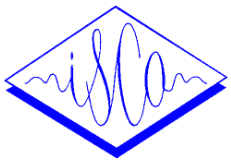

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

MULTIPLE MULTILABELING TO IMPROVE HMM-BASED SPEECH RECOGNITION IN NOISE

J. Hernando, J.B. Mariño and C. Nadeu

Department of Signal Theory and Communications. Polytechnical University of Catalonia
Ap. 30002, 08071 Barcelona, Spain
javier@tsc.upc.es

ABSTRACT

The performance of existing speech recognition systems degrades rapidly in the presence of background noise when training and testing cannot be done under the same ambient conditions. The aim of this paper is to propose the application of a simple multilabeling method, instead of the standard vector quantization -so called labeling-, as the front end for a speech recognizer based on the Vector Quantization (VQ) and Hidden Markov Models (HMM) approaches in order to increase its robustness to noise. Furthermore, not only cepstrum but also other features such as energy and dynamic parameters are evaluated and quantized independently in the multilabeling stage to represent more accurately characteristics of speech. The result of this process is a multiple multilabeling. Experimental results in the presence of additive white noise and car environment clearly demonstrate its good performance in isolated word recognition in noisy environments.

Keywords: *Noisy speech recognition, Vector Quantization, Hidden Markov Models.*

1.INTRODUCTION

Speech recognition in noisy environments remains an unsolved problem even in the case of isolated word recognition with small vocabularies. In order to develop a system that operates robustly and reliably in the presence of noise, many techniques have been proposed in the literature [1] for reducing noise in each stage of the recognition process, particularly, in feature extraction and similarity measuring.

A spectral estimation technique widely used in speech processing and, particularly, in speech recognition is linear predictive coding (LPC) [2], equivalent to an AR modeling of the signal. Concretely, it has been shown that the use of the liftered LPC-cepstral coefficients in the standard Euclidean distance measure lead to the best results of those obtained with this model in both noise free [3] and noisy [4] conditions. Moreover, it is well known that the use of a relatively high prediction order can provide more robust estimations in the presence of noise.

However, due to the sensitivity to the presence of additive noise of the standard LPC technique and the other parameterization techniques used in speech recognition, the performance of existing speech recognition systems degrades noticeably in the presence of background noise when training and testing cannot be done under the same ambient conditions.

On the other hand, Hidden Markov Models (HMM) have become the most popular automatic speech recognition tool because of its capability of representing speech variability as statistical parameters. These models can use discrete output probability distributions (DHMM) or continuous output pdf's (CHMM) [5].

The CHMM approach model the acoustic observation directly using estimated continuous pdf's. However, a mixture of large number of them are generally required, that increase not only the computational load, but also the number of free parameters that must be reliably estimated.

For the DHMM case, vector quantization (VQ) makes it possible to use a histogram oriented nonparametric characterization of the observed speech signals and solves those problems. Nevertheless, the standard VQ stage makes a hard decision as to which of its codewords is the best match for each acoustic observation, and so the information about the degree to which the input vector matches other codewords is discarded for the subsequent hidden Markov modeling. This information would be especially important in the case of noisy speech recognition, because that hard decision can be easily modified by the noise added to the speech.

To accommodate this information lost, several techniques have been proposed that use information provided in the neighbouring codewords, such as multilabeling [6], fuzzy VQ [7], semicontinuous HMM (SCHMM) [8], smoothing [9], etc.

The aim of this paper is to propose the application of a simple multilabeling method, instead of the standard vector quantization, as the front end for a speech recognizer based on the Vector Quantization (VQ) and Hidden Markov Model (HMM) approaches speech in order to increase its robustness to noise. This modeling approach will be named in this paper MLHMM (MultiLabeling HMM).

Furthermore, not only cepstrum but also other features such as energy and dynamic parameters are evaluated, applying regression analysis [10], and quantized

This work has been supported by the grant TIC 92-0800-C05/04

independently in the multilabeling stage to represent accurately characteristics of speech. The result of this process is a multiple multilabeling. Experimental results in the presence of additive white noise and real noisy car environment clearly demonstrate its good performance in isolated word recognition in noisy environments.

This paper is organized in the following way. In section 2 the multiple multilabeling method is briefly revised and its relationship with the other techniques mentioned above is discused. Section 3 reports the application of these techniques to recognize isolated words in a multispeaker task in order to compare their performance and gain some perspective of the merit of the multiple multilabeling method in noisy environments. Finally, in section 4 some conclusions are summarized.

2. MULTIPLE MULTILABELING

For improving the VQ accuracy in a simple manner, firstly a multilabeling method is introduced in this section. Unlike the standard VQ, this multilabeling method makes a soft decision about which codeword is closest to the input vector, generating an output vector whose components indicate the relative closeness of each codeword to the input.

Let the codewords of the multilabeling codebook be $\{v_k\}_{k=1..M}$, where M is the codebook size, and let the cepstral vector in the instant t be x_t. The multilabeling codebook maps the input vector x_t into an output vector $O_t=\{w(x_t,v_k)\}_{k=1..M}$ estimated as

$$w(x_t,v_k) = \frac{1/d(x_t,v_k)}{\sum_{m=1}^{M} 1/d(x_t,v_m)}, \qquad (1)$$

where $d(x_t,v_k)$ is the Euclidean distance between v_k and x_t. The expression (1) is the same that the fuzzy VQ rule with degree of fuzziness equals to 2 [7].

These components are positive, sum to 1 and are decreasing with $d(x_t,v_k)$. Thus O_t can be interpreted as a probability mass vector describing the probability that input vector x_t was drawn from the class represented by the codeword v_k.

The DHMM algorithms must be generalized to accommodate this multilabeling output. For a given state j of the HMM, the probability that a vector x_t is observed can be written as

$$b_j(x_t) = \sum_{k=1}^{M} w(x_t,v_k)\, b_j(k), \qquad (2)$$

where $b_j(k)$ denotes the discrete output probability associated with the codeword v_k and the state j.

Forward-backward and Viterbi algorithms are simply generalized using (2) instead of $b_j(k)$. With respect to the training problem, Baum-Welch reestimation formulas for the transition probabilities a_{ij} and initial state probabilities π_i are generalized in the same manner. However, in the case of the reestimation of $b_j(k)$, the maximum likelihood estimation yields this new formula for the case of a training sequence of length T, for k=1,...,M and j=1,..,N (number of states)

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)\,\frac{w(x_t,v_k)\,b_j(k)}{b_j(x_t)}}{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)}, \qquad (3)$$

where $\alpha_t(j)$ and $\beta_t(j)$ are the well known forward and backward probabilities, respectively.

Nevertheless, in the experiments reported in this paper an alternative reestimation formula has been used that only depends on the the term $w(x_t,v_k)$ and does not depend on the probability $b_j(k)$ in the iteration before:

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)\,w(x_t,v_k)}{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)}, \qquad (4)$$

This formula does not guarantee the convergence of the training process but, if only two or three iterations are performed, it yields better recognition rates than those obtained using maximum likelihood formula. It is due to formula (4) benefits the probability of the closest codewords to the input vector. The use of this reestimation formula reduces the amount of computational load in the training phase.

On the other hand, in practice, (2) can be simplified with K most significant values of $w(x_t,v_k)$ for each x_t without affecting the performance. Since K is of lower order than the codebook size M, this simplification also reduces the computational load in training and testing phases.

Because of those both factors, use of reestimation formula (4) and simplification of (2) with only K codewords, the MultiLabeling Hidden Markov Models (MLHMM) approach becomes extraordinary efficient.

The multilabeling method revised here is essentially the same described in [7], substituting the Itakura distortion measure between vectors of predictor coefficients by the Euclidean distance between cepstral vectors, fixing the degree of fuzziness to 2 and simplifying (2) with only K codewords. With respect to the multilabeling method described in [6], the main discrepancies are the the different generalization of HMM algorithms and that the number K of codewords considered in [8] is different for each input vector.

Compared with the closely related semicontinuous HMM approach (SCHMM) [8], the main difference is that the components of the output vector of the codebook, estimated from the deterministic viewpoint in the multilabeling method, are estimated from the stochastic viewpoint in the semicontinuous approach. Concretely, in the semicontinuous approach the codebook is modeled as a parametric family of mixture Gaussian densities. Furthermore, parameters of the codebook and models can be mutually optimized to achieve an optimal model/codebook combination. This mutual optimization of models and

codebook can be made also in the case of the multilabeling method, but in this paper this option has not been considered because of its computational complexity.

As it will be seen in section 3 the recognition accuracy obtained with both MLHMM and SCHMM approaches is similar. However, the multilabeling method proposed in this paper is more computationally efficient than the semicontinuous approach.

Also there is a relationship between these techniques and the smoothing techniques revised in [9] in the sense that all of them use information provided in the neighbouring codewords. Nevertheless, the way of this information is obtained and used is different.

The performance in noisy speech recognition of MLHMM, SCHMM and smoothing techniques will be compared in section 3 when only the static cepstrum of the speech signal is used.

Finally, in this work, in order to represent accurately characteristics of speech not only the static spectrum but also other features such as energy and dynamic parameters are evaluated, applying regression analysis [10], and quantized separately in the multilabeling stage. The result of this process is a multiple multilabeling and then each feature is considered statistically independent in the HMM framework. Experimental results shown in section 3 demonstrate that this process yields excellent results in noisy speech recognition.

This multiple multilabeling method is essentially different to the multidimensional multilabeling approach proposed in [8], in which each feature is also quantized separately in the multilabeling stage but their multilabels are combined at each frame.

3. RECOGNITION EXPERIMENTS

3.1. Speech databases and recognition system

Two different speech databases were used for each type of noise considered: additive white noise and real noisy car environment.

In the case of additive white noise, the database consists of ten repetitions of the Catalan digits uttered by seven male and three female speakers (1000 words) and recorded in a quiet room. Clean speech was used for training in all the experiments. Noisy speech for testing was simulated by adding zero mean white Gaussian noise to the clean signal so that the SNR of the resulting signal becomes ∞ (clean), 20, 10 and 0 dB. Firstly, the system was trained with half of the database and tested with the other half. Then the roles of both halves were changed and the reported results were obtained by averaging those results.

The database used in noisy car environment experiments is from ESPRIT-ARS project and consists of 25 repetitions of the Italian digits uttered by 4 speakers, 2 males and 2 females, inside a car in different noisy conditions: 5 repetitions with the engine and the fan off and 20 more with the engine on and different fan positions, 10 with the car stopped, 5 with the car running at 70 km/h and 5 with the car running at 130 km/h. The system was trained with the signal uttered with the engine and the fan off, i.e., in noise free conditions, and in the test phase the noisy signals were used.

In the parameterization stage, the speech signal, sampled at 8 kHz, quantized using 12 bits per sample, manually endpointed and preenfasized in the case of noisy car environment, was divided into frames of 30 ms at a rate of 15 ms and each frame was characterized by its liftered LPC-cepstral parameters. In some tests the log-energy and the dynamic parameters of the frame were also obtained. Each information was vector-quantized separately by means of a codebook of 64 codewords, using the standard VQ, the semicontinuous VQ or the multilabeling approaches. Each digit was characterized by a first order, left-to-right, Markov model of 10 states without skips. The parameters of the model were smoothed only in one of the experiments reported in this paper. Training and testing were performed using Baum-Welch and Viterbi algorithms, respectively.

3.2. Recognition results

The first experiments carried out with the above described speech recognition system consisted of empirically optimizing the prediction order and the type of cepstral lifter using only static cepstrum and standard VQ. The best results were obtained using prediction order equals to 12 and slope lifter in the case of white noise and prediction order equals to 16 and inverse of standard deviation lifter in the case of noisy car environment. These optimum orders and cepstral lifters were used in the experiments described bellow.

Using the multilabeling method to quantize the static cepstrum in the case of additive white noise, the best results were obtained when the five closest codewords to the incoming vector were considered (K=5). In table 1, multilabeling recognition rates (MLHMM) are compared with those obtained using standard VQ (DHMM) and semicontinuous VQ with K=5 (SCHMM) and also with those obtained applying the Parzen method on the discrete models (DHMM-Parzen), the best results provided by the smoothing techniques in our experiments.

Models /SNR(dB)	∞	20	10	0
DHMM	99.8	98.9	89.5	54.2
DHMM-Parzen	99.3	98.6	96.0	69.8
SCHMM	99.8	98.9	96.4	72.8
MLHMM	99.7	98.8	96.5	74.1

Table 1

It is clear from table 1 that the results of all the techniques that use information provided in the neighbouring codewords outperform noticeably DHMM results in noisy conditions. However, in noise free conditions only SCHMM and MLHMM results are comparable with DHMM results due to the increase of confusion associated with the smoothing techniques.

On the other hand, it also can be seen that SCHMM and MLHMM recognition rates are very similar in all the conditions considered in this study. Taking into account that the computational load in the multilabeling method is lower than in the semicontinuous approach, it is clearly preferable the multilabeling method in this application.

In figure 1, the results obtained adding energy and dynamic features to the DHMM and MLHMM approaches, also in the case of additive white noise, are shown: only

static cepstrum and standard VQ (a), cepstrum with energy and dynamic features and standard VQ (b), only static cepstrum and multilabeling (c), cepstrum with energy and dynamic features and multilabeling -i.e. multiple multilabeling- (d). Dynamic information consisted of delta-cepstrum and delta-energy, estimated using a window length of 90 ms, and delta-delta-cepstrum and delta-delta-energy, estimated using a window length of 120 ms.

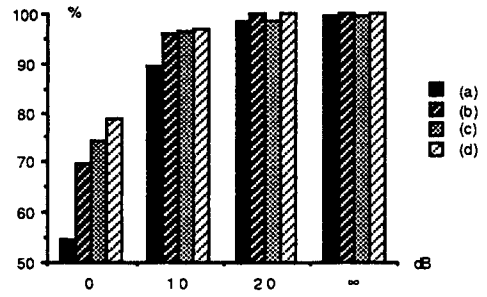

Figure 1

Finally, for the case of real noisy car environment, the recognition rates obtained using DHMM and MLHMM approaches with or without the addition energy and dynamic features are shown in figure 2: only static cepstrum and standard VQ (a), only static cepstrum and multilabeling (b), cepstrum with energy and dynamic features and standard VQ (c), cepstrum with energy and dynamic features and multilabeling -i.e. multiple multilabeling- (d). Dynamic information consisted of delta-cepstrum and delta-energy, estimated using a window length of 240 ms.

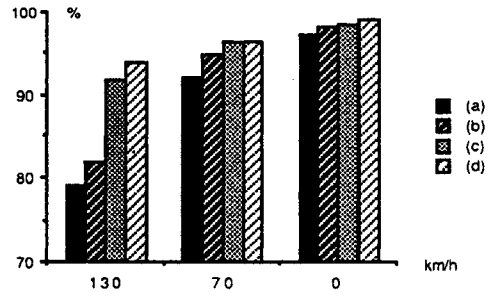

Figure 2

It is clear from figure 1 and 2 that the use of the multiple multilabeling, i.e. separate multilabeling of each feature of speech, yields excellent results in both noisy conditions: additive white noise and real noisy car environment. For more details about experimental results and conditions, see [11].

4. CONCLUSIONS

In this paper a multiple multilabeling method has been proposed to improve HMM-based noisy speech recognition. A multilabeling stage is used, instead of the standard VQ, as the front end for a speech recognizer based on the VQ and HMM approaches in order to increase its robustness to noise. This method is computationally very efficient, unlike the continuous HMM approach, and experimental results in the presence of additive white noise and real noisy car environment clearly demonstrate that its use in noisy speech recognition outperforms considerably the standard VQ. On the other hand, compared with the closely related semicontinuous HMM approach, the multilabeling recognition rates are slightly better than the semicontinuous results and the computational load is lower in our technique. Furthermore, not only the spectrum but also other features such as energy and dynamic parameters are evaluated and quantized independently in the multilabeling stage. Using this multiple multilabeling stage, excellent results have been obtained in isolated word recognition in noisy conditions.

ACKNOWLEDGMENTS

The authors would like to thank Jordi Cobo and Joan Dachs for their help in the software development.

REFERENCES

[1] B.H. Juang, "Speech Recognition in Adverse Environments", Computer Speech and Language, vol. 5, 1991, pp. 275-294.

[2] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. ASSP-23, 1975, pp. 67-72.

[3] B.H. Juang, L.R. Rabiner, J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recogniton", IEEE Trans. ASSP-35, 1987, pp. 947-954.

[4] B.A. Hanson, H. Wakita, "Spectral Slope Based Distortion Measures for All-Pole Models of Speech", IEEE Trans. ASSP-35, 1987, pp. 968-973.

[5] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications on Speech Recognition", Proc. IEEE-77, 1989, pp. 257-286.

[6] M. Nishimura, K. Toshioka, "HMM-Based Speech Recognition Using Multidimensional Multilabeling", Proc. ICASSP-87, Dallas, pp. 1163-1166.

[7] H.P. Tseng, M.J. Sabin, E.A. Lee, "Fuzzy Vector Quantization Applied to Hidden Markov Modeling", Proc. ICASSP-97, Dallas, pp. 641-644.

[8] X.D. Huang, "Phoneme Classification Using Semicontinuous Hidden Markov Models", IEEE Trans. SP-40, 1992, p. 1062-1067.

[9] R. Schwartz, O. Kimball, F. Kubala, M. Feng, Y.Chow, C. Barry, J. Makhoul, "Robust Smoothing Methods for Discrete Hidden Markov Models", Proc. ICASSP-89, Glasgow, pp. 548-551.

[10] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. ASSP, vol. 34, 1986, pp. 52-59.

[11] J. Hernando, "Técnicas de Procesado y Representación de la Señal de Vox para el Reconocimiento del Habla en Ambientes Ruidosos", Ph.D. Dissertation, Dept. of Signal Theory and Communications, Polytechnical University of Catalonia, Spain, 1993.