Real Time Context-Independent Phone Recognition Using a Simplified Statistical Training Algorithm

Othman Lachhab, Joseph Di Martino, El Hassan Ibn Elhaj, Ahmed Hammouch

REAL TIME CONTEXT-INDEPENDENT PHONE RECOGNITION USING A SIMPLIFIED STATISTICAL TRAINING ALGORITHM

*Othman LACHHAB,** *Joseph Di MARTINO,* *El Hassane Ibn ELHAJ,* *Ahmed HAMMOUCH,*

*INPT / ENSIAS
Madinat Al Irfane
Rabat, MOROCCO
othmanlachhab@yahoo.fr

*INRIA / LORIA
Vandoeuvre-lès-Nancy
FRANCE
jdm@loria.fr

*INPT
Madinat Al Irfane
Rabat, MOROCCO
ibnelhaj@inpt.ac.ma

ENSET
Madinat Al Irfane
Rabat, MOROCCO
hammouch_a@yahoo.com

***Abstract*—In this paper we present our own real time speaker-independent continuous phone recognition (Spirit) using Context-Independent Continuous Density HMMs (CI-CDHMMs) modeled by Gaussian Mixtures Models (GMMs). All the parameters of our system are estimated directly from data by using an improved Viterbi alignment process instead of the classical Baum-Welch estimation procedure. Generally, in the literature the Viterbi training algorithm is used as a pretreatment to initialize HMMs models that will be most often re-estimated by using complex re-estimation formula. In order to evaluate and compare the performance of our system with other previous works, we use the TIMIT database. The duration test of our recognition system for each sentence is between 2 seconds (for short sentences) to 12 seconds (for long sentences). We get, by combining the 64 possible phones into 39 phonetic classes, a phone recognition correct rate of 71.06% and an accuracy rate of 65.25%. These results compare favorably with previously published works.**

***Keywords-component*—**Real Time Automatic Speech Recognition (ASR) System, Continuous Speech Recognition, Continuous Density Hidden Markov Models (CDHMMs), Viterbi, Simplified Statistical Trainning Algorithm, Gaussian Mixture Models (GMMs).

I. INTRODUCTION

Implementation of a continuous speech recognition system is difficult because of the large amount of variability in the speech signal. However, there are a lot of possible acoustical units able to represent the speech, the most interesting one is probably the phone which can be considered as the smallest acoustical unit. To model these units, several techniques have been proposed, the connectionist approach with Neural Networks (NN), support vector machines (SVM) and finally the most popular in the field of Automatic Speech Recognition the statistical approach based on Hidden Markov Models (HMMs) [1][2]. The most recent works model the acoustic space by Gaussian Mixture Models (GMMs). Many researchers have introduced the formalism of this technique in their Automatic Speech Recognition (ASR) system [3][4][5][6][7], and they have proved that Continuous Density Hidden Markov Models (CDHMMs) permit to achieve better results than discrete HMMs. In this work, we shall describe our own speaker-independent continuous speech recognition system we call **Spirit**.

The purpose of this paper is not to provide the best phone recognition rates on the Timit database [8], but to demonstrate that by using a simple statistical training algorithm, we can reach similar or better context-independent phone recognition rates than those proposed in the literature.

This paper is organised as follows: in section 2, we explain our HMM training and recognition procedure; in section 3, we present experiments and results; finally in section 4, comparisons with other context-independent phone recognition systems, and some concluding and perspective works are given.

II. THE PHONE RECOGNITION

A. *Speech processing*

Before training the models, it is necessary to prepare the acoustic data by calculating the MFCC feature vectors. The signal is sampled at 16Khz and preeemphasized with a factor of 0.96. The static Mel-Cepstral vectors are computed from windowed time sections of 32ms duration and shifted every 10ms. Every calculated frame consists in 11 first static Mel-Cepstrum coefficients and the log energy(E), (the c_0 cepstrum coefficient was discarded). We also included the first and second order derivatives called dynamic coefficients (Δ and $\Delta\Delta$) in the same high dimentional feature vector. So we work with vectors of dimension d=36 $(11MFCC, E; 11\Delta MFCC, \Delta E; 11\Delta\Delta MFCC, \Delta\Delta E)$.

*This study has been realized in the framework of the INRIA Euro-Mediterranean 3+3 M09/02 OESOVOX project with help of the European COADVISE - IRSES (FP7) program.

B. *Context-independent HMM training*

Each phone of the system is represented by a left-to-right HMM composed of five states (but only three of them are emitting). Fig. 1 illustrates the topology and the type of HMM model used. Learning models is the starting point of any (ASR) system and certainly the most crucial. This consists in determining the optimal parameters $\tilde{\Theta} = \{A, \pi_i, B\}$.

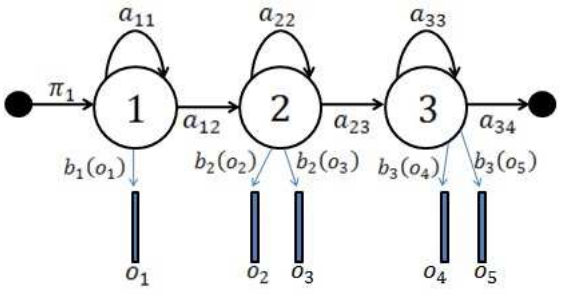

Fig. 1. Topology of the context-independent phonetic HMM.

- π_i : An initial state probability.
- $A = a_{ij}$: The state transition probability matrix.
- $B = b_i(\vec{o}_t)$: The distribution probability of emission of the observations $\vec{o}_t$ in state i.

In a CDHMM the output distribution $b_i(\vec{o}_t)$ for observation $\vec{o}_t$ in state i is generated by a Gaussian Mixture Model (GMM) which corresponds to a mixture of multivariate gaussian distributions of probability $\mathcal{N}(\vec{o}_t, \vec{\mu}_{ik}, \Sigma_{ik})$ with mean vector $\vec{\mu}_{ik}$ and covariance matrix Σ_{ik}:

$$b_i(\vec{o}_t) = \sum_{k=1}^{n_i} \frac{c_{ik}}{\sqrt{(2\pi)^d |\Sigma_{ik}|}} \exp(-\frac{1}{2}(\vec{o}_t - \vec{\mu}_{ik})^T \Sigma_{ik}^{-1} (\vec{o}_t - \vec{\mu}_{ik})) \tag{1}$$

Where n_i represents the number of gaussian components in state i and $\vec{o}_t$ corresponds to an observation at time t of dimension $d = 36$. The $\vec{\mu}_{ik}$ centroids are statistically computed in state i by using the LBG vector quantization algorithm [9] applied to the vectors associated with state i. Each k centroid in state i ($\vec{\mu}_{ik}$) is calculated (see Eq. 2) by an average of its associated cepstral vectors $\vec{o}_{ik}^{\,(n)}$.

$$\vec{\mu}_{ik} = \frac{1}{N_{ik}} \sum_{n=1}^{N_{ik}} \vec{o}_{ik}^{\,(n)} \tag{2}$$

Where N_{ik} represents the number of associated vectors for the k centroid in state i, and in Eq. 3, c_{ik} represents the mixture weight for the centroid k in state i estimated as follows:

$$c_{ik} = \frac{N_{ik}}{N_i} \tag{3}$$

With N_i is the total number of vectors associated with state i. Σ_{ik} is the covariance matrix associated with the gaussian k of state i which is computed directly from the data using the classical estimation formula (4):

$$\begin{aligned} \Sigma_{ik} &= E((X - E[X]).(Y - E[Y])) \\ &= \frac{1}{N_{ik}} \sum_{n=1}^{N_{ik}} (\vec{o}_{ik}^{\,(n)} - \vec{\mu}_{ik})(\vec{o}_{ik}^{\,(n)} - \vec{\mu}_{ik})^T \end{aligned} \tag{4}$$

It is important to say that the number of gaussian components associated to each state must be chosen, by making a compromise between a good modeling of the phonetic HMMs and the limited amount of training data. A too high number of gaussian components compared to the amount of available data leads to a bad learning because the training database has a limited number of samples for each phone. For this reason we optimize the number of gaussian components in each HMM state. We begin by setting the number of gaussian components for each state to 16. The actual optimum number of gaussian components is related to the number of MFCC vectors associated to each centroid: if the latter is less than the dimension d, then the associated gaussian component is removed because its covariance matrix will be non-invertible. The associated vectors with this removed gaussian component are then redistributed to the nearest remaining centroids.

The state transition probabilities were evaluated by Eq. 7. The same principle of this method has been successfully applied to various specialized tasks, such as speaker-independent alphabet letters recognition [7] and voice conversion [10] .

Let $\mathcal{X}$ be a random variable giving the number of times a HMM state is visited. If we consider event $\mathcal{S}_j$ "Staying j times in the same state" and $\mathcal{M}_j$ "Moving to the next state at the j-th time".
Then event $[\mathcal{X} = l]$ can be expressed by :

$$[\mathcal{X} = l] = \underbrace{\overbrace{S_1 \bigcap S_2 \bigcap \cdots \bigcap S_{l-1}}^{\mathcal{S}_j} \bigcap \overbrace{M_l}^{\mathcal{M}_j}}_{intersection\ of\ independent\ events}$$

Then the probability distribution of $\mathcal{X}$ is given by:

$$P(\mathcal{X} = l) = p_s^{l-1}.p_m \tag{5}$$

Where p_s is the probability to stay in the same state and $p_m = 1 - p_s$ is the probability to move to the next state.
Then by definition the expectation of $\mathcal{X}$ is given by:

$$E[\mathcal{X}] = \sum_{l=1}^{+\infty} l.p_s^{l-1}.(1 - p_s) = \frac{1}{1 - p_s} \tag{6}$$

Consequently :

$$p_s = \frac{E[\mathcal{X}] - 1}{E[\mathcal{X}]} \tag{7}$$

The expectation $E[\mathcal{X}]$ is calculated directly from data by the following formula:

$$E[\mathcal{X}] = \frac{N_{ip}}{R_p} \tag{8}$$

Where N_{ip} is the total number of vectors related to state i of phone p and R_p is the total number of samples of phone p in the training data space.

The Viterbi algorithm was applied to the acoustic vectors of each sentence to determine an optimal sequence of states which has produced the best sequence of observations. This process is iterated several times until a stability criterion, calculated from the paths returned by the viterbi process, has been reached. The maximal number of iterations was 20.

C. *Monophone HMM recognition*

Continuous speech recognition is a difficult process because we do not know the boundaries of the phones making up a sentence. Furthermore the monophone HMMs assume that speech is produced as a concatenation of phones, not affected by the phonetic context neighbors. To perform the recognition it is useful to infer the sequence of states that has generated the given observations. Actually, from the sequence of states we can easily find the phone string: this task is performed by the Viterbi decoding algorithm applied on each test sentences using the optimal parameters (A, π_i, B). To better carry out this task and find the adequate path, we built a bigram language model and a duration model on the phone durations witch we assume to follow a normal distribution ($\mathcal{N}(\mu, \sigma^2)$).

III. EXPERIMENTS AND RESULTS

The **Spirit** system has been implemented and tested on a linux machine with an Intel Pentium Dual CPU 1.86GHz and 2GB of RAM. We choose to evaluate our ASR system with the TIMIT database [8]. In this database a total of 64 phonetic labels, generally considered too detailed for learning HMMs models, has been reduced to 39 classes by K.F. Lee and H.W. Hon [11]. We used the same labeling in our system. 39 phonetic HMMs with the same topology described in Section 2.B (see Fig. 1) are used in the training and testing, with a total states of 3x39 = 117. These HMMs, the bigram model and the duration model are learned on 8 sentences "si" and "sx" of 462 speakers of the TIMIT database training part, corresponding to 3696 sentences. In the test 1344 sentences, pronounced by 168 speakers, corresponding to a total number of 50754 phones. The "sa" calibration sentences are excluded in both training and testing. In continuous speech recognition, the most common phone recognition evaluation measures are the phone error rate (PER), or the related performance metric, phone accuracy. These measures, calculated by Eq. 9 are used in this paper for making comparisons between the different phone recognition systems.

$$Accuracy = \frac{N-(S+D+I)}{N} \quad Correct = \frac{N-(S+D)}{N} \tag{9}$$

Where N is the total number of labels in the reference utterances and S, I and D (resp.) the Substitution, Insertion and Deletions errors, computed by a DTW algorithm (Dynamic Time Warping) between the correct phone strings (reference) and the recognized phone strings (test).

Table.1 presents the accuracy obtained by our system using the complete TIMIT test set. The phone recognition correct rate is 71.06% and the accuracy rate is 65.25%.

39 Monophone	Bigram	Bigram+Duration
Substitution	17.61% (8938)	17.25% (8756)
Deletion	10.46% (5310)	11.69% (5932)
Insertion	7.11% (3607)	5.81% (2951)
Correct	71.93% (36506)	71.06% (36066)
Accuracy	64.82% (32899)	65.25% (33115)

Table 1. Phone recognition results with our context-independent phone HMM system on all the TIMIT test set

Two other measures were chosen to evaluate the speed of our recognition system. The first is the average recognition time. In this case, the test sentences are classified into six categories (see Table. 2) according to their total number of phones N_{ph}. It is clear that the recognition time depends on the duration of the sentences to recognize. The second measure is the Real Time Factor (RTF) defined as the total computation time for recognition, divided by the total duration of the recorded speech processed. We obtain an acceptable speed with an RTF of 2.5.

Total number of phones N_{ph}	Average time in second (s)
$10 \leq N_{ph} < 20$	2 s
$20 \leq N_{ph} < 30$	5 s
$30 \leq N_{ph} < 40$	7 s
$40 \leq N_{ph} < 50$	9 s
$50 \leq N_{ph} < 60$	11 s
$60 \leq N_{ph} < 75$	12 s

Table 2. Average recognition times for the six categories of Timit test sentences.

Fig. 2 shows the evolution of the phone accuracy versus the number of iterations of the proposed training algorithm, by varying the shift from 8 to 10ms. We note that our (ASR) system is more efficient using a shift of 10ms. This behaviour can be explained by the fact that by decreasing the shift value, the number of insertion errors increased.

IV. COMPARISONS

Table.3 provides an accuracy comparison, between our **spirit** system with previously published results on the Timit database for the phone recognition task, using CI-CDHMMs. These systems differ by their learning approach of the phonetic model, level complexity, time computation etc; which makes this comparison a very difficult task. But we have demonstrated that by using a simple minded system, we can reach in real time a competitive accurracy in comparison with those obtained by other researchers.

V. CONCLUSION AND FUTURE WORKS

In this paper, we built a reference system for continuous speech recognition using context-independent phonetic HMMs. We show that the obtained results compare favorably with already published HMM technology. In the future we forsee to test our system using context-dependent phone models, and implement a new technique to locate the position of the insertion errors, in order to remove them.

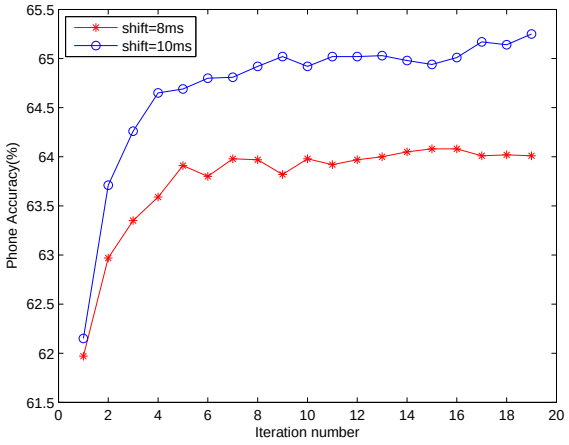

Fig. 2. Evolution of phone accuracy versus the number of iterations of the training algorithm.

System	Correct	Accuracy
discrete HMM (monophone) [11]	64.07%	53.28%
Tandem (monophone) [12] [13]	63.50%	61.48%
HTK (monophone) [6]	71.9%	62.8%
CDHMM (monophone) [5]	69.33%	63.05%
CDHMM (monophone) [4]	—	64.1%
CRF (monophone) [13]	66.74%	65.23%
CDHMM (monophone) this paper	**71.06%**	**65.25%**

Table 3. Phone accuracy comparisons using TIMIT

VI. REFERENCES

[1] J. Baker, "The dragon system–an overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 24–29, 1975.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *IEEE Proceedings*, vol. 64, pp. 532–556, Avril 1976.

[3] J.L. Gauvain L.F. Lamel, "High performance speaker-independent phone recognition using cdhmm," *Proc. Eurospeech*, vol. Berlin, pp. 121–124, September 1993.

[4] J. Chang J. Glass and M. McCandless, "A probabilistic framework for feature-based speech recognition," *in Proc. of the ICSLP*, pp. 2277–2280, October 2000.

[5] Z. Ben Hamida A. Ben Messaoud, "Cdhmm parameters selection for speaker-independent phone recognition in continuous speech system," *IEEE Mediterranean Electrotechnical Conference*, pp. 253–258, 2010.

[6] S.J. Young, "The general use of tying in phoneme-based hmm speech recognisers," *ICASSP*, vol. 1, pp. 569–572, 1992.

[7] J. Di Martino, "On the use of high order derivatives for high performance alphabet recognition," *ICASSP*, vol. Orlando, pp. USA, May 2002.

[8] L. Lamel W. Fisher J. Fiscus D. Pallet J. Garofolo and N. Dahlgren., "The darpa timit acoustic-phonetic continuous speech corpus cdrom. ntis order number pb91-505065," October 1990.

[9] A. Buzo Y. Linde and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications, Vol*, vol. COM-28, pp. No.1, January 1980.

[10] J. Di Martino S. Ben Jebara A. Werghi, "On the use of an iterative estimation of continuous probabilistic transforms for voice conversion," *ISIVC*, 2010.

[11] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. ASSP*, vol. 37(11), pp. 164–1648, November 1989.

[12] D. Ellis H. Hermansky and S. Sharma, "Tandem connectionist feature stream extraction for conventional hmm systems," *in Proc. of the ICASSP*, 2000.

[13] J. Moris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," *in Proc. of the InterSpeech*, pp. 597–600, 2006.