.

A set of experiments has been performed to set up the parameter values of an Energy-based EPD for providing a reliable detection of the temporal boundaries of a speech utterance inside a recording interval. Moreover, some heuristics have been inserted to account for the presence of noises, which are characteristic of the telephone network environment (e.g. dialling noises). In general, two possible kinds of segmentation errors can be easily encountered : leaving out some speech frames, or inserting a lot of noise frames. To cope with the latter error, which is more frequent than the former one, an optional leading and trailing noise state is appended to the word model, as pointed out before.

The complete evaluation of the performance of the EPD, considered as a component of the overall system, will be carried out both through a comparison of the differences among the automatically detected endpoint locations and the manually determined ones, and by the computation of the word accuracy obtained with automatically segmented speech data.

So far, a preliminary evaluation of the behaviour of the EPD has been carried out in terms of number of insertion and deletion errors, without considering the accuracy of the endpoint locations, as shown in Table 1); this is accomplished by a procedure which performs a temporal alignment among the manual segmentations and the automatic ones.

Triggers	*Insertion errors*	*Deletion errors*
18056	4	412

Table 1: Performance of the EPD

As there is a tight relationship between the behaviour of the EPD and the performance of the global system, it is better to privilege the deletion errors rather than the insertion errors.

The evaluation of the recognition accuracy on the digits vocabulary has been performed with different topologies of the Continuous HMMs to evaluate the influence of the HMM topology on the performance. The experiments have been carried out by using hand-labeled speech data.
In particular, the number of states for each HMM has been selected ranging from 8 to 14, and the number of Gaussian densities for each state is varying from 8 to 16. These experimentations have been performed by using 500 speakers for training and 500 different speakers for test.
Table 2 shows the results obtained by using just one model for each word, while Table 3 reports the performance with two models for each word (one for male and one for female).

Mixtures	*8 States*	*10 States*	*12 States*	*14 States*
8	97.1	97.3	97.6	95.7
10	96.8	97.6	97.9	96.1
12	97.0	97.2	97.5	96.0

Table 2: Continuous HMMs accuracy by a single model per word

One of the most critical problem that must be faced when switching from laboratory studies to real application prototypes is the presence of spurious sounds in conjunction with the application vocabulary words. In order to evaluate the capability of the system to reject these extraneous sounds, we have extracted the spurios portions of the database trig-

Mixtures	*10 States*	*12 States*
8	97.8	97.8
10	97.7	**98.0**
12	97.5	97.5
16	97.4	96.6

Table 3: Continuous HMMs accuracy by two models per word

gered by the EPD.
These tokens have been scored through the Viterbi decoding algorithm by using the HMMs of the digits, the background noise model, the "average" model and the HMM of the extra-linguistic phenomena. Around 91% of extraneous sounds (out of 847) have been correctly labeled.

5 Conclusion and future directions

The evaluation of the performance of the modules which are the components of a simulated ASR operating on speech data collected on the PSTN has been presented. The effects of different topologies of whole word HMM models have been evaluated. Some techniques have been investigated for rejecting both mispronounced or extraneous words.

Further work is planned to improve the performance and the reliability of the recognizer especially as far as the following topics are concerned :

- Speech data base integration with more utterances recorded directly during the field tests;
- Reduction of the computational complexity of Continuous HMMs;
- Experiments performed on the field environment;

As in a real application, the recognizer must be able to deal with words that do not belong to the application vocabulary, and its performance may be affected by extraneous sounds , further work will be devoted to this topic. Our final objective is the most unconstrained scenario, in which a user could utter a legal vocabulary word embedded inside a sequence of words. These situations need a word spotting technique to word recognition that is not currently pursued in this work.

In order to approach the items pointed out before, the databases will be collected during the man-machine interactions rather than in a controlled situation.

References

[1] A. Ciaramella, D. Clementino, and R. Pacifici. **A personal computer based continuous speech recognizer for large vocabulary applications.** In *Proc. of EUSIPCO*, Barcelona, Spain, 1990.

[2] K.H. Davis and P. Mermelstein. **Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences** . *IEEE Trans. Acoust., Speech and Signal Processing*, 28:357–366, 1980.

[3] L. Fissore, M. Codogno, and G. Pirani. **Isolated word recognition in the mobile-radio system: experiments and results.** In *Proc. of EUSIPCO*, Barcelona, Spain, 1990.

[4] B. H. Juang and L. R. Rabiner. **Mixture autoregressive Hidden Markov Models for speech signals.** *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33(6):1404–1413, 1985.

[5] M. Lennig. **Using Speech Recognition in the Telephone Network to Automate Collect and Third-number-billed Calls.** In *Proc. of Speech Tech'89*, pages 124–125, New York City, 1989.

[6] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi. **On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition.** *Bell System Technical Journal*, 62(4), 1983.

[7] L.R. Rabiner, J.G. Wilpon, and B.H. Juang. **A Segmental K-means Training Procedure for Connected Word Recognition Based on Whole Word Reference Patterns.** *AT&T Technical Journal*, 65(3):21–31, May/June 1986.

[8] Y. Tohkura. **A weighted cepstral distance measure for speech recognition.** *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-35(1):1414–1422, Oct. 1987.

[9] J. G. Wilpon and L. R. Rabiner. **Application of hidden Markov models to automatic speech endpoint detection.** *Computer Speech and Language*, 2(3/4):321–341, 1987.

[10] J.G. Wilpon and Rabiner L.R. **A modified K-means Clustering Algorithm for Use in Speaker Independent Isolated Word Recognition.** *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-33(3):587–594, Jun. 1985.