S4a.10

A Finite States Markov Quantizer for Speech Coding

A. Falaschi(*) M. Giustiniani (**) P. Pierucci(**)

(*) La Sapienza University of Rome (**) IBM Rome Scientific Centre

ABSTRACT

This paper presents a low bit rate codec based on an Ergodic Hidden Markov Model. A 256 states autoregressive gaussian EHMM has been trained on speech uttered by 8 different speakers, by mean of the Baum Welch algorithm; initial estimates are obtained from vector quantization. The resulting EHMM is then utilized for a Viterbi decoding of incoming speech data. The state sequence obtained is frame sincronously encoded. The bit rate is gradually lowered by cutting off low probability transitions, and thus reducing the destination state encoding bit allocation requirements. The encoded spectra sequence is used, on the receiver side of the codec, for LPC synthesis. Global entropy and distortion measures for different bit rate are reported, and compared to vector quantization results. Informal listening tests have been performed by doing a comparison between the proposed method at various bit rate and the same material encoded by VQ.

I - INTRODUCTION

Hidden Markov Models major application has been found in the field of automatic speech recognition systems, where the HMM capability to represent both local and temporal features of speech has allowed very accurate signal modelling, and very high recognition rates.

Initially, HMM has been regarded only as a convenient mechanism to deal with speech recognition substitutions, insertion and deletion events [1]. Subsequently, extention of the HMM parameter estimation algorithm to the continuous observation density case [2] [3] has added a more physical meaning to the HMM of linguistic units. Interpretation of HMMs as a generation model of the speech process taken as whole is reported in the pioneering work of Poritz [ ], in which an Ergodic (fully connected) HMM (EHMM) of speech is defined as the combination of local models (described by Autoregessive gaussian densities) with the EHMM transition probabilities matrix.

Viterbi alignment of incoming speech against the EHMM states revealed the capability of the model to automatically ascribe both broad phonetic classes to the model states, and phonotactical constraints to the transition probability values.

Since then, the spectrum parametric representation related to the HMM states decoded sequence has been often utilized to check the correctness of HMM estimates of linguistic units [4] [5]. Some EHMM based speech analysis-synthesis schemes have also been derived [6] [7], where the decoded state spectral parameters sequence is used to resynthesise the original speech.

A major result of such techniques is the good perceptual quality of this synthesis method, with respect to an equal size memoryless vector quantization technique, due to the "help" the transition probabilities give to the decoding of frequently occurring event sequences.

In the 1989 ICASSP proceedings, [8] proposed a very low bit rate speech spectra encoding method, named 'A Phonetic Vocoder'. It encodes speech at two levels of abstraction: the first one is the linguistic labeling obtained by means of an HMM phonetic recognizer, and the second is the phonetic HMM inner state sequence alignment with the original speech. The peculiarity of the left-to-right HMM transition matrices used allowed a state sequence coding with a bit rate on the order of 100 bit/sec.

The present work demonstrates that the phonetic decoding stage is unnecessary for such a low bit rate achievement, which can be more simply obtained by means of the execution of a Viterbi alignment procedure on a Pruned Transitions EHMM (PTEHMM). In fact, experience has shown that most of the EHMM transition probability matrix elements estimated in [ ] are labeled with very low probability values, because they connect spectral events that do not often occur sequentially in real speech utterances. This suggests the possibility to still obtain a good synthetic speech quality after having pruned out many of the less probable transitions. A decreased number of outgoing transitions from a state requires a lower number of bits to encode the state sequence, thus reducing the net bit rate. This approach has a certain resemblance with the one proposed in [9] for the Pruned Finite States Vector Quantization (PFTVQ) encoding technique, where the allowable codeword sequences are pruned on the basis of a minimum-distortion criterion.

The temporal continuity constraints embedded in the loop transition probabilities of the EHMM has been used in the present work to obtain a further bit rate reduction, by adopting a variable length transition encoding technique similar to the one introduced in [8]. It consists in sending a flag bit set to zero if the decoded state for the present frame is not changed from the previous one; a flag bit set to one plus the destination state code is sent if the state has changed. In the Pruned Transitions EHMM case, the next state code is the index in the next state lookup table function, table that must be replied at the receiver side of the channel. The present paper explores the case in which each state has the same, fixed, number of outgoing transitions, although a state-dependent number of transitions case could be dealt with, by some minor changes at the receiver next state lookup table selection component.

The rest of the paper is organized as follows. Section II describes the realization of the EHMM, its main features, and the procedure adopted for the transition matrix pruning. Section III reports some of the PTEHMM features, e.g. the minimum number of transitions needed to retain the model ergodicity, some considerations about the resulting information source entropy, and related performances. Section IV reports a discussion about the bit rate achieved during the preliminary experiments. Section V contains some quality evaluation results for the proposed encoding technique, obtained by subjective listening test and by average distortion measures, compared to classical memoryless VQ coding. Finally some alternative state sequence encoding methods are proposed.

CH2847-2/90/0000-0205 $1.00 © 1990 IEEE

II - MODEL DEFINITION

As foreworded in the introduction, our encoding scheme relies on the definition of an EHMM of speech, consisting of a set of N states, a set of N autoregressive gaussian observation densities (one for each state), an N*N transition probabilities matrix, allowing every state pair sequence, and an initial probability vector, giving the a priori probability of being in each of the EHMM states at the beginning of the encoding process. The experiments here reported refers to a state cardinality of N = 256.

The model parameters are estimated by means of the Baum algorithm [2], on about eight minutes of speech composed by the utterance of six hundred different phonemically compact words [12], by six different speakers.

The speech is sampled at a frequency of 10 KHz; linear prediction analysis has been performed; frame length is of 320 samples, and each frame is shifted of 80 samples from the preceeding.

The initial EHMM parameters are computed by means of the binary-splitting version of the Lloyd-Max vector quantization algorithm [10], utilizing as distance measure the Likelihood Ratio distortion measure, in order to be consistent with the gain-independent autoregressive gaussian densities [11] [7]. utilized as EHMM observation densities.

The choice of LPC spectral representation is motivated by the fact that such local model is very well suited to represent either the multivariate continuous observation densities utilixed by Hidden Markov Modeling of speech, as well as the parameters needed for synthetic speech spectral reconstruction.

A set of prediction coefficients is associated to each state, as the synthesis filter control parameters. Moreover the prediction coefficients autocorrelation function is needed for the evaluation of the HMM densities value.

The EHMM transition matrix is initialized by means of a smoothed co-occurrence count statistics of the VQ labels collected for the same training data. Figure 1 reports, in a graphical form, the resulting transition probabilities after 4 cycles of the Baum reestimation algorithm.

The periodicity of the resulting transition probability estimates results from the binary splitting method adopted for the VQ initialization of the EHMM. As evident, the initial transitions matrix structure is mainly preserved, and the resulting spectral shapes do not reveal too many differences among the initial and the final values. This result is acceptable, being the data utilized for VQ initialization and Baum reestimation the same. As a preliminary experiment, the PTEHMM has been obtained from the EHMM simply by zeroing the lowest probability transitions departing from each states, and properly rescaling the remaining ones.

III - MODEL FEATURES

Before the process of transitions pruning, a check has been done looking for the minimum number of transitions per state which preserves model ergodicity, i.e. which allows all of the resulting PTEHMM states to be reached form each other state, thus still obtaining a fully connected EHMM. This has been done by multiplication of the pruned incidence matrix (derived from the transition one after substitution of each non-zero probability value with one) for itself for $\log_2 N - 1$ times, and checking for the absence of all-zero columns. This assures that at least one path exists between all the states, whose maximum length, (i.e. the number of states visited along the path) is N. The minimum number of outgoing transition for our 256 state PTEHMM amounts to seven. The same result was obtained in another earlier experiment where a single speaker EHMM with 128 states was used.

We fixed the minimum number of undeleted transitions in the maximally pruned case to eight; this number will be referred in the following as M.

Once the PTEHMM has been obtained, a given speech utterance can be encoded by the state sequence labels obtained by execution of the traditional full search Viterbi algorithm decoding method applied to the resulting EHMM of speech. This sequence is then straigthforward coded by the indexes of the crossed transitions along the decoded state sequence.

In the maximally pruned case and without any other coding mechanism, 3 bits/frame are needed to completely encode the original 256 source symbols. This bit (62.5%) coding saving deserves some more information theory considerations. Let's first compare the bit rate requirements for a memoryless VQ scheme, with the effective information rate needed to encode a Markov source. In the VQ case, $\log_2 N$ bit/symbol are needed, because the a priori symbols probabilities are unknown; a Markov source model allows us to define an absolute H^0 and conditional H^1 intrinsic entropies, represented by

$$(1)\quad H^0 = -\sum_{i=1}^{N} P_i \log_2 P_i$$

$$(2)\quad H^1 = -\sum_{j=1}^{N}\sum_{i=1}^{N} P_j P_{i/j} \log_2 P_{i/j}$$

where P_i and $P_{i/j}$ are respectively the absolute and conditional probability of state S^i. It is clear that the true message entropy evaluation theoretically requires knowledges of an infinite numerable order statistics, but yet a simple one-memory Markov model allows an adequate reduction of the source information rate, as given in the following table .

Table 1.

Information Source	H^0	H^1
VQ	7.52	2.22
FSMQ	7.40	1.02

Absolute (H^0) and conditional (H^1) entropies of the information sources.

Let us now consider the PTEHMM, for which only $\log_2 M$ bits have to be transmitted for each symbol. Formula (2) becomes :

$$(3)\quad H^1 = -\sum_{j=1}^{N}\sum_{i=1}^{M} P_j P_{i/j} \log_2 P_{i/j}$$

Here below the H^1 values are given for different pruning factors, together with the bit rate required for a straightforward encoding method.

Table 2.

M	$\log_2 M$	H^1
128	7	1.02
64	6	1.02
32	5	1.03
16	4	1.06
8	3	1.10

Conditional Entropy H^1 of PT-FSMQ information source. M represents the maximum number of allowed output transitions from each state.

As it can be seen, H^1 only sligthly grows as the number of output transitions M decreases. The above figures demonstrates that the performed transitions pruning technique does not cause substantial changes of the intrisic source entropy.

IV - NET BIT RATE

As discussed in [5], the time continuity properties of the speech spectral features permit to encode the decoded state sequence in a more efficient way than simply emitting the index sequence of the undertaken transitions. The method relies on the consideration that frequently occurring states exhibit greater loop probability, resulting in longer runs of the decoded sequence. This suggest the use of a variable rate encoding for the transition sequence, using one bit to flag state changes, plus $\log_2 M$ bits for the transition index in the case of an inter-state transition. Such a technique permits to reach the encoding rates given in table 3, which have been computed from 60 seconds of experimental data.

Table 3.

Information Source	bit / sec
VQ	544
FSMQ	305
PT-FSMQ 128	283
PT-FSMQ 64	260
PT-FSMQ 32	239
PT-FSMQ 16	220
PT-FSMQ 8	205

Encoding bit/rate for different information sources.

As evident, the variable rate encoding method permits a remarkable bit economy also for VQ encoder, giving a better approximation of the intrinsic conditional entropy of the underlying Markov source. The different state probabilities suggest that a further bit rate reduction could be achieved by variable length state coding, using a lower number of bits for the higher probability states, with a Huffman like coding technique.

V - PERFORMANCES EVALUATION

At first, let us visually examine the quality of the proposed coding technique. Figure 2 reports the sonogram of a speech utterance (the word "massimo",maximum), with its PTEHMM and VQ versions.

Some subjective and objective quality evaluation measure have then been performed on a speech data base consisting of 20 words (not belonging to the training list) uttered by 4 speakers (2 of which are in the training set and two are not). The objective results are given as the average distortion rate, computed by means of the log likelihood distortion measure, for various pruning rates. As it can be noted, there are not significant quality differences for pruning factors greater then 16 transitions per state(figure 5).

The distortion figures, even if obtained for a different prediction order, compared with those given by [ ] , seem to reveal a little advantage for the here discussed PTEHMM encoding method. The advantages are more relevant if we consider also the values given in table 3, which report the effective bit rate of the encoder.

Subjective quality assessment have been performed by listening test, where 5 listeners are requested to express a preference opinion when comparing synthetic speech obtained by PT-EHMM versus VQ encoded words. As can be seen from figure 4, PT-EHMM subjectively performs better than VQ. Lowering the number of output transitions (M) from each state of the PT-EHMM reduces the preference score, that results close, for an M value of eight, to that of the VQ.

CONCLUSIONS

In this paper we proposed a novel technique for very low bit speech signal coding. Experimental evaluation of the model has shown that it is competitive with classical VQ approach, in terms of bit rate requirements and perceptual quality. In our opinion this is mainly due to the EHMM statistical framework used. In fact the EHMM states tends to assume phonetic identities; the relevant spectras are estimated from contiguous segments of signal, and the probability transition matrix reflects the language dependent phonotactical constraints. Some improvements can be made to the proposed method. First of all a beam search [13] decoding technique can substitute the Viterbi algorithm. A continuous backtracing procedure [14] can be applied, removing the present strong time delay limitation of the encoder. Obviously dependencies of the search width on the encoder delay and on the resulting distortion rate should be analyzed. As a second issue, a variable number of state outgoing trasitions can be considered. The lower bound of 7 can not be general for all the states; in fact we observed that more stationary events are related to states which exibits a lower number of output transitions. This way a further bit rate reduction can be obtained at the expense of a slightly more complex receiver architecture. Some iterations of the Baum Welch estimation EHMM parameters can be performed periodically during the transition pruning phase, allowing a continuous adaptation of the model features. Finally an alternative transition pruning method can be proposed. It comes from the examination of (3), where it can be noted that each transition gived its own (different) contribution to the source model conditional entropy. As a consequence the transition pruning procedure can be executed iteratively, deleting at each time the transitions j^1, i^1 for which :

(4) $j^1, i^1 = argmin\{P_j P_{i/j} \log_2 P_{i/j}\}$

discarding this way the less informative connections. Periodic Baum re-estimation will still mantain the EHMM parameters updated.

REFERENCES

[1] L.R.Bahl, F.Jelinek,
Decoding for Channels with Insertion, Deletions, and Substitutions with Applications to Speech Recognition
IEEE Trans IT-21, N.4
July 1975

[2] L.E.Baum,
An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes
Inequalities, vol III, pp.1-8, 1972

[3] L.A.Liporace,
Maximum Likelihood Estimation for Multivariate Observations of Markov Sources *IEEE Trans.IT-28, N.5,*
September 1982

[4] A.B.Poritz,
Linear Predictive Hidden Markov Models and the Speech Signal *ICASSP Proceedings 82, p.1291, Paris* 1982

[5] K.Ganesan, M.Marlot, P.Methta,
An efficent algorithm for combining vector quantization and stochastic modeling for speaker-independent speech recogniton *Proc of ICASSP Tokyo* 1986

[6] G.R.Doddington, J.Picone, J.Godfrey,
The LPC Trace as an HMM development tool *JASA, Suppl.1, Vol.84,* Fall 1988

[7] E.P.Farges, M.A.Clement,
An analysis-synthesis Hidden Markov Model of Speech
Proc of ths ICASSP, New York 1988

[8] A.Falaschi, M.Giustiniani, M.Verola, **A Hidden Markov Model Approach to Speech Synthesis** *Proc. EUROSPEECH, Paris* 1989

[9] J.Picone, G.R.Doddington,
A Phonetic Vocoder,
Proc. of the ICASSP, Glasgow 1989

[10] B.H.Juang,
Design and Performance of Trellis Vector Quantizers for Speech Signals
IEEE Trans. ASSP-36, N.9,
September 1989

[11] A.Buzo, A.H.Gray,Jr., R.M.Gray, J.D.Markel,
Speech coding *IEEE Trans ASSP-28, pp.562-574,*
October 1980

[12] B.H.Juang, L.R.Rabiner,
Mixture autoregressive Hidden *IEEE Trans ASSP-33, n.6,*
December 1985

[13] A. Falaschi,
Automatic selection of pphonologically compact
to appear at the Proc. of First SFA Congress, April 1990,

[14] B. Lowerre, R. Reddy,
The Harpy Speech Understanding System
W.A. Lea ed., Trends in speech recognition, Prentice Hall, N.J.,
1980

[15] C.Scagliola,
Language Models and Search Algorithms for Real *Int. Jou. of Man Machine Studies, n.22,*
1985

[16] B. Juang,
Design and Performance of Trellis Vector Quantizers for Speech Signals
Proc. of ICASSP
September 1988

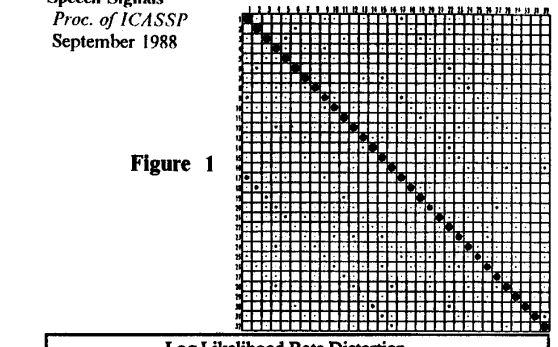

Figure 1

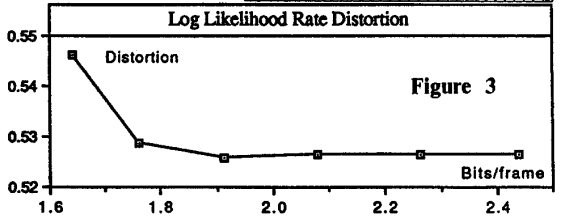

Figure 3

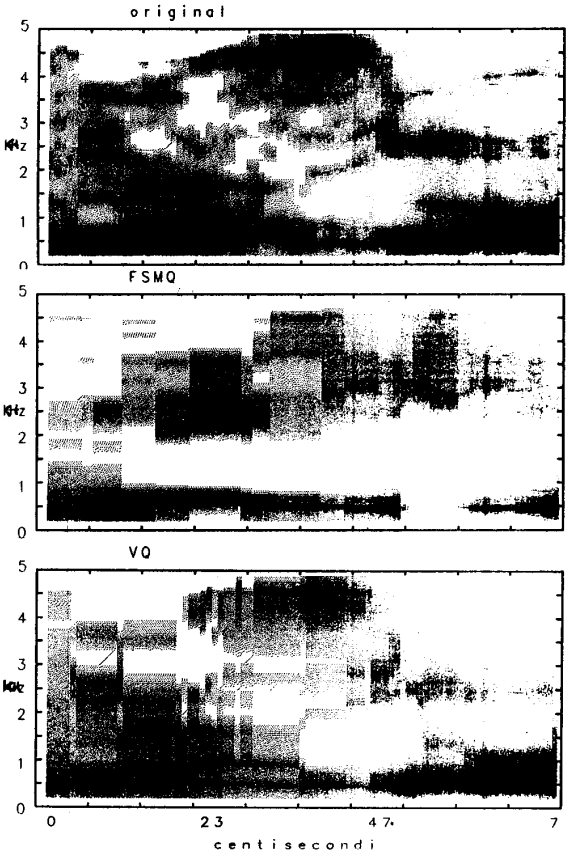

Figure 2

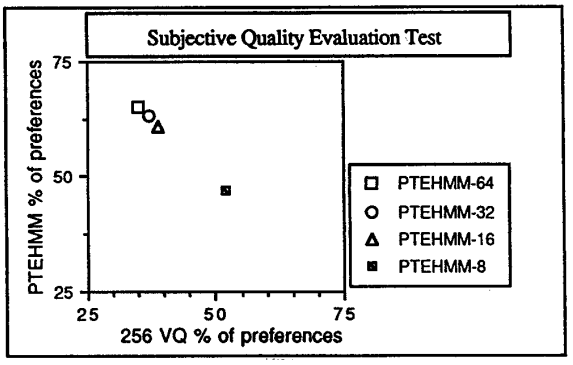

Figure 4