HMM-BASED METHODS FOR CHANNEL ERROR MITIGATION IN DISTRIBUTED SPEECH RECOGNITION

Antonio M. Peinado, Victoria Sánchez, José L. Pérez-Córdoba, José C. Segura, Antonio J. Rubio

Departmento de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071-Granada (Spain)
amp@ugr.es

ABSTRACT

Distributed Speech Recognition involves the development of techniques to mitigate the degradations that the transmission channel introduces in the speech features. This work proposes an HMM framework from which different mitigation techniques oriented to bursty channels can be derived. In particular, two MMSE-based and a new Viterbi-based mitigation procedures are derived under this framework. Several implementation issues such as the channel SNR estimation or the application of hard decision on the received signal vectors are dealt with. Also, different boundary conditions suitable for the speech recognition application are studied for the different mitigation procedures. The experimental results show that the HMM-based techniques can effectively mitigate channel errors, even in very poor channel conditions.

1. INTRODUCTION

The subject of Distributed Speech Recognition (DSR) has recently arisen allowing an efficient translation of the Automatic Speech Recognition technologies to mobile and IP network applications. DSR is based on the utilization of a local front-end and a remote back-end. This approach has clear advantages since voice features are not affected by the speech coder, language portability is facilitated and a simple front-end is utilized. An ETSI standard for DSR (ETSI-ES-201-108 v1.1.2) [1] has already been elaborated by the Aurora working group.

DSR systems can be affected by several degradation sources due to the acoustic environment and the digital channel. Although the processing of all these degradations can be carried out at the receiver, the distributed nature of the recognition process in DSR makes it more convenient the treatment and reduction of acoustic degradations in the local front-end, where the signal (speech plus noise) is fully available. Thus, at the remote back-end, only the errors introduced by the digital channel must be mitigated.

Several authors have already dealt with the problem of channel error mitigation. In the case of wireless channels, exponential feature weighting has been recently proposed [2]. Also, techniques based on Minimum Mean Square Error (MMSE) estimation [3, 4] have been adapted to DSR [4]. In this case, mitigation is performed before recognition at the back-end, modeling speech production as a first order Markov process in order to introduce temporal correlations in the MMSE-based mitigation. The present work makes a generalization of this idea by using an HMM formalism to perform mitigation. One of the advantages of this approach is that the large theoretical background associated with HMMs can be exploited. In particular, we present the MMSE-based mitigation procedures of our previous work [4] under this framework and propose a new procedure based on a Viterbi decoding. We also explore several aspects as the estimation of the channel SNR required for mitigation, the use of hard decision on the received signal vectors and the selection of a set of boundary conditions adapted to the speech recognition application.

Work supported by the Spanish CICYT project TIC-2001-3323

2. AURORA FRAMEWORK

The front-end used in this work is the one proposed in the ETSI standard [1]. This front-end provides a 14-dimension feature vector containing 13 MFCCs (including the 0th order one) plus log-Energy. These features are grouped into pairs and quantized by means of seven Split Vector Quantizers (SVQ). All codebooks have a 64-center size (6 bits), except the one for MFCC-0 and log-Energy, which has 256 centers (8 bits). The bitstream is generated by grouping frames into pairs (88 bits) that are protected by a 4-bit CRC.

At the back-end, error bursts are detected by means of a CRC checking and a consistency test. The Aurora mitigation algorithm can be summarized as follows: once a burst, containing 2B frames, is detected, the first B frames are substituted by the last correct frame before the burst and the last B ones by the first correct frame after the burst. In the case of a burst at the beginning of the utterance, the first correct frame after the burst is repeated in the degraded frames. A similar solution is applied for corrupted data at the end of the utterance.

The recognizer is the one provided by Aurora and uses eleven 16-state continuous HMM word models, (plus silence and pause, that have 3 and 1 states, respectively), with 3 gaussians per state. The training and testing data are extracted from the Aurora-2 database. Training is performed with 8440 clean sentences and test is carried out over set *A* (4004 clean sentences distributed into 4 subsets).

3. TRANSMISSION AND CHANNEL MODEL

Figure 1 shows a block diagram of our transmission scheme. After the SVQ quantization, each feature pair is represented by a vector $\mathbf{c}$ ($\mathbf{c} \in \{\mathbf{c}^{(i)}; i = 0, \ldots, 2^M - 1\}$) ($M$=6,8 in this work) that, after bit mapping, is represented by a bit sequence $\mathbf{x} = (x(0), x(1), \ldots$ $\ldots, x(M-1))$ ($\mathbf{x} \in \{\mathbf{x}^{(i)}; i = 0, \ldots, 2^M - 1\}$), where each bit is assumed to be bipolar ($x(k) \in [-1, +1]$). This sequence is transmitted, after channel encoding, through a digital channel.

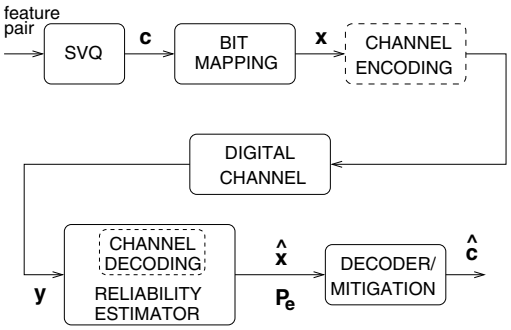

Fig. 1. *Transmission scheme for a feature pair.*

SNR_t (dB)	d (duration)	BER_t (%)
-3	657	10.90
-1	380	6.40
0	282	4.81
3	94	1.76
5	25	0.64

Table 1. *Correspondences of average SNR, burst duration (length in bits) and average Bit Error Rate.*

In the case of Aurora, the systematic CRC code has a very small error correction capability. Thus, we will work on the received information bits as if no channel encoding was used (except for error burst detection, as in Aurora).

As indicated above, the Aurora mitigation algorithm is suitable for bursty channels, and is tested in [5] under GSM error patterns EP1, EP2 and EP3. In this work, we present experimental results on these patterns, but, in order to experiment over a wider range of channel conditions, we also consider a simplified bursty channel model. In this channel, the received signal vector is obtained as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where the channel noise $\mathbf{n}$ is additive and obtained as the superposition of a background additive white gaussian noise (AWGN) (variance $N_g/2$) plus a sequence of AWGN noise bursts (variance $N_b/2 \gg N_g/2$) of fixed duration d (in number of bits), with a separation given by a Poisson variable of mean T_b [4]. The average variance of the channel noise is,

$$\frac{N_0}{2} = \frac{N_g}{2} + \frac{N_b}{2}\frac{d}{T_b} \tag{1}$$

and the average channel SNR can be computed as E_b/N_0 ($E_b = 1$). For our experiments, we consider $E_b/N_g = 6$ dB (BER=0.23 %), $E_b/N_b = -6$ dB (BER=24.59%), and $T_b = 1500$ bits. Thus, different values of the average SNR (E_b/N_0) have the meaning of different burst durations as it is detailed in table 1. Under these conditions, the EP3 pattern roughly corresponds to an average SNR between -1 and 0 dB.

4. DECODING AND MITIGATION USING HMMS

Since we are testing a bursty channel, a standard hard decoding is performed during error-free periods, and the proposed techniques are only applied during error bursts, replacing the Aurora mitigation algorithm. Thus, we consider a sequence of received observations $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$, where $\mathbf{y}_1$ and $\mathbf{y}_T$ are the last and first correctly received vectors before and after a burst, respectively. Isolated frame pairs with correct CRC are not allowed and are considered as included into the burst. We model each feature pair generation process by an ergodic HMM, where each state s_i represents an SVQ centroid $\mathbf{c}^{(i)}$ (or, equivalently, codeword $\mathbf{x}^{(i)}$). The transition probabilities a_{ij} can be obtained from an simple analysis of the training data. The observation probabilities $b_i(\mathbf{y}) = P(\mathbf{y}|\mathbf{x}^{(i)})$ can be computed from the instantaneous bit error probabilities $p_e(k)$ ($k = 0, \ldots, M-1$) (corresponding to the hard decoded bits $\hat{x}(k) = sign[y(k)]$) as (considering a memoryless channel),

$$b_i(\mathbf{y}) = C \prod_{k=0}^{M-1} P(x^{(i)}(k)|\hat{x}(k)) \tag{2}$$

where,

$$P(x^{(i)}(k)|\hat{x}(k)) = \begin{cases} 1 - p_e(k) & x^{(i)}(k) = \hat{x}(k) \\ p_e(k) & x^{(i)}(k) \neq \hat{x}(k) \end{cases} \tag{3}$$

In general, the computation of probabilities $p_e(k)$ depends on the transmission scheme. In the case of a fading channel with Gaussian noise and BPSK modulation, this probability can be obtained as [3],

$$p_e(k) = \frac{1}{1 + \exp(|L_c y(k)|)} \quad \text{with} \quad L_c = 4a\frac{E_b}{N_0} \tag{4}$$

where a is a fading factor ($a = 1$ in this work).

An MMSE estimation of the received parameter vector at time t (that considers the previous and subsequent received vectors) is obtained as,

$$\hat{\mathbf{c}}_t = E[\mathbf{c}_t|Y] = \sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} \gamma_t(i) \quad (1 < t < T) \tag{5}$$

with

$$\gamma_t(i) = P(\mathbf{x}_t^{(i)}|Y) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=0}^{2^M-1} \alpha_t(j)\beta_t(j)}$$

$$\alpha_t(i) = P(\mathbf{x}_t = \mathbf{x}^{(i)}|\mathbf{y}_1, \ldots, \mathbf{y}_t)$$

$$\beta_t(i) = P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\mathbf{x}_t = \mathbf{x}^{(i)})$$

where $\alpha_t(i)$ and $\beta_t(i)$ are the forward and backward conditional probabilities, respectively. These conditional probabilities can be computed from forward and backward recursions, respectively, as indicated in [4]. We will refer to this approach as FBMMSE (Forward Backward MMSE) estimation.

Although the forward probabilities could be computed while receiving data, the FBMMSE estimation necessarily introduces a delay in the decoding/estimation process, since it is required to wait until the end of the error burst to obtain the backward probabilities. The delay is the same as the one introduced by the Aurora mitigation. A simplified alternative consists in the use of only the forward probabilities in the expected value computation [3, 4]. Although clearly suboptimal, since only past received vectors are used, this approach has the advantage of a reduced computational cost in addition to delay suppression. This approach is referred to as FMMSE (Forward MMSE) estimation.

The proposed HMM framework leads us to the possibility of implementing a new mitigation procedure based on the use of the Viterbi algorithm. In this case, the estimated sequence of feature

pair vectors corresponds to the optimal state sequence $\hat{Q}$, that maximizes $P(Q, Y)$, where $Q = (q_1, \ldots, q_T)$ represents a given state sequence. The Viterbi algorithm implies the recursive computation of the joint probability $\delta_t(i)$ (later referred to as delta probability) of the best state sequence ending at state s_i at time t and the observation subsequence $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ [6]. Since we are obtaining a *Maximum A Posteriori* estimation of the best state sequence, we will refer to this approach as MAP estimation. This third approach presents several attractive points such as efficient implementations based on trellis and log-probabilities or the possibility of integration with the Viterbi word decoding applied for recognition, although it introduces the same delay as FBMMSE or Aurora.

5. IMPLEMENTATION OF HMM-BASED MITIGATION

5.1. About SNR estimation

We have previously shown that the computation of the observation probabilities from reliability measures requires an estimation of the channel SNR E_b/N_0. This can be a simple task in the case of channels for which the degradation is homogeneously distributed, such as AWGN or Rayleigh channels. In the case of a bursty channel, the noise is concentrated into bursts. In particular, the bursty channel model previously introduced can be viewed as AWGN with an SNR variable in time. In this situation, two different problems should be addressed. First, how to reliably estimate the average SNR from (1) for such non-stationary noise, for which error bursts can even have different durations and degradation levels in a real situation. Also, although the estimation of the average channel SNR could be carried out, this estimated value would be clearly high during bursts, where mitigation is applied.

In order to put some light about these questions, we have carried out an experiment consisting in testing the DSR performance (as Word Accuracy, WAcc) when an incorrectly estimated SNR is utilized for the probability computation of equation (4), transmitting through an AWGN channel. The applied mitigation is the one based on FMMSE estimation (FBMMSE is not suitable for this study). The results are depicted in figure 2. Each plot is obtained using a fixed value (-4, -2, 0, 2 and 4 dB) as estimated SNR for the computation of $p_e(k)$, and testing each for channel SNRs from -4 to 6 dB. Points labeled as "SVQ" (WAcc=99.04 %) and "Base" (WAcc=99.02 %) correspond to the use of error free quantized and original features, respectively.

These plots clearly show that a high value for the estimated SNR is more detrimental than a lower one. In the case of the -4dB plot, a very small degradation is obtained for high channel SNRs, while, on the contrary, an estimated SNR of 4dB rapidly decreases the performance when the channel conditions get worse. Thus, the utilization of a low fixed value for the SNR can provide quite an acceptable performance. A fixed SNR value equal to -2 dB is chosen for error probability computation in the rest of this work, thus avoiding the problem of SNR estimation.

5.2. A hard decision approach

When implementing a decoding stage, it is clear that the use of hard decision on the received signal vector implies a much simpler device. One possible solution [4], that allows the application of the proposed HMM-based methods along with hard decision, is to compute the observation probabilities as a function of the Hamming distance between the hard decoded bit sequence $\hat{\mathbf{x}}$ and

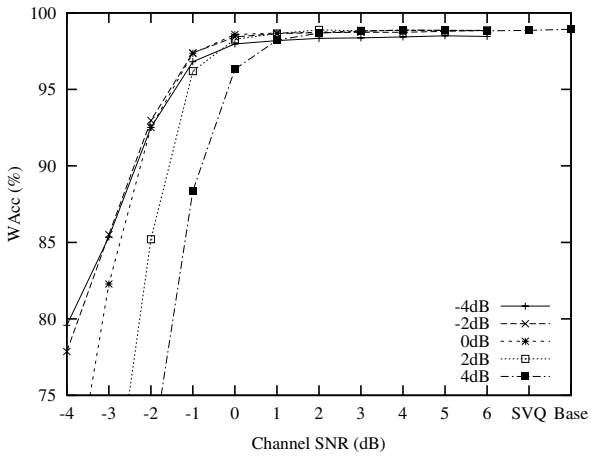

Fig. 2. *Word Accuracy (WAcc) results over an AWGN channel with reliabilities estimated with a fixed SNR value.*

codewords $\mathbf{x}^{(i)}$ as,

$$b_i(\mathbf{y}) = (1 - p_e)^{M - d(\hat{\mathbf{x}}, \mathbf{x}^{(i)})} + p_e^{d(\hat{\mathbf{x}}, \mathbf{x}^{(i)})} \tag{6}$$

where $d(\mathbf{a}, \mathbf{b})$ is the Hamming distance between codewords $\mathbf{a}$ and $\mathbf{b}$, and p_e is the average BER (we use the AWGN channel BER). Thus, only an estimation of the channel SNR is required to obtain p_e and, therefore, $b_i(\mathbf{y})$. Our previous work [4] shows that this hard decision approach provides WAcc results similar to those of soft decision in the case of a bursty channel using a fixed SNR value of -2 dB when applying an MMSE-based mitigation.

5.3. About boundary conditions

In our previous work [4], the forward and backward procedures are initialized in a standard way, by means of the following expressions (later referred as boundary conditions A),

$$\alpha_1(i) = P_i b_i(\mathbf{y}_1)/K_1 \quad (0 \leq i < 2^M) \tag{7}$$

$$\beta_T(i) = 1/2^M \quad (0 \leq i < 2^M) \tag{8}$$

where P_i is the *a priori* probability of source symbol i and K_1 is a normalization factor. It must be noted that we have normalized the backward probabilities to the number of states 2^M. This fact only implies the introduction of a constant that does not affect the results and that will ease the comparison with the modified boundary conditions introduced later in this section. The standard Viterbi algorithm can be initiated identically to the forward procedure and terminated by choosing the final state $\hat{q}_T$ as the one which maximizes probabilities $\delta_T(i)$.

In a speech recognition application, the beginning and end of the utterance to be recognized usually have specific characteristics. For example, silence or low energy segments can often appear. This knowledge can be integrated in the model topology by including starting and ending null states (I and F) that can be used when an error burst corrupts data at the beginning or end of an utterance. Otherwise, for bursts fully included in the utterance, it can be taken into account that at times $t = 1$ and $t = T$ the received data is correct, in order to initiate and terminate the mitigation procedures (in a similar manner as Aurora mitigation does). These considerations involve several changes in the strategies of initialization and termination of the mentioned procedures

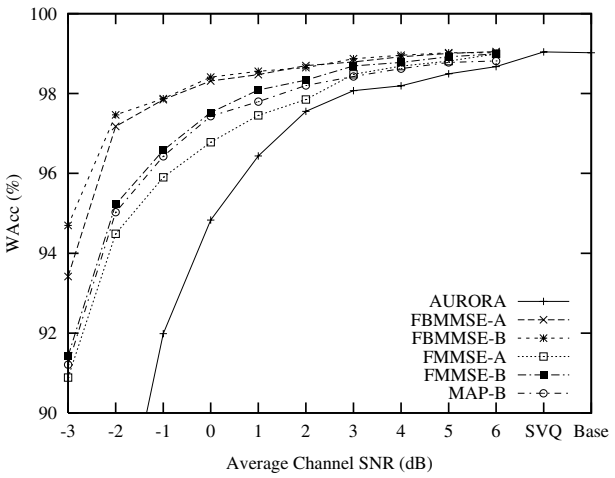

Fig. 3. *DSR over a bursty channel for conditions A and B.*

(boundary conditions B):

1) Initialization of forward and delta probabilities: in the case of detecting an error burst at the beginning of the utterance, the a priori probabilities P_i of each state are substituted by the probabilities of transition from state I. Otherwise, in the case of a burst fully contained in the utterance, it is taken into account that $\mathbf{y}_1$ is the last correct vector received before the burst, so $b_i(\mathbf{y}_1)$ is zero for all states s_i except the one corresponding to the hard decoded codeword $\hat{\mathbf{x}}_1$.

2) Initialization of the backward probabilities: in the case of a burst at the end of the utterance, the uniform initialization of (8) is replaced by assigning to each $\beta_T(i)$ the probability of transition to state F. Otherwise, the uniform initialization is maintained but considering that $b_i(\mathbf{y}_T)$ is zero for all states s_i except the one corresponding to the hard decoded codeword $\hat{\mathbf{x}}_T$.

3) Termination and backtracking of the Viterbi algorithm: in the case of a burst at the end of the utterance, an additional transition to state F is considered. Otherwise, the backtracking is initiated in the state corresponding to the hard decoded codeword $\hat{\mathbf{x}}_T$.

6. EXPERIMENTAL RESULTS

Figure 3 shows the performance of the different proposed techniques and the effect of boundary conditions A and B. First, it can be observed that FBMMSE is clearly superior to FMMSE or MAP, as it could be expected. A second aspect that is revealed is that, in general, conditions B provide better performance than A. This improvement is drastic in the case of the MAP estimation, for which boundary conditions A provides even worse results than the Aurora mitigation (the MAP results for conditions A do not appear in figure 3). This is a consequence of the introduced approximations (fixed SNR value and hard decision), for which MAP is clearly more sensitive than the MMSE estimations. However, the MAP results reach those of FMMSE with conditions B, showing that the MAP approach can provide a good performance with an appropriate set of boundary conditions.

We have also tested the FBMMSE, FMMSE and MAP techniques over the EP GSM error patterns, using the boundary condition set B. The results are shown in table 2. When the channel conditions correspond to patterns EP1 or EP2, the degradation is negligible for all the tested techniques. However, when the EP3 pattern is applied, the Aurora mitigation introduces more than 5 % of word accuracy reduction, while, on the opposite side, the FBMMSE technique obtains less than half a point of degradation. Also, the FMMSE and MAP techniques show quite a good behavior, since both introduce a degradation of less than 1 % of WAcc with respect to FBMMSE, providing the MAP procedure the second best results.

Wacc (%)	EP1 BER$\simeq$0%	EP2 BER=1.76%	EP3 BER=3.48%
AURORA	99.04	98.94	93.40
FBMMSE	99.04	99.01	98.66
FMMSE	99.04	99.01	97.95
MAP	99.04	99.01	98.13

Table 2. *Performance of the Aurora and HMM-based mitigation methods over GSM Error Patterns EP1, EP2 and EP3.*

7. CONCLUSIONS

In this paper we have proposed an HMM framework from which two MMSE-based channel error mitigation schemes can be derived for bursty channels on a DSR application. Also, a MAP mitigation procedure based on Viterbi decoding has been proposed. Several implementation aspects of these mitigation procedures have been studied. Thus, we have shown that it is possible to obtain a good performance without the need of SNR estimation or soft decision, utilizing a fixed SNR value and hard decision, respectively. We have also shown that an appropriate selection of the boundary conditions improves the performance, specially making the MAP option quite competitive for real implementations.

Acknowledgement: we would like to thank David Pearce (from Motorola Labs) for providing us with the GSM EP patterns.

8. REFERENCES

[1] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms", April 2000.

[2] A. Potamianos, V. Weerackody: "Soft-Feature Decoding for Speech Recognition over Wireless Channels". *Proceedings of ICASSP-2001*, May 2001.

[3] T. Fingscheidt, P. Vary: "Softbit Speech Decoding: A New Approach to Error Concealment". *IEEE Trans. on Speech and Audio Processing*, pp. 240-251, vol. 9, no. 3, March 2001.

[4] A. Peinado, V. Sánchez, J. Segura, J. Pérez-Córdoba: "MMSE-Based Channel Error Mitigation for Distributed Speech Recognition". *Proc. of EUROSPEECH-2001*, pp. 2707-10, Sept. 2001.

[5] D.Pearce: "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends". *AVIOS 2000: The Speech Applications Conference*, San Jose (USA), 2000.

[6] L. Rabiner, B. Juang: "Fundamentals of Speech Recognition". *Prentice-Hall*, 1993.