MMSE-Based Channel Error Mitigation for Distributed Speech Recognition

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

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

Abstract

Recently, the first version of an ETSI standard for Distributed Speech Recognition has been proposed. The main benefit of this approach is the possibility of maintaining a high recognition performance when accessing remote information systems. The use of a digital channel for transmission of the encoded speech parameters implies the introduction of several channel distortions. Our paper deals with the mitigation of such distortions. We study the application of MMSE estimation to this problem and propose a new MMSE procedure that obtains the probabilities needed for MMSE from a forward-backward algorithm. We show that MMSE estimation obtains better performance than the mitigation algorithm described in the ETSI standard under different channel conditions.

1. Introduction

Very recently, the problem of recognizing speech transmitted over digital channels has been addressed and an ETSI standard has been elaborated (ETSI-ES-201-108 [1]). The AURORA working group was the responsible for developing this first standard and a Distributed Speech Recognition (DSR) approach, that is, a local front-end and a remote back-end, was adopted. There are clear advantages in this approach: voice features are not affected by the speech coder, more robustness against channel errors, and access from different networks with a guaranteed performance.

An important issue being currently addressed is robustness against adverse environments (in which the front-end of a DSR system must operate). Also, robustness against transmission channel errors must be taken into account. This is not exclusively a channel coding problem. During the last years, several error mitigation (or concealment) techniques, that provide an improved decoding, have been studied for speech or image coding [2] [3]. These techniques usually exploit some kind of knowledge about the encoded parameters which is embedded in a soft decoding scheme. In the case of DSR, we find that the encoded parameters (MFCCs in the current version of the standard) differ from those normally utilized in speech coding. Moreover, the goal of DSR is completely different from subjective vision or hearing, since at the back-end we find an automatic speech recognition system. Therefore, the development of specific mitigation algorithms for DSR is clearly justified. The ETSI DSR standard already includes a basic mitigation algorithm that has been shown quite effective for medium and good quality channels on TETRA and GSM environments [4]. Error mitigation can be also interesting not only for DSR, but also for other applications such as speech reconstruction from the transmitted DSR speech features.

In this paper, we address the problem of mitigating channel errors, studying the performance of mitigation algorithms based on an MMSE (Minimum Mean Square Error) philosophy. In particular, we propose a new MMSE mitigation algorithm that utilizes correct frames received before and after the frame being estimated. The different proposed techniques are developed using the AURORA ETSI standard front-end, although they could be straightforwardly extended to other encoding schemes. The proposed mitigation algorithms affect only to the decoding stage of the ETSI standard. For the sake of simplicity, we will assume a BPSK modulation and test two different data channels (AWGN and bursty). The recognition experiments are performed on the Aurora-2 speech database.

The paper is organized as follows. First, we briefly summarize the ETSI DSR standard and its error mitigation algorithm. Sections 3 and 4 are devoted to the study of several mitigation techniques over AWGN and bursty channels, respectively. Finally, the conclusions of this work are summarized.

2. Revision of the DSR ETSI standard and Aurora framework

The standard ETSI ES 201 108 (v1.1.2) [1] describes the speech processing, transmission and quality aspects of a DSR system. Although it allows 3 different sampling frequencies (8, 11 and 16 KHz), we will only use to 8 KHz, since this is the one Aurora-2 uses. Frames are 25 ms long and shifted 10 ms. Each frame is represented by a 14 dimension feature vector containing 13 MFCCs (including the 0th order one) plus log-Energy. These features are quantized using a Split Vector Quantizer (SVQ) that groups them into pairs (MFCCs 1 and 2, MFCCs 3 and 4, ..., MFCC 0 and log-Energy). Each pair has its own codebook that is generated utilizing a weighted distance measure. All codebooks have a 64-center size (6 bits), except the one for MFCC-0 and log-Energy, that has 256 centers (8 bits).

The bitstream is organized into a sequence of multiframes. Each multiframe contains a 2-octet synchronization sequence, a 4-octet header (containing different informations and a multiframe counter), and a 138-octet frame packet stream which contains 24 frames grouped into pairs encoded with 88 information bits followed by a 4-bit CRC.

After decoding, an error mitigation algorithm is applied. There are two tests for error detection: a CRC checking and a data consistency test. This last test tries to determine whether the frames in a frame pair have a minimal continuity. When an incorrect CRC is detected, the corresponding frame pair is classified as errored. Besides, if the previous frame pair is "inconsistent" is also labelled as errored. From this point on, all frame pairs are classified as errored until one is received that

passes the CRC and consistency tests. In this way, we have an effective procedure for detecting error bursts. 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.

This standard will be compared with the different techniques introduced along the paper under different channel conditions. It must be observed that our work is exclusively concerned with error mitigation on the feature packet stream and that headers are not used due to several reasons. First, the standard document does not specify how to decode them. Also, they are not necessary to carry out our experiments (assuming that the order of the received multiframes is perfectly known). Besides, it must be considered that a reliable decoding can be performed on them since the 16 header information bits are protected with other 16 parity bits.

The Aurora-2 database is based on the TI-Digits database (connected digits) decimated to 8 KHz. The recognizer is the one provided by Aurora and uses eleven 16-state continuous word HMM models (except silence and pause, that have 3 and 1 states, respectively), with 6 gaussians per state. Training is performed with 8440 clean sentences and test is carried out over set *A* (4004 clean sentences distributed into 4 subsets).

3. Applying MMSE to DSR decoding

Let us consider a quantized parameter 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 $\mathbf{x} = (x(0), x(1), \ldots, x(M-1))$ ($\mathbf{x} \in \{\mathbf{x}^{(i)}; i = 0, \ldots, 2^M - 1\}$), where each bit is assumed to be bipolar ($x(k) \in [-1, +1]$). This sequence is transmitted (typically, after some type of channel encoding) through a digital channel, which can be degraded by noise and fading. At the receiver, an MMSE estimation $\hat{\mathbf{c}}$ of the encoded parameter $\mathbf{c}$ is obtained from the received signal vector $\mathbf{y}$ as [3],

$$\hat{\mathbf{c}} = E[\mathbf{c}|\mathbf{y}] = \sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} P(\mathbf{x}^{(i)}|\mathbf{y}) = \frac{\sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} P(\mathbf{y}|\mathbf{x}^{(i)}) P_i}{\sum_{j=0}^{2^M-1} P(\mathbf{y}|\mathbf{x}^{(j)}) P_j} \tag{1}$$

where P_i is the *a priori* probability of symbol i. This estimation can be applied to the SVQ vectors of a DSR system, taking into account that seven different SVQ codebooks must be managed. An important point is the estimation of probabilities $P(\mathbf{y}|\mathbf{x}^{(i)})$. Assuming that soft decisions are applied to the channel output $\mathbf{y}$, these probabilities could be obtained, for example, by applying the SOVA algorithm [5] to the 92 bits of a frame pair, using a trellis as the one described in [6]. For each decoded bit $\hat{x}(k)$, SOVA provides a reliability value $|L(k)|$ that allows the computation of the instantaneous bit error probability as,

$$p_e(k) = \frac{1}{1 + exp(|L(k)|)} \tag{2}$$

Therefore, the probability that bit $x^{(i)}(k)$ was transmitted given that $\hat{x}(k)$ has been received is,

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

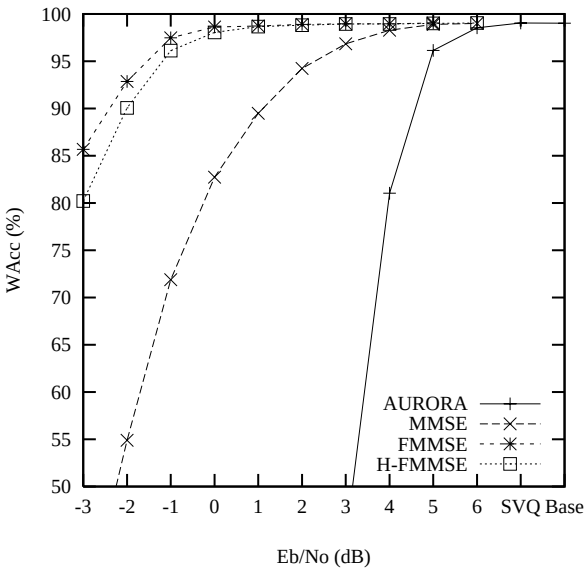

Figure 1: *DSR over an AWGN channel.*

Considering a memoryless channel and that $P(\mathbf{y}|\mathbf{x}^{(i)}) = P(\hat{\mathbf{x}}|\mathbf{x}^{(i)})$, it is finally obtained that,

$$P(\mathbf{y}|\mathbf{x}^{(i)}) = C \prod_{k=0}^{M-1} P(x^{(i)}(k)|\hat{x}(k)) \tag{4}$$

where C is a constant independent of subindex i.

The recognition performance (Word Accuracy) results of such a decoding procedure over an AWGN channel are depicted in figure 1 (MMSE plot). The performance of the Aurora mitigation algorithm for the same channel is also shown. Although Aurora is clearly inferior to the MMSE estimation, it is difficult to establish any comparison, since the Aurora procedure is not optimized for this type of channel, but for bursty channels. As references, the word accuracies of the recognition system without channel errors are also depicted at SNR points "Base" (original features) and "SVQ" (quantized features). It is assumed that the channel SNR (E_b/N_0) can be reliably estimated in order to obtain the bit error probabilities $p_e(k)$.

The decoding method described above can be further improved if the previously received vectors are considered in the MMSE estimation as in [3]. In order to do so, the MMSE estimation (at time t) must be modified as,

$$\hat{\mathbf{c}}_t = E[\mathbf{c}|\mathbf{y}_1, \ldots, \mathbf{y}_t] = \sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} \alpha_t(i) \tag{5}$$

where

$$\alpha_t(i) = P(\mathbf{x}_t = \mathbf{x}^{(i)}|\mathbf{y}_1, \ldots, \mathbf{y}_t) \tag{6}$$

is the (a posteriori) probability of receiving bit sequence $\mathbf{x}^{(i)}$ at time t given that $\mathbf{y}_t$ is received at time t and that vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$ have been previously received. This probability can be computed by modeling speech as a first order Markov source and by means of the following forward recursion:

1. Initialization: $t = 1$

$$\alpha_1(i) = P_i P(\mathbf{y}_1|\mathbf{x}^{(i)})/K_1 \tag{7}$$

2. Recursion: $2 \leq t \leq T$

$$\alpha_t(i) = \left[\sum_{j=0}^{2^M-1} \alpha_{t-1}(j) a_{ji} \right] P(\mathbf{y}_t|\mathbf{x}^{(i)})/K_t \quad (8)$$

where K_t is the normalization factor at each time t, and a_{ji} is the transition probability from source symbol j to i.

The results of this concealment technique are also depicted in figure 1, labelled as FMMSE (Forward MMSE). It can be observed that at a channel SNR of -3 dB, the DSR system still has a reasonably behavior (more than 80% of recognition accuracy), meanwhile a simple MMSE estimation is severely degraded.

3.1. A hard decision variant

Both, the simple MMSE and the FMMSE estimations are based on soft decisions on the received bits that allow the estimation of instantaneous bit error probabilities. But they can also be applied in the case of hard decisions. Since the channel code is systematic, the information bits are directly available from the input bitstream performing hard decision, and, then, it is possible to assign them the average error probability of an AWGN channel, that is,

$$p_e(k) = \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0}}\right) \quad (9)$$

In this case, the reliability of the different possible SVQ vectors $\mathbf{c}^{(i)}$ is related to the hamming distance of the corresponding code $\mathbf{x}^{(i)}$ to the code of the optimal (hard-decided) vector $\hat{x}$. Utilizing estimation (9) in (3) (instead of (2)), it is possible to perform an FMMSE estimation using the same expressions detailed above. The results for an AWGN channel are also depicted in figure 1 (labelled as H-FMMSE, Hard decision FMMSE). As it could be expected, the performance is inferior to that of soft decision using source time correlation although meaningfully better than a simple MMSE estimation.

4. Bursty channels and Forward-Backward MMSE

Very often, transmission systems must work over channels in which errors are grouped into bursts. This fact must be taken into account when designing mitigation algorithms. This is clearly the case of the Aurora mitigation algorithm, which in [4] is tested with 3 different GSM bit error patterns (EP1, EP2 and EP3) representing 3 different channel conditions (from acceptable to very poor quality). The FMMSE estimation can also be easily adapted to bursty channels. As it can be seen in Aurora, a key point is to detect the beginning and end of bursts. In our MMSE-based algorithms, it is considered that a burst starts when an erroneous CRC is received, and finishes when at least two consecutive correct CRCs are received. Once the burst end is detected, the mitigation procedure is initialized with the last correct frame received before the burst and it is performed during all the burst. While no bursts are received, a standard hard-decision decoding is performed.

We consider a simplified bursty channel model that introduces an additive noise consisting of a background AWGN noise of variance $N_g/2$ plus a sequence of bursts of fixed duration d with a separation given by a Poisson variable of mean T_b [7]. Inside a burst, the noise is also gaussian distributed with variance $N_b/2$. Of course, it is expected that $N_b \gg N_g$. The

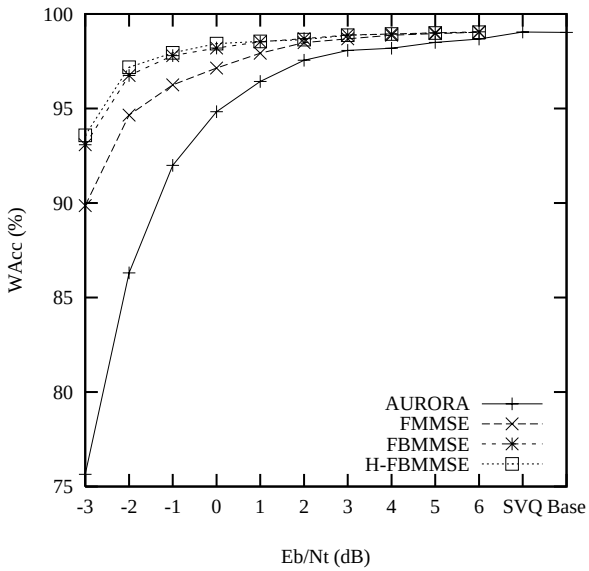

Figure 2: *DSR over a bursty channel.*

average energy of the channel noise can be computed as,

$$\frac{N_t}{2} = \frac{N_g}{2} + \frac{N_b}{2}\frac{d}{T_b} \quad (10)$$

For our experiments, we consider $SNR_g = 6$ dB (BER=0.23%), $SNR_b = -6$ dB (BER=24.59%), and $T_b = 1500$ bits. Figure 2 shows the performance of Aurora and the modified FMMSE techniques versus the total SNR (E_b/N_t). It must be pointed out that the SNR values from -3 dB to 5 dB have the meaning of different burst durations (from 657 to 25 bits, respectively). In order to have a point of reference, the EP3 condition can be roughly considered as equivalent to this bursty channel with a total SNR between -1 and 0 dB (with the proposed channel parameters). It must be taken into account that the modified FMMSE technique also requires an estimation of the channel SNR in order to compute bit error probabilities. This can be a difficult task for a bursty channel (the total SNR does not correctly describe the amount of noise at each time) and, besides, it is not the goal of this work. Therefore, we have utilized a fixed SNR (-2 dB), since preliminary experiments have shown that it is much better to provide SNR values closer to worst channel condition. This SNR value is utilized in all MMSE-based techniques along the rest of this work. Although it is clear that Aurora reaches a much better performance under this type of channel than with an AWGN channel, the FMMSE mitigation still provides better results.

In spite of the fact that Aurora still yields the worst result, it has an interesting property: it utilizes the correctly received frames that delimit a burst to rebuild the degraded frames. This idea can be translated to the MMSE estimation if it is carried out as,

$$\hat{\mathbf{c}}_t = E[\mathbf{c}|Y] = \sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} \gamma_t(i) \quad (1 < t < T) \quad (11)$$

with

$$\gamma_t(i) = P(\mathbf{x}_t^{(i)}|Y) \quad (12)$$

	EP1 BER≃0%	EP2 BER=1.76%	EP3 BER=3.48%
AURORA	99.04	98.94	93.40
H-FBMMSE	99.04	99.02	98.58

Table 1: *Performance of Aurora and H-FBMMSE over GSM Error Patterns 1,2 and 3.*

where $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ ($\mathbf{y}_1$ and $\mathbf{y}_T$ are the last and first correctly received vectors before and after the considered burst, respectively). The a posteriori probabilities $\gamma_t(i)$ can be obtained as,

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=0}^{2^M-1} \alpha_t(j)\beta_t(j)} \qquad (1 < t < T) \tag{13}$$

where,

$$\beta_t(j) = P(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T | \mathbf{x}_t = \mathbf{x}^{(i)}) \tag{14}$$

These probabilities can be obtained by means of the following backward recursion:

1. Initialization: $t = T$

$$\beta_T(i) = 1 \tag{15}$$

2. Recursion: $t = T-1, T-2, \ldots, 2$

$$\beta_t(i) = \sum_{j=0}^{2^M-1} a_{ij} P(\mathbf{y}_{t+1} | \mathbf{x}^{(j)}) \beta_{t+1}(j) \tag{16}$$

In the same way as with the forward probabilities, it is convenient to apply at each step in the recursion a normalization factor in order to avoid underflows.

This proposed Forward-Backward MMSE estimation (FB-MMSE) is also depicted in figure 2, and provides the best results obtained so far. A hard decision version (H-FBMMSE) is also depicted. In this case, hard decision does not imply any performance degradation of Forward-Backward MMSE. This fact can be due to several reasons as the power of the Forward-Backward MMSE technique and the short duration of the bursts (no more than 8 frames).

Since the H-FBMMSE technique makes hard decisions over the input bits, we also compare its performance with that of Aurora when applying the EP error masks. Word accuracy results are shown in table 1. The corresponding bit error rates are also shown. The H-FBMMSE technique also obtains better performance than Aurora (more than 5%) and only suffers a very slight degradation inferior to 0.5% (with respect to the baseline no-errored system) when applying the EP3 mask.

5. Summary

This paper is devoted to the application of MMSE to DSR channel error mitigation. This has been carried out by means of performing an MMSE estimation of the received parameters. In all cases, MMSE estimation has provided better results than those provided by the mitigation procedure of the Aurora standard. First, we have studied the behavior of a simple MMSE estimation, based on soft decisions on the channel outputs, over an AWGN channel. In order to obtain the a posteriori probabilities required by the MMSE estimation, it is necessary to have a reliability measure of the received information bits (in our case, by means of SOVA). A much larger improvement can be obtained on the same channel if the previously received signal vectors are considered in those a posteriori probabilities (FMMSE technique). This can be carried out by considering a first order Markov model of the speech source, and the a priori probabilities are obtained from a forward algorithm. The results make clear that a large amount of information can be extracted from time correlations of the utilized speech features (MFCCs and log-Energy).We also show that even in the case of performing hard decisions on the channel outputs, FMMSE provides an excellent result.

In order to test the behavior of MMSE under more realistic conditions, we have also checked the utilization of a bursty channel model. In this case, the MMSE technique requires some adaptation to this type of channel. First, we have implemented an easy mechanism to detect bursts. Besides, the idea of Aurora of utilizing the correct frames before and after the burst to carry out the mitigation is also incorporated to the MMSE estimation, which it is performed by means of a forward-backward procedure (FBMMSE technique). This new method obtains the best results over a bursty channel, and its hard decision version (H-FBMMSE) does not imply any loss of accuracy. Finally, this H-FBMMSE mitigation is compared with Aurora applying the GSM error patterns EP1, EP2 and EP3 to the coder output. The most noticeable difference is observed with the EP3 pattern, for which our technique obtains an word accuracy improvement of 5% over Aurora.

6. Acknowledgement

This paper has been supported by the Spanish CICYT project TIC-99-0583 "Speech Recognition in GSM environments".

7. References

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

[2] C.G. Gerlach: "A probabilistic framework for optimum speech extrapolation in digital mobile radio". *Proceedings of ICASSP-93*, vol. 2, pp. 419-422, 1993.

[3] T. Fingscheidt, P. Vary: "A Universal Approach to Bit Error Concealment". *Proceedings of ICASSP-97*, pp. 1667-70, April 1997.

[4] D.Pierce: "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.

[5] J. Hagenauer, P. Hoeher: "A Viterbi algorithm with soft-decision outputs and its application". *Proc. of GLOBECOM-89*, pp. 1680-86, 1989.

[6] J.K. Wolf: "Efficient Maximum Likehood Decoding of Linear Block Codes Using a Trellis". *IEEE Trans. on Information Theory*, vol. 24, no. 1, pp. 76-80, Jan. 1978.

[7] W.J. Ebel, W.H. Tranter: "The Performance of Reed-Solomon Codes on a Bursty-Noise Channel". *IEEE Trans. on Communication*, vol. 43, no. 2/3/4, pp. 298-306, February-April 1995.