Band-Pass Filtering of the Time Sequences of Spectral Parameters for Robust Wireless Speech Recognition

J. Vicente-Peña, A. Gallardo-Antolín, C. Peláez-Moreno, F. Díaz-de-María

Dpto. de Teoría de la Señal y Comunicaciones
EPS-Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911-Leganés (Madrid), SPAIN
Phone: +34 91 624 9170
Fax: +34 91 624 8749

Abstract

In this paper we address the problem of automatic speech recognition when wireless speech communication systems are involved. In this context, three main sources of distortion should be considered: acoustic environment, speech coding and transmission errors. Whilst the first one has already received a lot of attention, the last two deserve further investigation in our opinion. We have found out that band-pass filtering of the recognition features improves ASR performance when distortions due to these particular communication systems are present. Furthermore, we have evaluated two alternative configurations at different Bit Error Rates (BER) typical of these channels: band-pass filtering the LP-MFCC parameters or a modification of the RASTA-PLP using a sharper low-pass section perform consistently better than LP-MFCC and RASTA-PLP, respectively.

Key words:
Robust speech recognition, Wireless speech recognition, Transmission errors, Modulation spectrum, RASTA-PLP

1 Introduction

Robustness in Automatic Speech Recognition (ASR) systems has always been an extremely important issue since the first attempts to transfer this technol-

Email address: jvicente, carmen, gallardo and fdiaz@tsc.uc3m.es.

ogy from research laboratories to real world applications. According to Junqua (2000) we can distinguish three sources of variability in speech that affect the performance of ASR systems: *task* and *speaker* is the first broad class, the second is the *acoustic environment* and last, *transducers* and *transmission channels*.

In this paper we are interested in dealing with the distortion produced by the new transmission channels that have emerged in voice transmission. This technology has experienced an enormous revolution in the past decade and still continues. These systems have evolved from the classical and sole transmission channel provided by Public Switched Telephone Network (PSTN) into a wide range of alternatives that include wireless cellular systems, VoIP, Bluetooth, wireless local and personal area networks of even a mixture of them.

Besides, the pervasiveness of all these means of voice transmission has triggered the creation of multiple new information providing services that users can access through these networks. These services can greatly benefit from the use of automatic dialog systems for which an improved performance of the ASR subsystem over the particular underlying transmission channels can significantly reduce the need to resort to a human operator in many situations.

In this context, our work focuses on improving the robustness of ASR systems that are accessed through a wireless network. Thus, in this scenario the speech signal is transmitted through the corresponding wireless standard channel and is recognized at a remote server. This is not the only approach to this problem. Either embedded or Distributed Speech Recognition (DSR) face up the same problem from a different point of view. Though not considered in this paper these alternatives are briefly reviewed and their drawbacks and advantages compared with those of the option considered here.

In this paper, we pay attention to two typical sources of distortion of wireless channels: lossy speech coding and transmission errors. Our work is inspired in previous works that suggested the filtering of the modulation spectrum of the speech features to deal with channel-distorted or noisy speech (Hermansky and Morgan (1994), Hanson and Applebaum (1993) or Nadeu et al. (1997) are good examples). We have applied and adapted these ideas to the distortions typical of wireless speech communications.

As a starting point, we consider two well-known parameter sets, namely: MFCC and LP-MFCC. Further on, we focus on LP-MFCC since our experiments reveal that it performs better than MFCC in presence of coding distortion and transmission errors. We also compare our proposal with RASTA-PLP (Hermansky and Morgan (1994)), a well-known filtering-based parameter set.

We show, conceptually and experimentally, that a band-pass filtering of the time sequences of the spectral parameters is beneficial to deal with distortions

due to transmission errors. Specifically, we suggest two configurations: the first one, called BPF-LP-MFCC, consists on a band-pass filtering of the LP-MFCC parameters; the second one is a modified version of RASTA-PLP, called M-RASTA-PLP, using a sharper low-pass section. In both cases, we obtain significant improvements with respect to LP-MFCC or the original RASTA, respectively, when transmission errors are considered.

The paper is organized as follows: section 2 presents the problem of ASR in wireless communication systems; section 3 describes the previous works on filtering the spectral parameters and discusses the reasons (either given by other authors in other contexts, or presented in this paper for wireless speech) for which we propose to improve and adapt this technique to the wireless speech communication scenario; section 4 describes the experimental setup, the baseline systems, and the experimental assessment of the filtering-based proposed techniques in comparison to well-known robust parameterization methods; finally, conclusions and directions of further work are summarized in section 5.

2 ASR in wireless environments

The enourmous success of the wireless cellular systems makes the analysis of the distortion caused by them a relevant issue of research. With this purpose, we can identify the main sources of distortion originated by these systems that affect the performance of speech recognizers as:

- *Acoustic environment*: though strictly speaking this is not a distortion caused by the wireless system itself we have included this category into the classification to reinforce the idea that the wireless nature of these networks have broaden dramatically the variety of situations or acoustic environments in which voice is likely to be originated. Therefore, though indirectly, it poses a new challenge on the speech recognition systems.
- *Speech coding distortion*: the wireless bandwidth is a very expensive resource due to the increasing number of emergent wireless services that has only made worse the saturation that already existed in the radio-electric spectrum. Therefore, to optimize the productivity of the spectral bands that allow the transmission using electronic devices of mass production, extremely smart bandwidth sharing protocols have been devised. As part of these efforts to maximize the utilization of the spectrum the use of medium and low-rate speech coders plays a fundamental role in the feasibility of these networks in the market place. This aggressive compression of the speech signal produces a distortion that damages the speech recognizer operation.
- *Transmission errors*: due to the unreliable and variable nature of the radio-channel, transmission errors are much more likely to happen in this type of

networks than in wired ones. This issue is partly addressed by the channel coders that aim at minimizing their effects. However, some errors remain affecting once more the performance of speech recognizers.

To overcome the effects of these sources of distortion three main approaches can be encountered in the literature, namely: local, distributed and remote speech recognition. We will outline their main strengths and limitations in the next subsection, paying special attention on the way in which they cope with the mentioned distortions.

2.1 System architectures for speech recognition over wireless cellular systems

This taxonomy was established by Digalakis et al. (1999) attending to the distribution of the processes of feature extraction (front-end) and decoding (back-end) between the *local user device* and the *service provider* computing system.

2.1.1 Local or embedded speech recognition

When both front-end (FE) and back-end (BE) modules are allocated in the user device we usually refer to it as *local* or *embedded* speech recognition (Junqua (2000)). This is indeed, the best way to avoid both coding distortion and transmission errors, since no transmission of the speech signal is needed: the speech transcription is sent to the server end as text data.

The main drawback of these systems is the limited capability of the devices, normally small, that makes the embedding of a speech recognition application extremely challenging and only allows the deployment of restricted vocabulary tasks. In fact, this is a very interesting problem per se and currently a topic of active research. However, it will not be treated in this paper.

2.1.2 Distributed speech recognition

Under the Distributed Speech Recognition (DSR) approach the BE (the most computationally demanding of the two processes) is situated in the server side, while the FE still resides in the user device (the client).

The advantages of this approach rely on the fact that the bandwidth required to transmit the features for recognition is very small, while the computational effort needed for their extraction is not so high and therefore can be accomplished by modest devices. Besides, a more protected data channel can be used for the transmission of the features instead of the speech channels used for the

coded speech transmission.

Nonetheless, in the typical voice-enabled services the amount of data sent to the server (usually an information request) is not very high and therefore it does not make a significant difference in bandwidth usage to send the features for recognition or the coded speech (provided a Discontinuous Transmission -DTX- system is used).

On the other hand, there is still a lot of research going on about the design of FE for speech recognition that can be difficult and expensive to fit in a conventional user device (see for example Chen et al. (2004)) and nonetheless provide important improvements when included in the server side.

However, in order for the FE to match the BE the user and the server must agree in the type of features that are going to be computed and therefore an important effort has been taking place to come up with the appropriate standards. The earlier standard (ETSI ES 201 108 (2003)) was found to behave poorly in noisy environments and thus recently a second Advanced Front-End (AFE) (ETSI ES 202 050 (2004)) has been defined. This ETSI initiative has produced an enormous advance in the understanding of noise influence on speech recognizers and many proposals have been shown to improve significantly the performance under those conditions. Fortunately, most of those techniques can be implemented as well in the server end (though with the impairment caused by the coding distortion) and therefore are not tied to the use of DSR approaches.

Still, there are some remaining issues for the implementation of those FE in the user devices as discussed in Kiss et al. (2003) related with infrastructure changes and application adaptations.

2.1.3 Remote speech recognition

On the other hand the remote speech recognition approach does not require the local user device to do any processing of the speech signal further than the usual encoding and transmission already embedded in the majority of them. The whole recognition process takes place at the server end.

In this paper we have chosen this approach for several reasons:

- It provides the server with the ability to choose the FE that better matches a particular application and even update it when needed.
- It does not impose restrictive conditions on the client terminal capabilities nor does it create the need for special setting or agreements between client and server.
- It preserves the transmission bandwidth requirements and the compatibility

with the existing standard-based voice applications.
- It is possible to recover the uttered speech signal with the quality provided by the speech coders employed.
- The new Adaptive Multi-Rate (AMR) speech coders balance appropriately the amount of bandwidth employed for the actual transmission and the protection of that transmission taking into account the conditions of the channel, while these amounts are fixed in the case of DSR. Besides, Tandem Free Operation (TFO) systems, when available, limit the speech coding distortion to one stage (ETSI TS 128 062 V6.1.0 (2004-12)).

As for the drawbacks of these remote recognition systems we can name two: the coding distortion and the transmission errors. We will take a closer look to these problems and the solutions provided so far in the next section.

2.2 Wireless transmission distortions

The main transmission distortions caused by the wireless cellular communication networks are the coding-decoding distortion and the transmission errors.

Thus, from the early works by Euler and Zinke (1994) and Lilly and Paliwal (1996) to the more recent by Hirsch (2002) we learned that for medium to low rate speech coders the loss of recognition accuracy becomes important but that this impairment can be greatly reduced by training the recognizer with the same speech coder (matched conditions). It is important to realize that in contrast with the environment noise distortions when the perfect matching is almost impossible in real implementations, the set of types of coding distortions is very small (the number of speech coders employed) and given that the information of the coder employed is always signaled in the communications protocol it is therefore feasible to consider the matched situation.

Bitstream-based solutions have been proposed (Peláez-Moreno et al. (2001); Kim et al. (2002); Gallardo-Antolín et al. (2005)) to cope with the transmission distortions. The principle behind those solutions is to avoid the coding distortion and an important part of the transmission errors by extracting the bits that carry the information needed for recognition before the decoding stage. This approach takes advantage of the Unequal Error Protection (UEP) of the channel coding that makes the spectral envelope of the speech signal much more robust to transmission errors than the rest of the signal. The main drawback of this approach is that the feature extraction module needs to have direct access to the bitstream.

Therefore, in our opinion, specific solutions for the channel and coding distortions compatible with the application of noise-robust FE should be considered. We have analyzed these problems in this paper considering both the channel

and the source coding which has led us to the proposal of an enhancing filtering of the modulation spectra of the speech features.

3 Filtering the time sequences of spectral parameters for wireless speech recognition

3.1 Modulation Spectra

Figure 1 illustrates the well-known process of obtaining a set of parameters from the speech signal. In particular, the speech signal is analyzed in a frame by frame basis and a N-dimensional parameter vector is obtained for each frame. Besides, we have represented the modulation spectra of each coefficient which is defined as the Fourier transform of its temporal evolution (see (Nadeu et al. (1997)) for more details).

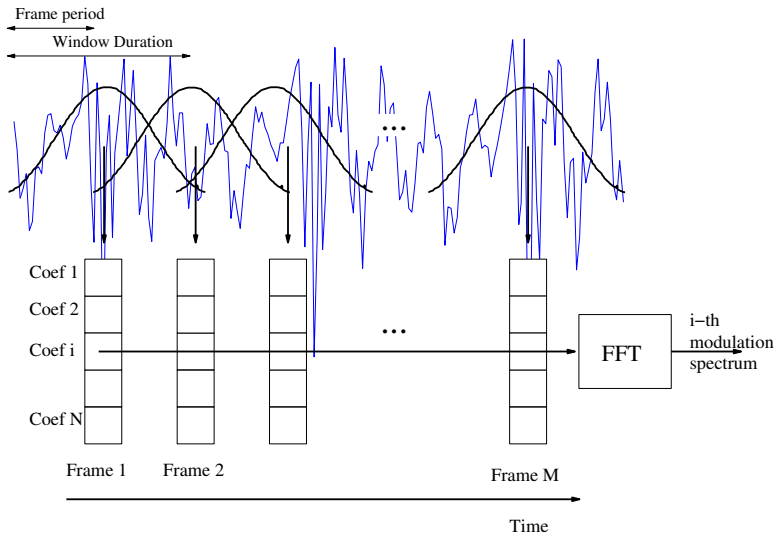

Fig. 1. The modulation spectra of the $i-th$ coefficient is the spectrum of the signal defined by the time evolution of this coefficient

In this paper we propose filtering the temporal evolution of each component to achieve robust ASR systems in wireless environments. In this section we first review the main previous works involving a filtering of the modulation spectra and, after that, we propose a novel filtering method conceived to improve the robustness of the ASR systems dealing with wireless speech transmission.

In real world applications, ASR systems often encounter situations in which a mismatch between training and testing conditions exists (e.g. noise, transmission channel or the intra- or inter-speaker variations). In such cases, there is a dramatic degradation of the recognizer accuracy.

During the past decades, a variety of techniques has been proposed for dealing with this type of problems, such as robust parameterizations, feature vector adaptation or model compensation. In this paper, we have focused on the first approach, i.e., extracting robust speech features that are relatively insensitive to different sources of degradation.

For that purpose, it would be desirable that the front-end of the speech recognition system was able to keep the linguistic information (the relevant part in terms of intelligibility) contained in the speech signal and reject the irrelevant information (for example, signal distortions due to channel or the presence of noise). This idea is directly related to the phenomena observed in several perceptual experiments in which it is shown that the intelligibility of speech mostly relies on some bands of the modulation spectra, while the rest does not seem to contribute considerably (Drullman et al. (1994), Greenberg (1996)). Typically, the suppression of the less important components of the modulation spectra is accomplished by filtering of time trajectories of feature vectors.

The RelAtive SpecTrAl technique (RASTA) (Hermansky et al. (1992), Hermansky and Morgan (1994)) is one of the pioneering techniques developed in this context. RASTA basically consists in a band-pass filtering applied in the log-subband domain, which keeps the modulation frequencies in the range between 1 and 12 Hz. The low-pass filtering helps to smooth some of the fast frame-to-frame spectral changes appearing in the spectrum due to short-term analysis artefacts. The high-pass filtering was initially designed for minimizing the influence of convolutional noise (such as distortions due to microphones or fixed-telephone channels). This effect can be viewed as that of a linear system, producing a non-desired component which is additive in the log filter-bank energies domain. As the spectrum of this kind of noise varies in a different way than the speech spectrum, it can be removed efficiently by means of the RASTA technique. In fact, Hermansky and Morgan (1994) showed that the reduction of this irrelevant information in the parametric representation of speech signals significantly improves the performance of the recognition system.

Hanson and Applebaum (1993) extended the RASTA approach by applying, in the cepstral domain, either a high-pass or a band-pass filter. Moreover, they dealt with distorted-channel, additive noise and Lombard speech style. They

showed that both, log-subband and cepstral high-pass filtering can improve the ASR system performance when a mismatch between training and testing conditions exists. Both approaches produced similar results because cepstral coefficients are computed using a Discrete Cosine Transform (DCT), which is a linear transformation of the logarithmic filter-bank energies. This result is very appealing because it allows to successfully apply filtering techniques in parameterizations where filter-bank energies are not available, such as LPC-based front-ends, as it has been also shown by Smolders and Van Compernolle (1993).

Other authors have proposed more sophisticated filters. For example, in Nadeu et al. (1997) a cascade of a first-order equalizer and a band-pass filter (a FIR-Slepian filter) was applied as well to the cepstrum-LPC domain. The authors encountered that the enhancement of modulation frequencies around 3 Hz (corresponding roughly to the average syllable rate of the used database) has a beneficial influence on the ASR system performance.

Kanedera et al. (1998) provided an interesting study about the relevance of some bands of modulation spectrum from the recognizer accuracy point of view. The main conclusions extracted in this work were the following:

- In clean environments, most of the useful information is contained in the frequency band between 1 and 16 Hz of the modulation spectrum.
- The band around 4 Hz is the most useful component in both, clean and noisy conditions (this result is similar to the one obtained in Nadeu et al. (1997)).
- In noisy environments, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility. In particular, the band below 1 Hz contains mostly information about the environment (e.g. the effects due to the transmission channel). Therefore, the recognition performance can be improved by suppressing this band in the parameterization process.

Some authors (Hanson and Applebaum (1993), Nadeu et al. (2001)) have stressed the relationship between the time filtering of speech parameters and the classical first time-derivative or regression coefficients and acceleration coefficients (Furui (1986)). In fact, dynamic features can be seen as a high-pass (in the target bandwidth) filtering of the static parameters in the cepstral domain, in which the components around 10 Hz are enhanced. This interpretation explains their effectiveness to cope with both, convolutional and additive noises.

Although more extensively explained later, we find convenient to mention at this point that the transmission errors due to wireless communications reach the speech decoder in two forms: either as residual bit errors (those still present after channel decoding) or as frame erasures. In this paper both are jointly considered, since it is the channel decoder which decides whether the frame is discarded (and substituted) or not, depending on the number of bit errors and the sensitivity of the erroneous bits. Thus, once the bitstream has been evaluated by the channel decoder, the source decoder receives either a clean frame, a frame with residual errors, or a bad frame indication. In the last case, this flag triggers the corresponding frame error concealing mechanism.

The residual bit errors produce unpredictable changes in the speech spectral features. Thus, the whole bandwidth of their modulation spectra may be eventually affected. In other words, the residual bit errors add certain level of randomness to the spectral features, i.e., noisy variations in their time evolution. These time variations generate spurious components in the modulation spectra.

With respect to frame erasures, we presume that the spectral envelope (almost exact) repetition performed by the error concealment mechanism produces both low and high frequencies in the modulation spectrum. The former due to the steadiness of the repeated segment, and the later when, after successive repetitions, a reliable frame reaches the decoder, likely producing an abrupt time change.

The previous conjectures indicate that a band-pass filtering of the modulation spectra could help to focus on the modulation frequencies which, being relevant for speech intelligibility, are less contaminated by the transmission errors.

Furthermore, in order to prove these arguments, we have estimated the bandwidth of the modulation spectrum for each MFCC coefficient extracted from both, clean speech and speech that has suffered from transmission errors. For these experiments, we consider the bandwidth as the frequency range where the 90 % of the signal energy is contained. Finally, a channel with a Bit Error Rate (BER) equal to $5 \cdot 10^{-2}$ has been used [1]. Figure 2 represents the histogram of the bandwidth computed for the first six MFCCs where each coefficient was analyzed in a window-by-window basis.

As shown in Figure 2, the effect of transmission errors in the modulation spectra of MFCCs depends on the coefficient order. Particularly, observing

[1] More details about the channel simulation and the bandwidth estimation will be provided in section 4.

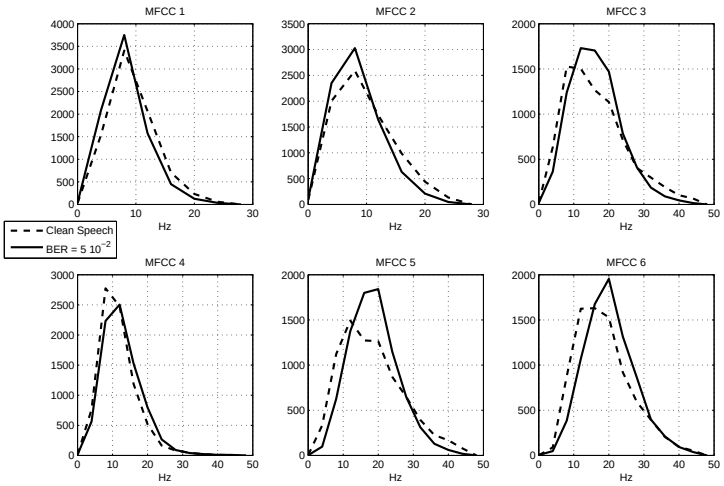

Fig. 2. Histogram of the modulation spectrum bandwidth for the first six MFCC parameters extracted from either clean speech (dashed-line) or speech that has suffered from transmission errors (solid-line).

the bandwidth histograms for the first two coefficients we see that, because of the transmission errors, a larger number of windows exhibit a lower bandwidth (the histogram is slightly left-shifted). This fact indicates that low-frequency components appear in those modulation spectra due to transmission errors, since the same percentage of the energy is concentrated in a smaller bandwidth. On the contrary, observing the bandwidth histograms for the higher coefficients we see that a larger number of windows show higher bandwidth (the histograms are slightly right-shifted due to errors).

Therefore, in order to reduce the effects of transmission errors, we propose band-pass filtering the time trajectory of the spectral parameters to attenuate or remove these undesired low or high frequencies that appear in their modulation spectra. Although, taking into consideration the histograms displayed in Figure 2, the optimal solution seems to be to high-pass filter the lower coefficients and low-pass filtering the higher ones, this issue has been left for further work (preliminary experiments using individual filters did not work as expected). Instead, we have chosen to perform the same band-pass filtering for every coefficient. There are two reasons to proceed in that way. First, it is easier to implement. And second, the histograms are showing just a trend and we have found it better to remove those frequency bands that can be contaminated and, at the same time, are not determinant from the intelligibility point of view (in presence of degradations, modulation frequencies above 10 Hz worsen the recognition performance and frequency components under 2 Hz do not yield any improvement and, furthermore, could even degrade it (Kanedera et al. (1998))).

In this paper we put forward that both low- and high-pass filtering signifi-

cantly improve the recognition performance in presence of transmission errors. In particular we suggest the replication of the well-known and well-established high-pass section of the RASTA filter, and the design of a new low-pass section to achieve the best balance between preserving relevant modulation frequencies and mitigating the effect of the transmission errors.

4 Experimental Results

4.1 Experimental Setup

4.1.1 Database

The database employed in our experiments is the well-known Resource Management RM1 Database NIST (1992), which has a vocabulary of 991 words. We have used the speaker independent data which is divided into two groups: the training corpus which consists of 3990 sentences uttered by 109 speakers and the test set which contains 1200 sentences from 40 different speakers and corresponds to the compilation of the first four official test sets (February and October, 1989, February, 1991 and September, 1992). We have used a downsampled version (at 8 KHz) of the database (originally recorded at 16 kHz in clean conditions using a high-quality desktop microphone). The orthographic transcription of the data is based on the SRI Resource Management dictionary (provided in the same distribution by NIST) which has been modified for adapting it to the CMU phone set as suggested in the RM task defined in HTK.

Since the database was recorded in a clean environment, it is possible to study the effects of the transmission errors without any interference caused by other sources of distortions.

4.1.2 Wireless channel model

For the purspose of testing the performance of our proposal in realistic conditions, we have simulated a complete GSM scenario which includes not only a channel model but also the GSM channel coding/decoding processes. The behavior of the GSM channel has been simulated for different conditions using a hybrid model combining both empirical measures (for modelling shadowing effects produced by the presence of obstacles like buildings in urban areas) and theoretical results (for the Rayleigh fading phenomena related to the mobile speed). The GSM channel coding/decoding has been implemented following the ETSI/GSM specifications for half-rate traffic channels (ETSI Recommen-

dation GSM 6.20 (1999)). It includes implementations of the channel coding (cyclic, convolutional coding) and the blocks relevant to the arrangement of the digital TDMA GSM stream (reordering, partitioning, interleaving and burst formatting). More details about the overall GSM channel simulator are given in Gallardo-Antolín et al. (2005).

The channel model inserts bursty transmission errors in the bitstream according to the desired Bit Error Rate (BER). The channel decoder is able to detect and correct some of these errors or even substitute a seriously damaged frame by an attenuated version of the last reliably received one. Therefore, two different types of errors appear at the input of the speech decoder: frame erasures and residual bit errors. The first one is measured in terms of the Frame Erasure Rate (FER) which is the percentage of erroneous frames that were replaced by the concealing mechanism and the second one is characterized by the Residual Bit Error Rate (RBER) which is the percentage of remaining transmission errors not corrected or detected in the channel decoding stage.

Following this procedure, we have designed five different half-rate GSM channels corresponding to different channel conditions ($BER = 0$, 10^{-3}, 10^{-2}, $2.5\,10^{-2}$ and $5\,10^{-2}$). The FER and RBER values of each channel are listed in Table 1. FER and RBER are not theoretical values, but experimentally computed ones for the database we have employed.

Table 1
Characteristics of the half-rate GSM channels used in the experimentation. BER, FER ("Frame Error Rate") and RBER ("Residual Bit Error Rate") are shown for each channel

Channel	**BER**	**FER**	**RBER**
Channel0	0	0 %	0 %
Channel1	10^{-3}	0.015 %	0.0265 %
Channel2	10^{-2}	0.479 %	0.2753 %
Channel3	$2.5 \cdot 10^{-2}$	2.9296 %	0.8061 %
Channel4	$5 \cdot 10^{-2}$	12.333 %	2.3222 %

These channel conditions have been chosen taking into account the eight quality bands defined in the GSM standard (ETSI ETS 300 578 (1999)) shown in Table 2. These quality bands are defined in accordance to the BER estimated before the channel decoding.

The fourth band is considered as the one representing an expected average quality. For this reason we have chosen channels with a BER around this band. Specifically, a channel in the third ($BER = 10^{-2}$), fourth ($BER = 2.5 \cdot 10^{-2}$) and fifth ($BER = 5 \cdot 10^{-2}$) quality bands have been chosen. Besides, we have tested two channels belonging to the best band: an error-free ($BER = 0$, only

Table 2
Quality bands in GSM

Quality Band	**BER**
0	$\text{BER} < 2 \cdot 10^{-3}$
1	$2 \cdot 10^{-3} < \text{BER} < 4 \cdot 10^{-3}$
2	$4 \cdot 10^{-3} < \text{BER} < 8 \cdot 10^{-3}$
3	$8 \cdot 10^{-3} < \text{BER} < 1.6 \cdot 10^{-2}$
4	$1.6 \cdot 10^{-2} < \text{BER} < 3.2 \cdot 10^{-2}$
5	$3.2 \cdot 10^{-2} < \text{BER} < 6.4 \cdot 10^{-2}$
6	$6.4 \cdot 10^{-2} < \text{BER} < 1.28 \cdot 10^{-1}$
7	$1.28 \cdot 10^{-1} < \text{BER}$

coding distortion) and a low error one ($BER = 10^{-3}$).

4.2 Baseline Experiments

4.2.1 Front-End

Two parameter sets have been used for the baseline experiments: MFCCs and LP-MFCCs and figure 3 illustrates the way we have implemented them.

As can be observed, the difference between both parameter sets relies on how the speech spectrum is obtained. In particular, the "Spectral Analysis" step in the MFCC computation is replaced by the "Pole Modeling" and "Spectrum Envelope Computation" steps in the LP-MFCC case. In this last case, the order of the all-pole model has been experimentally chosen. In particular, we have considered (always for clean speech) 8, 10, 12, 14 and 16, obtaining very similar results above 10. Consequently, we have chosen that order for our experiments.

In both cases, we use a 25 ms Hamming analysis window, obtaining 12 coefficients every 10 ms. These static features are extended with the log-energy and the corresponding first order delta parameters.

In those experiments (described further on) where an additional filtering stage is introduced, the delta features are calculated from the filtered sequence of spectral parameters (either MFCCs or LP-MFCCs). Hanson and Applebaum (1993) show that calculating the regression parameters from the filtered ones yields better results. Their experiments involved filtering either the log-subband energies or the cepstral coefficients obtained from a PLP

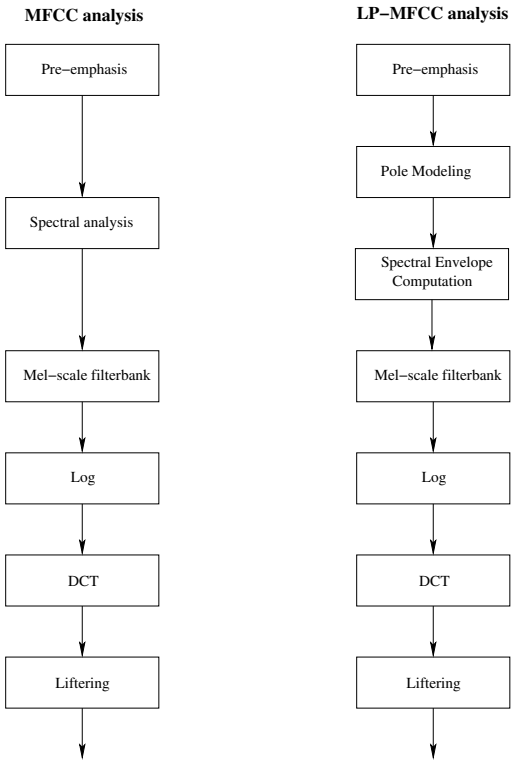

Fig. 3. MFCC and LP-MFCC analysis

(*Perceptually-based linear prediction*) analysis, but similar results can be expected for MFCCs or LP-MFCCs.

4.2.2 Back-End

The back-end is based on HMMs (Hidden Markov Models). The HTK toolkit Young et al. (1995) has been used to build the system. Context-dependent acoustic models have been used, namely: cross-word triphones. A three-state, three-mixture per state model is used to represent each triphone. The synthesis of unseen triphones in the training set was performed through a decision tree method of state clustering. Models are obtained using either clean speech (just for reference experiments) or coded speech, without transmission errors. These last models are used for every wireless speech recognition experiment, with or without transmission errors. Finally, the standard word-pair grammar is used as the language model.

It is important to note that when the temporal trajectories of the coefficients are filtered the acoustic models are trained using those filtered parameters. Thereby, we avoid any possible mismatch introduced by that filtered stage.

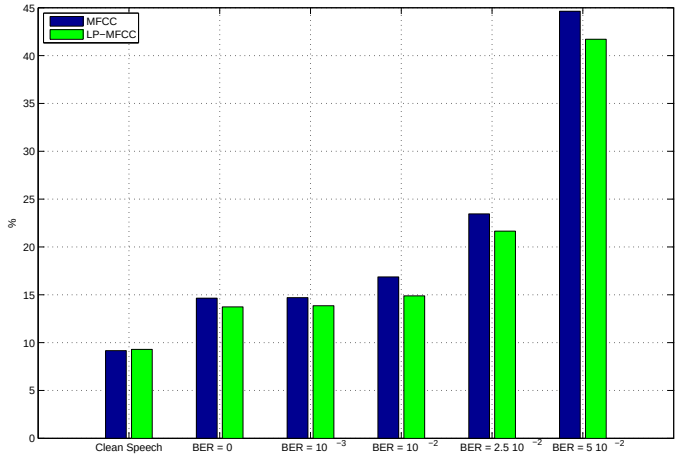

Fig. 4. Baseline results: WER(%) for two parameter sets, MFCCs and LP-MFCCs, and several channel conditions

4.2.3 Baseline results

Figure 4 shows the recognition results, in terms of Word Error Rate (WER), for the two parameter sets considered and several channel conditions, namely: clean speech, speech under coding distortion ($BER = 0$) and speech under coding distortion and transmission errors ($BER = 10^{-3},\ 10^{-2},\ 2.5 \cdot 10^{-2}$ and $5 \cdot 10^{-2}$). These results will be taken as the baseline for future comparisons.

From these experiments we draw our first conclusion: MFCCs achieve slightly better performance for clean speech but LP-MFCCs are superior when coding distortion and transmission errors are considered. Furthermore, as the channel conditions worsen the performance improvement becomes more significant. This is very likely to the smoother spectral envelope obtained due to LP analysis carried out as a part of the LP-MFCC parameterization procedure.

4.3 MFCC Bandwidths

Before designing low- and high-pass sections to filter the time sequences of the MFCCs, we have analysed which are the most relevant bands of their modulation spectra (we assume that those corresponding to LP-MFCCs will be quite similar). A block diagram of the process involved in that bandwidth estimation for every coefficient is represented in figure 5 and summarized as follows (Peláez-Moreno et al. (2002)):

- First, twelve MFCC coefficients ($MFCC\_i[n_f]$, where $i = 1, 2 \ldots 12$ and n_f is the time index) are extracted from clean speech. This process is similar to the one explained in figure 1 but, in this case, a very small frame period

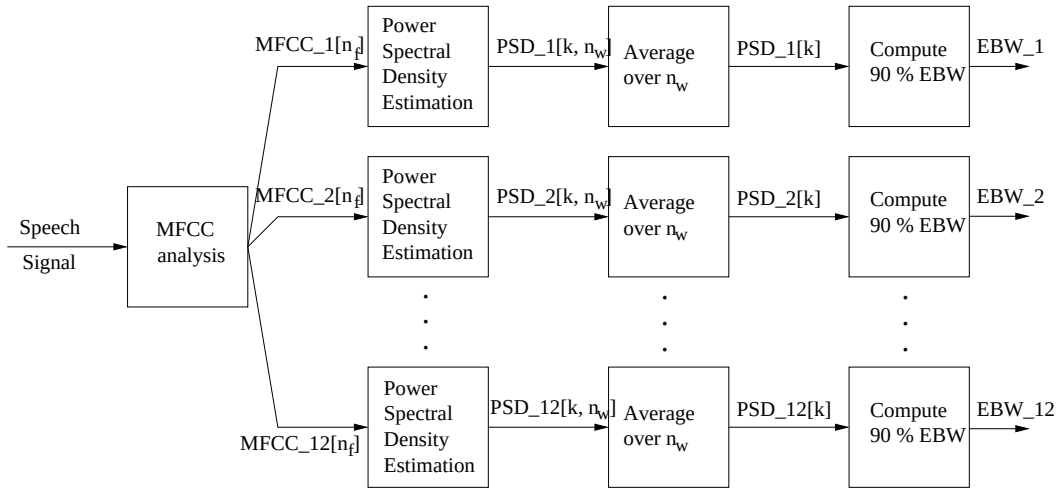

Fig. 5. 90 % Effective Bandwidth (EBW) estimation process

.

is used. As a result, the temporal trajectory of the MFCC parameters is oversampled in order to avoid any possible aliasing effect that could affect the bandwidth estimation process. Note that this frame period is just used for bandwidth estimation and not for speech recognition.

- Second, we analyze the individual time sequences corresponding to every MFCC coefficient using long Hamming windows. Specifically, we use windows with a duration equal to 2 seconds and with a 50 % of overlap between neighboring windows. On the one hand, such a long window imposes a very poor time resolution; but, on the other hand, the frequency resolution is high, making possible to estimate bandwidths with an accuracy around 1 Hz as required by this problem.

  Then, the power spectral density is computed for each window. Those signals are represented in figure 5 under the notation $PSD\_i[k, n_w]$ $(i = 1, 2 \ldots 12)$ where k represents the frequency modulation index and n_w represents the time index that corresponds to the current window.
- Third, we compute the mean power spectral density $(PSD\_i[k], i = 1, 2 \ldots 12)$ making an average over all the power spectral densities. From this mean power spectral density we compute what we call Effective Bandwidth (EBW), that is, the bandwidth within which a specific fraction of the spectral power is concentrated. For example, a 90 % EBW refers to the bandwidth containing the 90 % of the energy of the current signal (EBW_i, $i = 1, 2 \ldots 12$ in figure 5).

A similar estimation procedure is employed with the log-energy coefficient extracted from the speech signal.

Finally, in figure 6 we have depicted the 90 % EBWs for the log-energy and the twelve MFCCs. From that figure, it is clear that the EBW of the MFCCs increases with the coefficient order, starting around 9 Hz for the log-energy

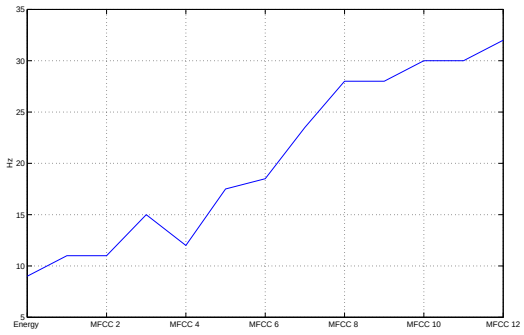

Fig. 6. 90 % Effective Bandwidth (EBW) of the log-energy and twelve MFCCs

and 11 Hz for the first MFCC and ending around 32 Hz for the 12th MFCC.

4.4 Low-pass filtering

In section 3.3 we advocated the convenience of band-pass filtering the modulation spectrum of each coefficient in order to find a robust parameter set. We design our filter in two steps: first, a tailored low pass section is build to cope with transmission errors and, second, a high pass section is added. This two stage filter design allow us to weight up the contribution of each section into the final results.

The design of the high-pass section will be presented in section 4.5 while the low-pass section is introduced in the current section.

4.4.1 FIR filters

Though previously reported results concerning the MFCC bandwidths suggest to use a different filter for every coefficient, some preliminary results did not indicate a clear advantage of using different filters. Actually, similar results were found using the same filter for every coefficient that do not justify the increment of complexity involved in the use of different filters.

Thus, using the same filter for every coefficient, we have assessed the effectiveness of low-pass filtering the two reference parameter sets, MFCC and LP-MFCC, for several channel conditions. We have finally employed an FIR filter; nevertheless, the use of an IIR filter is briefly examined in the next subsection. Figures 7 and 8 illustrate the whole schema embbeding the filtering stage of MFCC or LP-MFCC parameter set, respectively.

To gain insight on the more desirable characteristics (order and cutoff frequency) of the sought filter, its effectiveness (in terms of recognition accuracy)

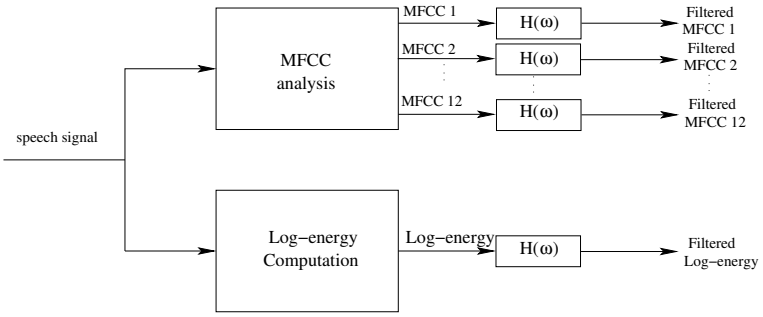

Fig. 7. MFCC filtering

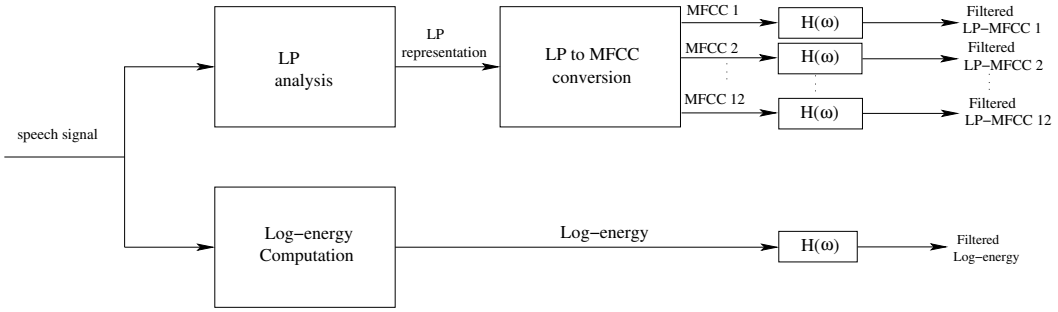

Fig. 8. LP-MFCC filtering

has been assessed for the following set of parameters and conditions:

- Order: 10, 20 and 30
- Cutoff frequency (Hz): 8, 10, 12, 18, 24, 30
- Environment: Clean speech, $BER = 0,\ 10^{-2},\ 2.5 \cdot 10^{-2},\ 5 \cdot 10^{-2}$

concluding that a 20th-order filter with a cutoff frequency of 12 Hz is the one that achieves the best results. Although a 20th-order filter seems to be too high considering the potential "time spreading" (the impulse response extends over 200 ms.), the achieved improvement (as will be shown below) under coding distortion and transmission errors is high enough to consider the selected filter order as a good trade-off.

It is worth noting that the selected cutoff frequency is close to the ones found by other authors like Nadeu et al. (1997) or Kanedera et al. (1998). If we relate this cutoff frequency with the bandwidth estimation made in the previous subsection (figure 6), we observe that the chosen cutoff frequency allows almost the whole spectral power of the first four coefficients to remain, while turning out quite selective for the remaining coefficients. In other words, this low-pass filter, in addition to removing the high frequencies of the modulation spectra, performs some type of liftering by attenuating the higher-order coefficients.

Figure 9 shows the results for the two filtered parameter sets, LPF-MFCC (Low-Pass Filtered MFCC) and LPF-LP-MFCC (Low-Pass Filtered LP-MFCC).

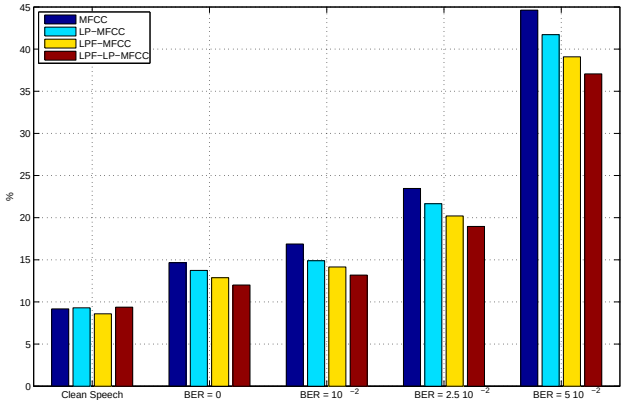

Fig. 9. WER achieved for LPF-MFCC and LPF-LP-MFCC and the corresponding unfiltered versions, MFCC and LP-MFCC

Besides, for comparison purposes, the results corresponding to the unfiltered ones are also shown.

From these results, we extract the following conclusions:

- Low-pass filtering the parameter set is beneficial from the recognition point of view. Even without any distortion, filtering the MFCC coefficients leads to some improvements. And what is more important, improvements increase as channel conditions worsen.
- When some kind of distortion is present, the best results are achieved by LPF-LP-MFCC. In particular, the relative reduction of the WER with respect to LP-MFCC has a mean of 11.9 %.

4.4.2 IIR filters

Lower-order IIR filters can be as selective as the chosen 20th-order FIR filter. Therefore, at least from the computational point of view, it is worth trying IIR filters. However, as long as the design of a computationally efficient implementation is not the aim of this paper, we have only conducted some preliminary experiments to explore if IIR filters should be considered in the future.

In particular, we haved tested a 5th order Butterworth IIR filter with a cutoff frequency of 12 Hz. This filter was assessed only for the LP-MFCC parameter set (so far, the most successful). The results are slightly lower than the ones obtained with a FIR filter. Therefore, IIR filters could be considered as an alternative and computationally more efficient implementation.

4.5 Band-pass filtering

In this section we have evaluated the benefits of the inclusion of a high-pass section to deal with coding distortion and transmission errors.

We start evaluating the performance of the RASTA-PLP method, a very well-known technique which performs a band-pass filtering in the log-spectral domain. In order to make the differences between the filtering stage in RASTA-PLP and the proposed filtering of MFCCs or LP-MFCCs clear, we briefly compare RASTA-PLP with LPF-MFCC and LPF-LP-MFCC in the following subsection.

We have also empirically compared RASTA-PLP with LPF-LP-MFCC for several channel conditions. As shown below, RASTA performs better for lower BERs (easy channels) while LPF-LP-MFCC turns out to be superior for higher BERs (difficult channels). These results reveal two conclusions: 1) the high-pass section of RASTA-PLP is beneficial; and 2) the low-pass section should be sharper for medium and high BER channels.

4.5.1 RASTA-PLP (RelAtive SpecTrAl-Perceptually-based Linear Prediction)

Figure 10 illustrates the block diagram of RASTA-PLP (Hermansky and Morgan (1994)) computation. Below follows a brief review of the goal of each block and its relation with those involved in MFCC and LP-MFCC computation (figure 3).

- Spectral analysis: the same as the one used in the MFCC analysis.
- Critical band analysis: This stage matches up with the "Mel-scale filter-bank" although the weights are different.
- LOG (logarithm): it transforms a convolutional distortion into an additive one.
- Band-pass filtering filters the time trajectories of the log-subband energies. The RASTA filter has the next transfer function:

$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - \rho z^{-1}} \tag{1}$$

  As shown in this expression, the RASTA filter has four zeros and one pole (ρ). Originally, this pole was set to 0.98 by the authors of RASTA (Hermansky and Morgan (1994)) but they also tried with different values.
- EXP (exponential): inverse of the previous logarithmic operation.
- Equal-loudness pre-emphasis: this step can be compared to "pre-emphasis" in MFCC. The goal is the same in both cases: to take into account the different sensitivity of the human hearing system to different frequency bands.

- Intensity-loudness power law: the spectrum magnitude is comprised aiming at replicating the human hearing behaviour by simulating the relationship between intensity and tonality (perceived intensity). In MFCC and LP-MFCC parameter sets this goal is pursued by the "Log" operator.
- Pole modeling: the spectral envelope is estimated. In our experiments, the model order was experimentally chosen to be 12 after having tested several orders (8, 10, 12, 14 and 16) for both clean and distorted (coding distortion and transmission errors) speech.
- Cepstral analysis: the same as the one in the MFCC or LP-MFCC analysis.

In comparison with PLP, RASTA includes a filter stage in the logarithmic spectral domain, which is implemented through three steps: "LOG", "band pass filtering" and "EXP".

With respect to MFCC and LP-MFCC, the main differences are found in how the speech spectral envelope is estimated and how the human hearing behavior is taken into account.

Several pole values of the RASTA filter (ρ in eq. (1)) have been experimentaly tested. We test values from 0.5 till 0.98 for both, clean and distorted - coding and transmission errors- speech. We observed that as the pole position gets closer to one, the results improve. However, for the highest pole positions the differences were not significant and, consequently, we have fixed the pole position, ρ, to a value equal to 0.98.

Figure 11 compares the word error rates achieved by PLP, RASTA-PLP and LPF-LP-MFCC for several channel conditions. On the one hand, the RASTA band-pass filter yieds clear improvements with respect to PLP when coding distortion and transmission errors are considered. On the other hand, LPF-LP-MFCC turns out to be better than RASTA-PLP for channels with BERs equal or higher than $2.5 \cdot 10^{-2}$, while RASTA-PLP is the best solution for lower BERs.

The last results allow us to draw two main conclusions:

- RASTA filter is effective to deal with coding distortion and transmission errors.
- Low-pass section of RASTA filter is not as selective as required for medium and high BERs.

The last conclusion leads us to propose a new approach combining a high-pass section similar to that of the RASTA filter, and a low-pass section similar to the one suggested for filtering the LP-MFCCs.

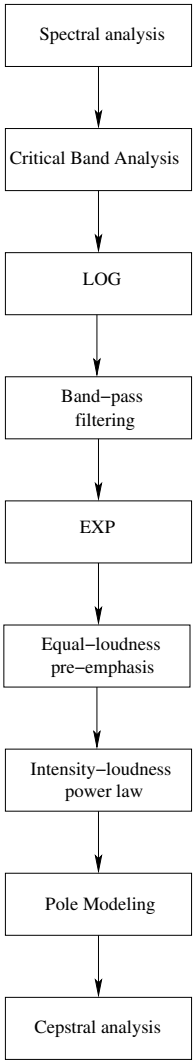

Fig. 10. RASTA-PLP analysis

4.5.2 Combining the high-pass section of RASTA-PLP with a sharper low-pass section

A band-pass filter has been designed combining the high-pass section of the RASTA filter (with the pole at $\rho = 0.98$) and a low-pass section similar to that of the 20th-order FIR filter proposed for LPF-LP-MFCC. The band-pass filter so conceived has been implemented using 20 zeros and 1 pole. Figure 12 shows the frequency amplitude response of this filter. The phase has been chosen to be linear.

We have tested this new filter in two different configurations: 1) for filtering the LP-MFCC parameter sets in the same way that we had proposed the first low-pass filtering, as illustrated in Figure 8. Henceforth, we call this approach BPF-LP-MFCC (Band-Pass Filtering LP-MFCC). And 2) as an alternative to the band-pass filtering of RASTA-PLP. From now on M-RASTA-PLP

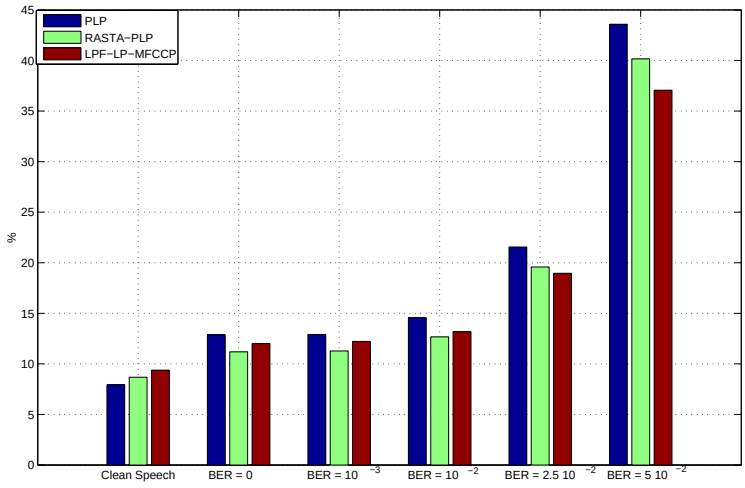

Fig. 11. WER: Comparative assessment of PLP, RASTA-PLP and LPF-LP-MFCC

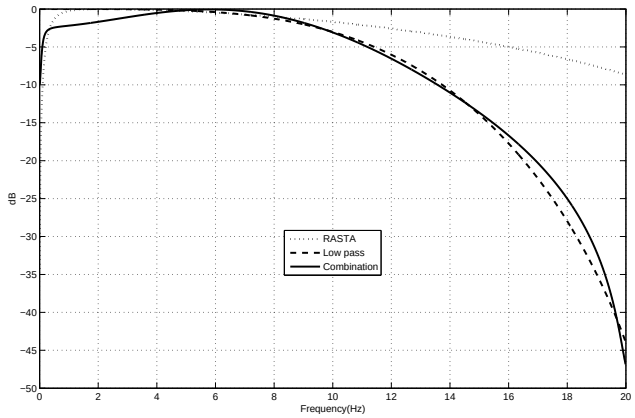

Fig. 12. Frequency amplitude response of the RASTA filter, 20th-order low-pass filter and the band-pass filter designed from the combination of them

(Modified-RASTA-PLP).

Figure 13 shows the performance, in terms of word error rate, of BPF-LP-MFCC in comparison with LPF-LP-MFCC for several channel conditions (results for LP-MFCC have also been depicted for reference). As it can be observed, BPF-LP-MFCC always yields the best results, with relative improvements (with respect to LPF-LP-MFCC) going from 2 % for a BER of 0 to 14 % for a BER of $5 \cdot 10^{-2}$ (for clean speech the relative improvement is equal to 6 %). Therefore, it can be concluded that the high-pass section of the filter is also beneficial for speech recognition in wireless environments. Furthermore, the advantage due to the high-pass section is higher as the channel conditions

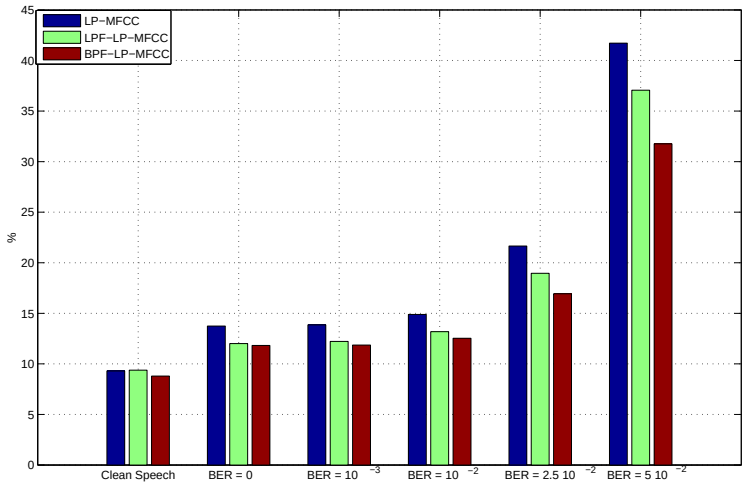

Fig. 13. WER: Comparative assessment of BPF-LP-MFCC and LPF-LP-MFCC. Results for LP-MFCC have also shown for reference

worsen. In particular, the improvements are statistically significant [2] for BERs of $2.5 \cdot 10^{-2}$ and $5 \cdot 10^{-2}$.

The results, again in terms of word error rate, corresponding to M-RASTA-PLP in comparison to RASTA-PLP for several channel conditions are shown in figure 14 (those achieved by PLP have also been included for reference). In this case, parallel conclusions can be drawn: M-RASTA-PLP is always better than RASTA-PLP and the improvement due to the new low-pass section of the filter is higher as the BER of channel increases. In particular, the improvements are statistically significant for BERs of $2.5 \cdot 10^{-2}$ and $5 \cdot 10^{-2}$.

Finally, Figure 15 shows a comparison between the two approaches, BPF-LP-MFCC and M-RASTA-PLP. Although the differences are not statistically significant, the trends are very clear: BPF-LP-MFCC is superior for higher BERs while M-RASTA-PLP is the best solution for lower BERs. We think that these results are due to the place where the pole modeling is performed. In the PLP parameter set, the pole modeling is done in its latest stages while, in the LP-MFCC parameter set, it takes place at the beginning. Although further work should be done for extracting a clear conclusion, our first intuition is that the smoothing step performed by the pole modeling should be done earlier for channels with high BERs.

[2] We have stated the statistical significance of the results calculating the confidence intervals, for a confidence of 95 % (see Weiss and Hasset (1993), pp. 407-408, for details).

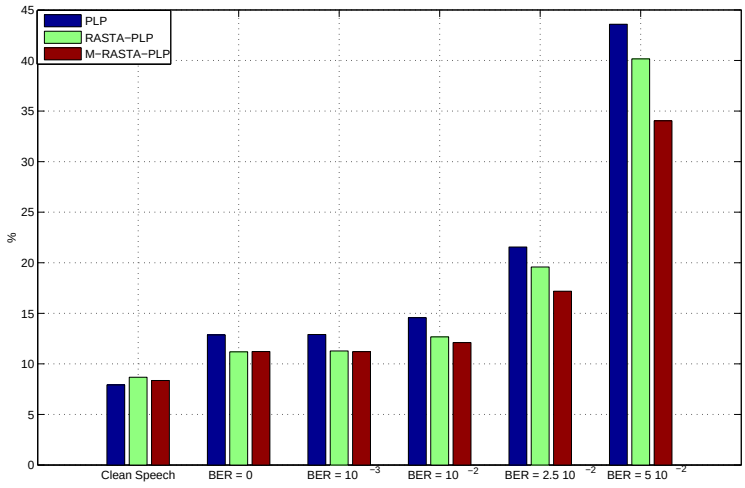

Fig. 14. Comparative assessment of M-RASTA-PLP and RASTA-PLP. Results for PLP have also been included for reference

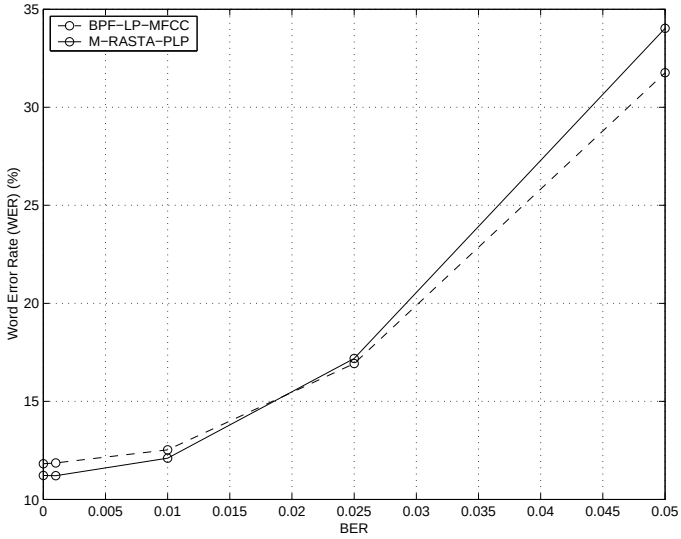

Fig. 15. WER: Comparison of the two best approches for dealing with transmission errors: BPF-LP-MFCC and M-RASTA-PLP

5 Conclusions and directions of further work

In this paper we have tackled the problem of speech recognition in a wireless environment, paying special attention to coding distortion and transmission errors. Our work is inspired in previous works that suggested filtering the modulation spectra of the recognition features to deal with channel-distorted or noisy speech. We have applied and adapted these ideas to the distortions typical of wireless speech communications.

Our work starts proposing a low-pass filtering of the recognition features to remove the potential artificial high frequencies appearing in their modulation spectrum due to transmission errors. To assess our proposal, we establish two baseline parameter sets, namely, MFCCs and LP-MFCCs. Those experiments, using either MFCCs or LP-MFCCs, showed that LP-MFCC is preferable to MFCC when coding distortion and transmission errors are considered. And, furthermore, the corresponding low-pass filtered parameter sets, called LPF-MFCC and LPF-LP-MFCC, turn out to be better than the original (unfiltered) ones. In particular, LPF-LP-MFCC provides the best results.

In order to compare our proposal with other previous works, we have also assessed the performance of RASTA-PLP, a well-established filtering-based technique, in the context of wireless speech communications. The results achieved by RASTA-PLP in comparison to those of LPF-LP-MFCC allow us to draw the following conclusions: 1) the high-pass section of the RASTA-PLP band-pass filter yields improvements in the recognition performance in presence of coding distortion and transmission errors; and 2) the low-pass section of the same filter is not sharp enough to deal with this type of distortions, especially for medium and high BERs.

Motivated by this last conclusion, we have designed a band-pass filter combining the high-pass section of RASTA-PLP with the low-pass section that we had proposed for filtering LP-MFCC. This novel filter has been applied in two configurations: 1) as an alternative to the low-pass filtering proposed to filter the LP-MFCC, called BPF-LP-MFCC and 2) as an alternative to the band-pass filter of RASTA-PLP, leading to what we have called M-RASTA-PLP.

The experimental results indicate that the novel band-pass filter provides better results than previous filters, in both configurations, when coding distortion and transmission errors are considered, especially for medium and high BERs. In particular, M-RASTA-PLP is superior to RASTA-PLP for almost every channel conditions and BPF-LP-MFCC is always better than LPF-LP-MFCC. In both cases, the improvements are statistically significant for the two highest BERs.

Finally, we have compared M-RASTA-PLP and BPF-LP-MFCC to conclude that, although both parameter sets yield similar results, it seems clear that M-RASTA-PLP should be selected for low BERs while BPF-LP-MFCC is the best option for high BERs. Although further work is needed to extract a clear conclusion, our first impression points at the position of the pole modeling stage as the responsible of that behavior: when BER is high (the distortion is high), it seems better to carry out the pole modelling at the first stages (as in LP-MFCC) of the feature extraction procedure.

We suggest four lines of research for further work. First of all, we would like to extend the experiments using the new AMR speech coder. Second, we plan to assess the proposed filter when, besides coding distortion and transmission errors, additive noise is also present. Third, the use of IIR filters should be explored in more detail. In addition, the use of different filters for each coefficient should be further investigated.

Acknowledgements

This work has been partially supported by Spanish CICYT grant TIC2002-02025 and Spanish Regional grants CAM-07T-0031-2003 and CAM-DR-SAL-0472-2004.

The authors would like to thank their colleague Ana García-Armada for her helpful discussions on realistic modeling of the wireless channel.

References

Chen, B., Zhu, Q., and Morgan, N., "Learning long-term temporal features in LVCSR using neural networks", Proc. International Conference on Spoken Language Processing (INTERSPEECH-2004), pp. 925-928, 2004.

Digalakis, V. V., Neumeyer, L.G. and Perakakis, M., "Quantization of cepstral parameters for speech recognition over the World Wide Web", IEEE Journal on Selected Areas in Communications, vol. 17, no. 1, pp. 82-90, Jan. 1999.

Drullman, R., Festen, J. M. and Plomp, R., "Effect of temporal envelope smearing on speech perception", J. Acoust. Soc. Amer., vol. 95, pp. 1053-1064, 1994.

ETSI ETS 300 578, "Digital cellular telecommunications system (Phase 2); Radio subsystem link control (GSM 05.08 version 4.22.1) ", March 1999.

ETSI TS 100 909 Ver 8.6.1 ,"Digital cellular telecommunications system (Phase 2+); channel coding", 1999.

ETSI Recommendation GSM 6.20, "Digital cellular telecommunications systems; Half Rate speech; Part 2: Half Rate Speech Transcoding", 1999.

ETSI ES 201 108 Ver. 1.1.3, "Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", 2003.

ETSI ES 202 050 Ver. 1.1.1, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2004.

ETSI ES 202 212 Ver. 1.1.1, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-

end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm", 2004.

ETSI TS 128 062 V6.1.0 "Digital cellular telecommunications system (Phase 2+);Universal Mobile Telecommunications System (UMTS);Inband Tandem Free Operation (TFO) of speech codecs;Service description", 2004.

Euler, S. and Zinke, J. "The influence of speech coding algorithms on automatic speech recognition", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-94), Adelaide, Australia, vol. I, pp. 621-624, 1994.

Fingscheidt T., Aalburg S., Stan S. and Beaugeant C., "Network-based vs. distributed speech recognition in adaptive multi-rate wireless systems", Proc. International Conference on Spoken Language Processing (ICSLP-02), pp. 2209-2212, 2002.

Furui, S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. on Speech and Audio Processing, vol. 34, pp. 52-59, 1986.

Gallardo-Antolín, A., Peláez-Moreno, C. and Díaz-de-María, F., "Recognizing GSM digital speech", IEEE Trans. on Speech and Audio Processing, vol. 13, no. 6, pp. 1186-1205, Nov. 2005.

Greenberg, S., "Understanding speech understanding - towards a unified theory of speech perception", Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pp.1-8, 1996.

Hanson, B.A. and Applebaum, T.H., "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93), vol. 2, pp. 79-82, Apr. 1993.

Hermansky, H., Morgan, N., Bayya, A. and Kohn, P., "RASTA-PLP speech analysis technique", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-92), vol. 1, pp. 121-124, 1992.

Hermansky, H. and Morgan, N., "RASTA processing of speech", IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 587-589, Oct. 1994

Hirsch H-G., "The influence of speech coding on recognition performance in telecommunication networks", Proc. International Conference on Spoken Language Processing (ICSLP-02), pp. 1877-1880, 2002.

Junqua, J. C., "Robust speech recognition in embedded systems and PC applications", Ed. Kluwer Academic Publishers, 2000.

Kanedera, N., Hermansky, H. and Arai, T., "On properties of modulation spectrum for robust automatic speech recognition" Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98), vol. 2, pp. 613-616, May 1998.

Kelleher H., Pearce D., Ealey D., Mauuary L., "Speech recognition performance comparison between DSR and AMR transcoded speech", Proc. International Conference on Spoken Language Processing (ICSLP-02), pp. 1873-1876, 2002.

Kim, H. K., Cox, R. V. and Rose, R. C., "Performance improvement of a

bitstream-based front-end for wireless speech recognition in adverse environments", IEEE Trans. on Speech and Audio Processing, vol.10, no. 8, pp. 591- 604, 2002.

Kiss, I., Lakaniemi, A., Yang, C. and Viikki, O., "Review of AMR speech codec- and distributed recognition-based speech-enabled services", Proc. of ASRU, pp. 613-618, 2003.

Lilly, B. T. and Paliwal, K. K., "Effect of speech coders on speech recognition performance", Proc. International Conference on Spoken Language Processing (ICSLP-96), Philadelphia, USA, vol. 4, pp. 2344-2347, 1996.

Nadeu, C., Pachès-Leal, P.and Juang, B.-H, "Filtering the time sequences of spectral parameters for speech recognition", Speech Communication. vol. 22, no. 4, pp. 315-32, Sep. 1997.

Nadeu, C., Macho, D. and Hernando, J., "Time and frequency filtering of filter-bank energies for robust HMM speech recognition", Speech Communication. vol. 34, pp. 93-114, 2001.

NIST, The Resource Management Corpus (RM1). Distributed by NIST, 1992.

Peláez-Moreno, C., Gallardo-Antolín, A. and Díaz-de-María, F., "Recognizing voice over IP: A robust front-end for speech recognition on the World Wide Web", IEEE Trans. on Multimedia, vol.3, no. 2, pp. 209-218, 2001.

Peláez-Moreno, C., A. Gallardo-Antolín, A., Vicente-Peña, J., and Díaz-de-María, F., "Filtering the spectral parameters to mitigate the influence of transmission errors on ASR systems", Proc. International Conference on Spoken Language Processing (ICSLP-02), pp. 2217-2220, Sep. 2002.

Smolders, J. and Van Compernolle, D., "In search for the relevant parameters for speaker independent speech recognition", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93), vol. 2, pp. 684-687, Apr. 1993.

Weiss, N. A. and Hasset, M. J., "Introductory statistics", Third Edition. Reading, MA: Addison-Wesley, 1993.

Young, S. et al, "HTK-Hidden Markov Model Toolkit (ver. 2.1)", Cambridge University, 1995.