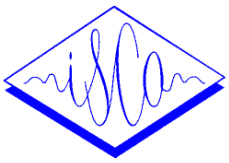

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

HIDDEN MARKOV MODELS FOR NOISY SPEECH RECOGNITION

Adam Wrzoskowicz

Speech Acoustic Laboratory, Institute of Fundamental Technological Research, Polish Academy of Science, Warsaw, Poland

ABSTRACT

This paper presents the development of the speech recognition systems based on Hidden Markov Models. HMM-based isolated words and continuous speech recognizers are presented. The network of connected HMM has been applied to improve accuracy of continuous speech recognition. The HMM network is build on a priori knowledge, for instance: phonetic description, phonological rules, syntax. The recognizers has been studied in an noisy environment. We have proposed two methods of creation Hidden Markov Models for noisy speech recognition. Both can be used in noise auto-adapting systems.

Keywords: *HMM, Noisy Speech Recognition,*

10.21437/Eurospeech.1993-357

1. INTRODUCTION

Hidden Markov Models are parametric models particularly suitable for describing speech events. The success of HMMs is largely due to the Forward-Backward reestimation algorithm. Every iteration of the algorithm modifies the parameters to increase the probability calculated for training data until the local maximum has been reached. The theory of Hidden Markov Models has been widely described in [3][4].

Three recognizers based on HMM are presented. The first is a simply isolated word recognizer designed for small (40 words) vocabulary. This vocabulary is intended to control a scientific calculator (10 digits and 30 commands). The others are continuous speech recognizers. They are based on 39 Hidden Markov Models of Polish phones. They have been trained and studied on phonetic-balanced utterances. The third uses a network of connected HMM.

The main problem is concerned with application of HMM-recognizers in noisy environment. A speech recognizer designed or trained under clean or low SNR conditions suffers serious performance degeneration when used in environment with the different noise characteristic. This problem can be solved in various ways. The most popular is including the noise during training. Every time when noise characteristic has changed this procedure must be repeated. This solution can afford good results but it is very slow. The new method of HMMs modification has been proposed. The algorithm which creates new HMMs for noisy environment from HMMs of clean speech and HMM of noise is described.

2. HIDDEN MARKOV MODELS

A Hidden Markov Model λ is a triple $\lambda=(\pi,A,B)$, where π denotes a initial state probability vector. A is the state transition probability matrix with components a_{ij} = probability of the transition to the state j (at the next instant), given that the system is currently in the state i. The vector π and the matrix A describe a N-state Markov chain.

In our system each segment is modeled by a chain of Markov states. Transitions between them are allowed only in left-to-right direction with no skipping of states.

$$a_{i,j} = 0 \quad \Leftrightarrow \quad i < j \ \vee \ j > j+1 \qquad (1)$$

B is a set of observation probability densities. The probability density function (pdf) of observed vector of signal parameters is connected witch each state. We assume that components of observed vector O are statistically uncorrelated [4]. Thus pdf can be expressed as single multivariate Gaussian function

$$b_i(O) = \frac{\prod_{d=1}^{D} \exp \frac{-(O_d-\mu_{j,d})^2}{2\sigma_{j,d}^2}}{(2\pi)^{D/2} \sqrt{\prod_{d=1}^{D} \sigma_{j,d}^2}} \qquad (2)$$

where, for i-th state, O_d is the d-th component of the observation vector, μ_{id} is d-th component of mean value vector, σ_{id}^2 is d-th component of variance vector, and D is number of components of O.

The Forward-Backward algorithm for training and the Viterbi algorithm for recognizing are used. Those are typical HMM algorithms used for training and recognizing. All procedures are implemented applied HTK Toolkit V1.2 [5].

3. HMM-BASED SPEECH RECOGNIZERS

Three speech recognizers based on HMM have been developed.

The first is a HMM-based isolated word recognizer. It can recognize a vocabulary of 40 Polish words. Each word is modeled as a Markov chain. We have used two states for phone in a word.

The second is a continuous speech recognition system based on Hidden Markov Models of 39 Polish phones. It works with no grammar.

The last one is the most sophisticated. It is a continuous speech recognizer using a network of connected HMM. The HMM network is build automatically on the basis a priori knowledge, for instance: phonetic description, phonological rules, syntax. Generally, the system can work in two modes. The first mode is intended for automatic labelling speech signal using the information provided by grapheme-to-phoneme conversion of the processed utterance. This subject is presented in the other paper [1]. The second mode corresponds to the continuous speech recognition supported by the network modelling of the dictionary and language of communication. The network structure of the whole dictionary is created automatically. Every word is represented by a sub-network of connected phonetic HMMs which includes supplementary information about the word's structure: its orthographic and phonetic transcription, conjugation, declination and irregular forms etc. The system connects sub-networks and builds HMM network used by the recognizer for all admitted utterances.

Both continuous speech recognizers use this same set of Hidden Markov Models of 39 Polish phones. Each phone is represented by a 3-state model.

4. NOISY SPEECH RECOGNITION

The HMM based isolated word recognizer has been studied in noisy environment. In order to show results of noisy speech recognition several experiments were performed. A scheme of the experiments has ben presented on Figure 1.

The speech signals have been recorded in environment without noise. 120 utterances (40 words, 3 repetition) spoke by one man voice have been stored. The speech has been sampled at 10 Khz, and pre-emphasized with a filter whose transform function is $1\text{-}0.95z^{-1}$. All stored signals have been degenerated by Gaussian white noise for SNR= 35, 30, 25 20, 15, 10 and 5 Db. The waveform has been blocked into frames. Each frame has spaned 25.6 msec. Consecutive frames have overlapped by 15.6 msec. Each frame has been multiplied by a Hamming window. From these smoothed speech samples the set of 12 LPC cepstral coefficients have been computed. The following experiments have been performed.

Experiment 1: The conventional HMM-based isolated words recognizer has been studied in noisy environment. Models have been builded using only clean signal. Then, they

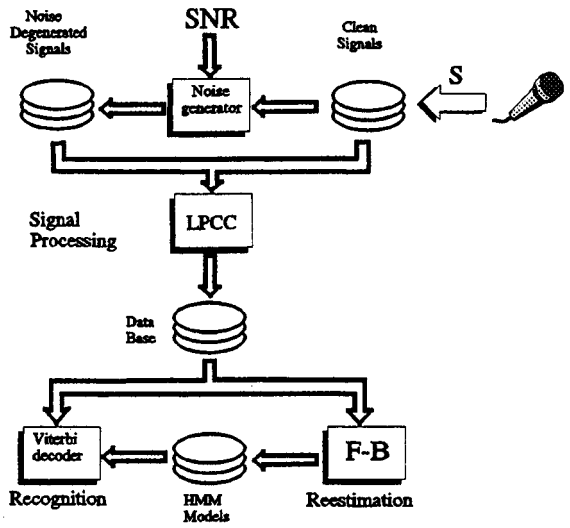

Figure 1

have been used to noisy speech recognition. The results have been presented on Figure 2. It have worked correctly only for clean speech. When SNR≤30dB Correctness<61%.

Experiment 2: The noise adaptive HMM-based isolated words recognizer has been build and studied. Training utterances have been recorded in a clean environment. They have been used to create HMM models for clean speech recognition. When system has had to work in a noisy environment the reestimation procedure has been applied. In the first step a noise has been recorded and than the training utterances have been modified by adding recorded noise to each utterance. Then the HMM models have been modified by the Forward-Backward procedure. In this case only parameters connected with probability density functions (2) have been reestimated. The transitions of the Markov chain have not been modified. The results of recognition have been presented on Figure 2. When SNR≥15dB Correctness>98% and for SNR=5dB Correctness=92%.

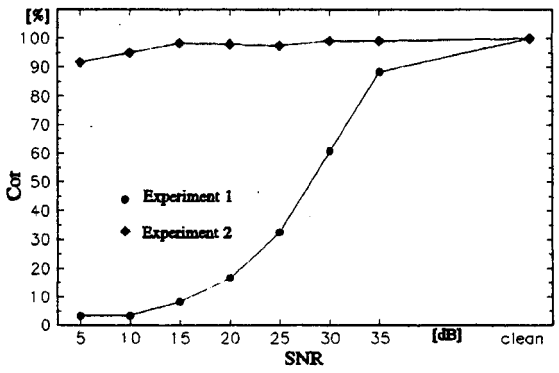

Figure 2

Experiment 3: In this experiment we have been looking for an universal HMM set which is able to work correctly both for clean speech and when the SNR=5dB. The training data base used in reestimation procedure has contained several sets of recorded utterances for different noise level. The creation the universal HMM has appeared impossible but we have found two set of models :HMM1 and HMM2.

HMM1 has been trained using clean and degenerated (SNR= 15, 20, 25, 30, 35 Db) signals. It is able to work correctly when SNR ≥ 20 Db. HMM2 has been trained on noisy signals (SNR= 5, 10, 15, 20 Db) and it can recognize signals when SNR= 5 ÷ 30dB. When SNR= 20 ÷ 30dB both sets of models work correctly. The created by us recognizer can identify a noise level and decided which set of models is to be applied. The results of recognition have been presented on Figure 3. When SNR≥15dB Correctness=100% and for SNR=5dB Correctness=90%.

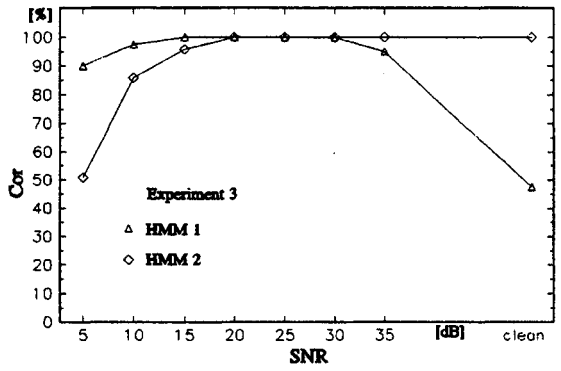

Figure 3

5. Hidden Markov Models Modification

The experiments presented in the previous section show that the creation of HMM-recognizer designed for noisy environment is possible. The proposed procedure can afford good results but if it is applied in auto-adaptive system the HMM must be reestimated when the noise characteristic in a operating room has changed. The reestimation procedure is very slow and it is a main disadvantage of this recognizer. A new Hidden Markov Models modification procedure is proposed which is able work more quickly.

Let $x = [x_0\ x_1\ \dots\ x_{L-1}]$ be a sequence of speech waveform samples which are said to be generated by an autoregressive source satisfying the following relationship:

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n \tag{3}$$

where e_n is a Gaussian white noise sequence and a_i is a component of the well-known LPC inverse filter polynomial.

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \tag{4}$$

The LPC cepstrum of the waveform x is related to the inverse filter polynomial through the expression

$$C(z) = \sum_{i=1}^{k} c_i z^{-i} = -\ln A(z) \tag{5}$$

The transfer function of the system is:

$$F(z) = \frac{K}{A(z)} = K \exp C(z) \tag{6}$$

The signal power E (the energy per frame) can be expressed:

$$E = \int_{-\pi}^{\pi} |F(z=e^{j\omega})|^2 \frac{d\omega}{2\pi} = \\ = K^2 \int_{-\pi}^{\pi} \exp\left(2\sum_{i=1}^{P} c_i \cos(i\omega)\right) \frac{d\omega}{2\pi} \tag{7}$$

We assume that the observed speech waveform contains a disturbance signal v.

$$y_n = x_n + v_n \tag{8}$$

Were x is a clean speech, v is a noise and y is a observed signal.

The new transfer function of generating speech signal and noise can be presented in the z-domain as a sum of transfer functions of speech and noise.

$$F_y(z) = \frac{K_x}{A_x(z)} + \frac{K_v}{A_v(z)} = \\ = K_x \exp C_x(z) + K_v \exp C_v(z) \tag{9}$$

The new set of LPC cepstral coefficients can be computed:

$$C_y(z) = \ln F_y(z) \tag{10}$$

From (9) and (10) using Taylor's theory the following recursive formula can be derived:

$$c_{y,n} = \frac{d_n}{n!} - \sum_{k=1}^{n-1} \frac{(n-k)(n-k)!}{n!} c_{y,n-k}\, d_k \tag{11}$$

where

$$d_n = \frac{K_x}{K_x+K_v} d_{x,n} + \frac{K_v}{K_x+K_v} d_{v,n} \tag{12}$$

and

$$d_{i,n} = n!\, c_{i,n} + \sum_{k=1}^{n-1} (n-k)\, k!\, c_{i,k}\, d_{i,n-k} \tag{13}$$
$$i = \{ x, v \}$$

The parameters K_x and K_v can be computed from (7).
The power of observed signal can be expressed as a sum of the speech power and a the noise power:

$$E_y = E_x + E_v \tag{14}$$

We have proposed a following approximation of the new variances parameters. After linearization of equations (11)-(13) the new variances can be computed:

$$\sigma_{c_{y,n}}^2 = \frac{d_n}{n!}\,\sigma_{d_n}^2 + \sum_{k=1}^{n-1} \frac{(n-k)(n-k)!}{n!}\left[\,|c_{y,n-k}|\,\sigma_{d_k}^2 + |d_k|\,\sigma_{c_{y,n-k}}^2\,\right] \tag{15}$$

where

$$\sigma_{d_n}^2 = \frac{K_x}{K_x+K_v}\,\sigma_{d_{x,n}}^2 + \frac{K_v}{K_x+K_v}\,\sigma_{d_{v,n}}^2 + \left|\frac{d_{x,n}(2K_x+K_v)+d_{v,n}K_v}{(K_x+K_v)^2}\right|\sigma_{K_x}^2 + \left|\frac{d_{x,n}K_x+d_{v,n}(K_x+2K_v)}{(K_x+K_v)^2}\right|\sigma_{K_v}^2 \tag{16}$$

and

$$\sigma_{d_{i,n}}^2 = n!\,\sigma_{c_{i,n}}^2 + \sum_{k=1}^{n-1}(n-k)\,k!\left[\,|d_{i,n-k}|\,\sigma_{c_{i,k}}^2 + |c_{i,k}|\,\sigma_{d_{i,n-k}}^2\,\right] \tag{17}$$

$$i = \{x, v\}$$

The variance of the parameter K can be approximated using the power variance.

$$\sigma_{K_i}^2 = \frac{K_i}{E_i}\,\sigma_{E_i}^2 \qquad i = \{x, y\} \tag{18}$$

The variance of the new power can be computed simply:

$$\sigma_{E_y}^2 = \sigma_{E_x}^2 + \sigma_{E_v}^2 \tag{19}$$

The system designed to noisy environment is presented on Figure 4. The HMM models of clean speech segments are builded as usual. After recording all utterances the signal processing is performed. For each frame the set of LPC cepstral coefficients and E are computed. Then, the set of HMMs for clean speech using reestimation procedure is created. Each model contains the set of means and variance of LPCC coefficients and E for each state. These HMMs can be applied for clean signals recognition.

For a noisy environment the previously calculated HMM models must be modified. In first step HMM model of noise is created in similar way like HMM modes of clean signal. The noise model is very simply. It contains only one state. If it is possible a representative set of noise signal samples in operating room is recorded to create credible noise model. This model contains means and variances of LPCC and E for noise.

In the second step the HMM model for noisy speech are calculated (eq. (11)-(19)). These modified Hidden Markov Models can be used in noisy speech recognition.

The system based on this procedure is currently tested and final results will be presented on the conference.

ACKNOWLEDGMENT

The software HTK Toolkit was kindly supplied by S.Young from Cambridge University, England.

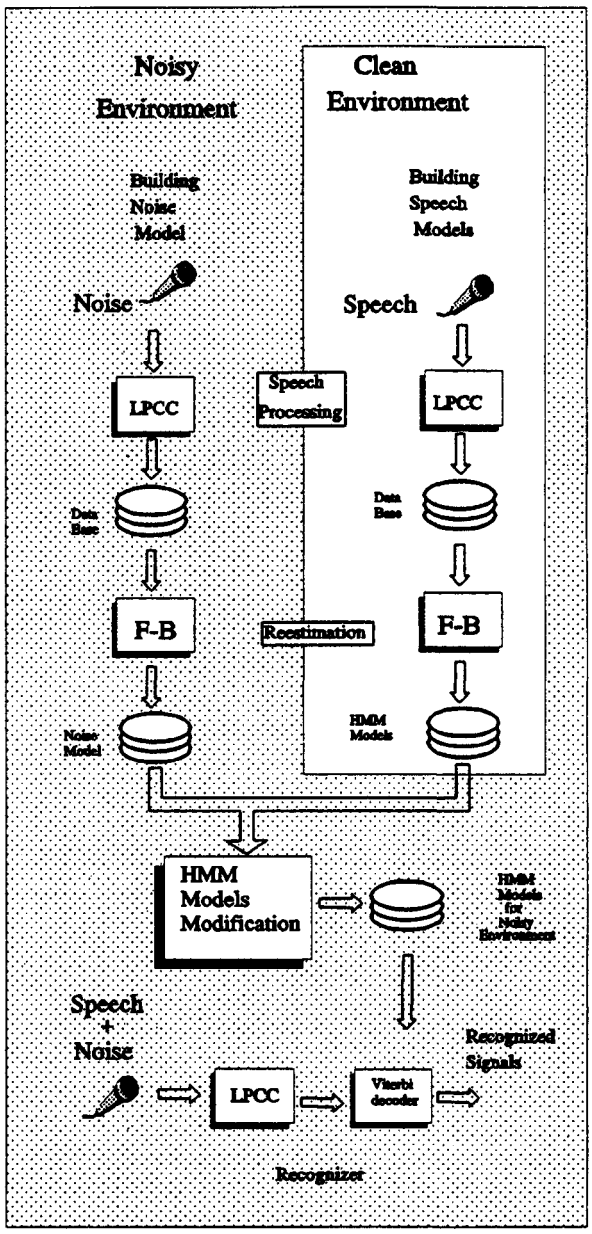

Figure 4

REFERENCES

[1]:*R.Gubrynowicz, A.Wrzoskowicz,* "LABELLER - A System for Automatic Labelling of Speech Continuous Signal", EUROSPEECH '93.

[2]:*B-H.Juang, K.K.Paliwal,* "Hidden Markov Models with First_Order Equalization for Noisy Speech Recognition", IEEE Transaction on Signal Processing, Vol. 40, No. 9, September 1992, pp. 2136-2143.

[3]:*K-F.Lee,* "Automatic Speech Recognition, The Development of the SPHINX System", Kluwer Academic Publishers, 1989.

[4]:*L.R.Rabiner, J.G.Wilpon, F.K.Soong,* "High Performance Connected Digit Recognition Using Hidden Markov Models", IEEE Transaction on Speech and Signal Processing, Vol. 37, No. 8, August 1989, pp. 1214-1225.

[5]:*S.J.Young,* "Hidden Markov Model Toolkit, Reference Manual", CUED, December, 1990.