MFCC-BASED REMOTE PATHOLOGY DETECTION ON SPEECH TRANSMITTED THROUGH THE TELEPHONE CHANNEL
Impact of Linear Distortions: Band Limitation, Frequency Response and Noise

Rubén Fraile, Nicolás Sáenz-Lechón, Juan Ignacio Godino-Llorente, Víctor Osma-Ruiz
Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid
Carretera de Valencia Km 7, 28031 Madrid, Spain

Corinne Fredouille
Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse
339, chemin des Meinajaries, 84911 Avignon Cedex 9, France

Keywords: Speech analysis, Pattern classification, Biomedical signal analysis, Communication channels.

Abstract: Advances in speech signal analysis during the last decade have allowed the development of automatic algorithms for a non-invasive detection fo laryngeal pathologies. Performance assessment of such techniques reveals that classification success rates over 90% are achievable. Bearing in mind the extension of these automatic methods to remote diagnosis scenarios, this paper analyses the performance of a pathology detector based on Mel Frequency Cepstral Coefficients when the speech signal has undergone the distortion of an analogue communications channel, namely the phone channel. Such channel is modeled as a concatenation of linear effects. It is shown that while the overall performance of the system is degraded, success rates in the range of 80% can still be achieved. This study also shows that the performance degradation is mainly due to band limitation and noise addition.

1 INTRODUCTION

The social and economical evolution of developed countries during the last years has led to an increased number of professionals whose working activity greatly depends on the use of their voice. It has been reported that this number has reached one third of the total labor force and, in parallel, that approximately 30% of the population suffers from some kind of voice disorder along their lives (Sdersten and Lindhe, 2007). In this context, methods for objective assessment of vocal function have a relevant interest (Umapathy et al., 2005) and, among them, speech analysis has the additional features of being non-invasive and allowing easy data colection (Baken and Orlikoff, 2000).

Speech assessment for the detection of pathologies has been traditionally realised through the analysis of global distortion and noise measurements taken from records of sustained vowels (Umapathy et al., 2005) (Baken and Orlikoff, 2000). Classification performances over 90% in terms of success rates have been reported for automatic pathology detection systems based on such parameters (e.g. (Boyanov and Hadjitodorov, 1997)). Recently, alternative approaches based on Mel-frequency Cepstral Coefficients (MFCC) with similar performance (Godino-Llorente and Gomez-Vilda, 2004) have also been proposed. These approaches have the advantage of relaying on robust parameters whose calculation does not require prior pitch estimation (Fraile et al., 2008a). Moreover, analysis in cepstral domain for this application is further justified by the presence of in the cepstrum information about the level of noise (Murphy and Akande, 2005). Additional reasons that support the specific processing involved in MFCC calculation can be found in (Fraile et al., 2008a), (Godino-Llorente et al., 2006) and (Fraile et al., 2008b).

From another point of view, remote diagnosis is one of the foreseen applications of telemedicine (TM Alliance Team, 2004). In this context, the use of a non-invasive diagnosis technique such as speech analysis is well suited to that application. Moreover, since the analogue wired telephone network is one of the most mature and widely extended communications infrastructures, it seems reasonable to expect

Fraile R., Sáenz-Lechón N., Godino-Llorente J., Osma-Ruiz V. and Fredouille C. (2009).
MFCC-BASED REMOTE PATHOLOGY DETECTION ON SPEECH TRANSMITTED THROUGH THE TELEPHONE CHANNEL - Impact of Linear Distortions: Band Limitation, Frequency Response and Noise.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 41-48
DOI: 10.5220/0001534200410048

that it will become one of the supporting technologies for that medical service. However, the feasibility of such application will heavily depend on the ability of voice analysis to extract significant information from speech signals even after the distortion caused by the communications channel.

Up to now, some preliminary works on this issue have been carried out and published. In the first place, pathology detection on voice transmitted over the phone has been shown to experiment a performance degradation figure around 15% when detection is based on traditional acoustic parameters (Moran et al., 2006). Secondly, the impact of several speech coders on voice quality has been studied, but without regarding the additional degradation introduced by communications channels (Jamieson et al., 2002). Last, the problem of analysing the effect of the analogue telephne channel on a MFCC-based system for pathology detection has also been approached (Fraile et al., 2007), but without differentiating among the different distortions introduced by the channel and without accounting for noise distortion.

Considering all above-mentioned aspects, that is, the adequateness of MFCC for automatic pathology detection and the interest of analyzing the impact of the analogue telephone channel on speech quality, this paper offers a detailed report on the effect of the distortions introduced by the telephone channel on the performance of automatic pathology detection based on MFCC. More specifically, a study more complete than that of (Fraile et al., 2007) is provided in which the effects of band limitation, frequency response of the channel and additive noise are analysed separately. This way, the results of the study are useful, not only for remote diagnosis applications such as the one described before, but also for setting minimum conditions, in terms of bandwidth and noise levels, for speech recording in clinical applications.

The rest of the paper is organised as follows: section 2 contains the specific formulation of MFCC and the values for related parameters used in the study, section 3 describes the model of telephone channel that has been considered, in section 4 the database, classifier and procedure used for the experiment are detailed, results are reported in section 5 and, last, section 6 is dedicated to the conclusions.

2 MFCC FORMULATION

As argued in (Fraile et al., 2008a), the variability of the speech signal is specially relevant in the presence of pathologies, thus justifying the use of short-term signal processing. A framework for such short-term processing in the case of speech is provided in (Deller et al., 1993). Within this framework, the short-time MFCC definition given in (Fraile et al., 2008b), which is slightly different from the original proposal in (Davis and Mermelstein, 1980) but it has an easier interpretation, is used:

$$c_p[q] = \frac{1}{M+1}\sum_{k=1}^{M} \log\left|\widetilde{S}_p(k)\right| \cdot \cos\left(\frac{\pi k}{M+1}\cdot q\right) \quad (1)$$

where p is the frame index, q is the index of the MFCC that ranges from 0 to M, M is the number of Mel-band filters used for spectrum smoothing and $\left|\widetilde{S}_p(k)\right|$ is the estimate of the spectral energy of the speech signal in the k^{th} Mel band. Specifically:

$$\widetilde{S}_p(k) = \sum_{f_i^m \in I_k^m} \left(1 - \frac{\left|f_i^m - F^m \cdot \frac{k}{M+1}\right|}{\Delta f^m/2}\right) \cdot \left|S_p(i)\right| \quad (2)$$

where $S_p(i)$ is the i^{th} element of the short-time discrete Fourier transform of the p^{th} speech frame, f_i^m is its associated Mel frequency,

$$I_k^m = \left[F^m \cdot \frac{k-1}{M+1}, F^m \cdot \frac{k+1}{M+1}\right] \quad (3)$$

is the k^{th} band in Mel-frequency scale, $\Delta f^m/2$ is the width of these Mel bands and F^m is the maximum frequency in Mel domain, which corresponds to half the sampling frequency of the speech signal. The frequency transformation that allows passing from linear to Mel scale is:

$$f^m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

For the herein reported application, speech frame duration has been chosen to be 20 ms, which allows capturing the spectral envelope of speech for fundamental frequencies above 50 Hz, thus covering the cases of both male and female voices (Baken and Orlikoff, 2000). Overlap between consecutive frames was 50%. The number of Mel band filters M has been made equal to 31, since that value has shown to exhibit good preformance (Fraile et al., 2008b) and vectors of 21 MFCC, that is $q \in [0,20]$, have been used as feature vectors for each speech frame.

3 TELEPHONE CHANNEL MODEL

The task of assessing the impact of the analogue telephone channel on the performance of a MFCC-based pathology detector was done bearing in mind

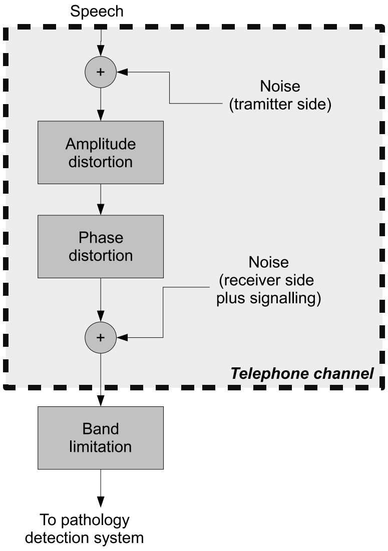

Figure 1: Block diagram of the analogue telephone channel model.

the same modeling methodology as in (Fraile et al., 2007). Such methodology comprises the main aspects of the model proposed in (Dimolitsas and Gunn, 1988). Namely, the linear effects of the channel have been assumed to be the dominant ones: amplitude, phase and noise distortions. Normative restrictions on amplitude and phase distortion imposed by (ITU, 1998) have also been taken into account. The block diagram of the overall channel model is drawn in figure 1 and it consists of the following elements:

- *Amplitude Distortion.* Its limits are normalised in (ITU, 1998) for the 300-3400 Hz band and no restrictions are imposed outside that band.
- *Phase Distortion.* Its limits for the 300-3400 Hz band are also specified in (ITU, 1998) and they are mainly referred to the phase effects at the edges of that band.
- *Noise Distortion.* This distortion can be split in noise at the transmitter side, which undergoes the same amplitude and phase distortion as the speech signal, and noise at the receiver side that does not suffer that distortion.
- *Bandwidth Limitation.* This has to be carried out as the first stage of the detector due to the uncertainty about the distortion out of the 300-3400 Hz band. Another reason for this limitation is that the telephone network adds some signalling in the 0-300 Hz band (ITU, 1998).

3.1 Amplitude Distortion

The analogue telephone channel acts as a band-pass filter. Attenuation of high frequencies comes from the low-pass behaviour of the transmission line while attenuation of low frequencies (below 300 Hz) allows the use of out-of-band signalling. Limits recommended by (ITU, 1998) for the amplitude response of the channel are represented as continuous lines in figure 2.

The simulation of the amplitude and phase distortion of the channel has been realised separately, as proposed in (Dimolitsas and Gunn, 1988) and illustrated in figure 1. Within such a setup, the amplitide distortion has been modeled as a band-pass linear-phase system, hence achieving null phase distortion in this stage, implemented by means of a symmetric FIR filter. Bearing in mind restrictions in (ITU, 1998), a 176-order filter has been designed that has the frequency response plotted in figure 2 (dashed line).

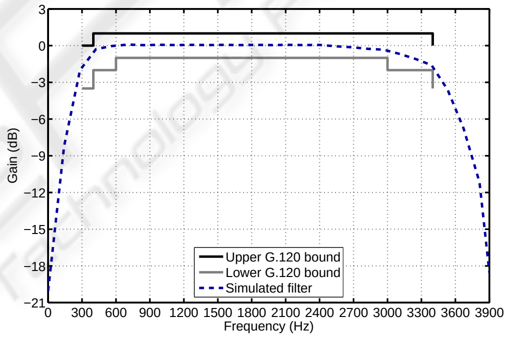

Figure 2: Amplitude response of the channel: restrictions (continuous line) and model (dashed line).

3.2 Phase Distortion

Regarding phase distortion, (ITU, 1998) imposes limits to group delay variations within the pass band. Namely, different limits are specified for the low and high parts of the band, as represented by the thick lines in figure 3. A simple procedure to obtain an all-pass filter that achieves phase distortion around certain frequencies is to design an IIR filter having zeros and poles in the frequencies at which phase distortion has to be greatest. For the filter to be all-pass, zero and pole modules must be symmetric with respect to the unit radius circle of the z-plane. Specifically, the implemented filter corresponds to the following transfer function:

$$H(z) = H_{ap}(z; f_{low}) \cdot H_{ap}(z; f_{high}) \qquad (5)$$

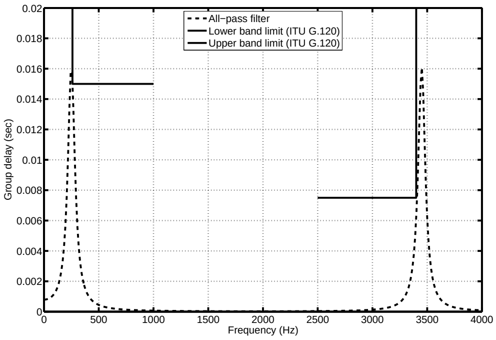

Figure 3: Phase response of the channel: restrictions (continuous line) and model (dahsed line).

$$H_{ap}\left(z;f_x\right) = \left[\frac{1-rz^{-1}e^{j2\pi\frac{f_x}{f_s}}}{1-\frac{1}{r}z^{-1}e^{j2\pi\frac{f_x}{f_s}}}\cdot\frac{1-rz^{-1}e^{-j2\pi\frac{f_x}{f_s}}}{1-\frac{1}{r}z^{-1}e^{-j2\pi\frac{f_x}{f_s}}}\right]^2 \tag{6}$$

where $r = 1.01$, $f_{low} = 250$ Hz, $f_{high} = 3450$ Hz and f_s is the sampling frequency of the speech record. The obtained frequency-dependent group delay is depicted in figure 3. It can be noticed that the maximum phase distortion happens at the limits of the pass band of the FIR filter, as specified by (ITU, 1998).

3.3 Band Limitation

The above-mentioned specifications for the frequency response of the telephone channel only cover the band between 300 and 3400 Hz, thus leaving uncertainty as for the distortion that the speech signal undergoes out of that band. In addittion, as specified by (ITU, 1998), out-of-band signalling is allowed in the 0-300 Hz band. This adds the possibility of narrow-band noise distortion to the lack of normalisation of the response of the channel within that band. These facts make it logical to perform a band limitation of the speech signal prior to its analysis, as indicated in figure 1. In this way, only the 300-3400 Hz band of the signal is further processed. This band limitation procedure is of common use in other speech processing applications (Reynolds et al., 1995).

The band limitation has a direct effect on the computation of MFCC. Specifically, the Δf^m parameter in 2 depends on both the bandwidth of the signal and the number of mel-band filters used for MFCC calculation. When limiting the frequency band of the signal, two strategies may be followed in the subsequent analysis: either maintaining the number of mel bands, hence reducing Δf^m, or keeping Δf^m approximately equal by reducing the number of bands. The performance of these two options will be analysed in section 5.

3.4 Additive Noise

The fourth modeled distortion of the telephone channel is noise. Although more complex models exist for telephone noise modelling (Dimolitsas and Gunn, 1988), herein a simpler approach, similar to (Reynolds et al., 1995), has been chosen. Namely, noise has been considered to be additive and white Gaussian (AWGN). Yet, a differentiation has been made between noise that suffers the same channel effects as the speech signal, accounting for the transmitter side, and noise that does not pass through the channel, hence the receiver side. In both cases, signal-to-noise ratio (SNR) has been controlled by tuning the power of noise to the specific power of each processed signal.

4 SIMULATION PROCEDURE

4.1 Database

All the herein reported results have been obtained using a well-known database distributed by Kay Elemetrics (MEE, 1994). More specifically, the utilized speech records correspond to sustained phonations of the vowel /ah/ (1-3 s. long) from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders in different stages (from early to mature). The subset taken corresponds to that reported in (Parsa and Jamieson, 2000) and it corresponds to 53 records from healthy patients (normal set) and 173 to ill patients (pathological set).

The speech samples were collected in a controlled environment and sampled at sampling rates equal to either 50 or 25 kHz with 16 bits of resolution. A down-sampling with a previous half band filtering has been carried out over some registers in order to adjust every utterance to the sampling rate of 25 kHz.

4.2 Classifier

The chosen classifier consists of a 3 layered Multilayer Perceptron (MLP) neural-network (Haykin, 1994) with 40 hidden nodes having logistic activation functions (as in (Godino-Llorente and Gomez-Vilda, 2004)) and two outputs with linear activations. The use of two linear outputs allows obtaining two values for each speech frame, characterised by its MFCC vector $\mathbf{c_p}$. In the training phase of the MLP, one output is trained to produce a value of “0” for pathological voice frames and “1” for normal voice frames, while the other output is trained to produce a “0” for normal data and a “1” for pathological data. In the testing phase, each output value is an estimation of the likelihood of that frame to be either normal $L_{nor}(\mathbf{c_p})$ (first output) or pathological $L_{pat}(\mathbf{c_p})$ (second output).

These likelihoods, whilst not probabilities, give an idea of how feasible is that any particular frame corresponds to each class or set. Their precise values depend on the value of the feature vector components and on the learned parameters of the MLP. Since the orders of magnitude of both likelihoods may significantly differ, it is more usual to compute log-likelihoods; the classification decision for the p^{th} frame is, then, based on the difference between log-likelihoods, as described in (Bimbot et al., 2004):

$$\log[L_{nor}(\mathbf{c_p})] - \log[L_{pat}(\mathbf{c_p})] > \theta \qquad (7)$$

If the previous condition is met, then the speech frame is classified as normal, if not, it is considered pathological. In ideal conditions, that is, if the likelihoods could be perfectly estimated by the classifier, then the value for the threshold θ should be θ =0. In practice, however, this is not the case and the choice of θ helps to make the decision system more or less conservative. Nevertheless, since decisions in this case should not be taken at the frame level, but at the record level, a mean log-likelihood difference is computed and this is the value actually compared to the threshold:

$$\frac{1}{N_{frames}} \cdot \sum_{p=1}^{N_{frames}} \log[L_{nor}(\mathbf{c_p})] - \log[L_{pat}(\mathbf{c_p})] > \theta \qquad (8)$$

where N_{frames} is the number of frames of the speech record.

4.3 Testing Protocol

The testing of each detection scheme consists of an iterative process. Within each iteration 70% of the available speech records have been randomly chosen for training the classifier, that is, to estimate the likelihood functions mentioned above. Among the remaining 30% of records, one third (10%) have been used for cross-validation during training in order to get an objective criterion for finishing the training phase (Haykin, 1994). The rest (20%) have been used for testing. For each testing record, a decision according to the previously described framework has been taken. Last, with the decisions corresponding to all the testing records, misclassification rates for different values of θ and the corresponding iteration have been computed. Twenty iterations with independently chosen training, validation and testing sets have been repeated.

5 RESULTS

There are several performance indicators for the evaluation of detection systems. A summary of the most typically used for speech applications can be found in (Bimbot et al., 2004). Among these indicators, the DET plot (Martin et al., 1997) and the Equal Error Rate (EER) have been chosen for this study as graphic and quantitative indicators, respectively. For the DET plot, *false alarm* has been defined as the event of detecting a normal voice as pathological, while *miss* means the event of detecting a pathological voice as normal. In this context, the DET curve represents the relationship between miss and false alarm rates as the threshold θ in (7) and (8) changes and the EER is the point at which the DET curve crosses the diagonal of the graph, i.e. the value of miss and false alarm rates when θ is tuned so that they coincide. In all experiments, the results have been computed both at frame and record levels, corresponding to (7) and (8).

5.1 Effect of Band Limitation

As indicated in figure 1, the first step in the speech analysis after transmission through the telephone channel is band limitation. This involves taking only the spectral energy between 300 Hz and 3400 Hz for spectrum smoothing using the Mel filter bank. Such bandwidth reduction can be achieved in two different ways. The first of them consists in maintaining the number of filters (M=31), thus reducing their individual widths. The second option, instead, involves maintaining the filter width by reducing the number of filters. It can be checked that if the band is split in 16 Mel bands (M=16), very similar Mel-filter widths are achieved. However, this means reducing the number of MFCC from 21 ($q \in [0,20]$) to 16 ($q \in [0,15]$), since $q < M$ due to the periodic nature of the discrete-time Fourier transform.

In figure 4, the different performances of both alternatives are represented by means of the averaged empirical EER and their 95% confidence intervals. The results indicate, on the one hand, that a significant increase in EER is produced by the band limitation inherent to the telephonic channel. Such observation is complementary to results reported in (Pouchoulin et al., 2007), where it was shown that the most relevant band for dysphonia detection was between 0 and 3000 Hz. The herein reported results indicate that there is significant information within the lower part of that band, that is, below 300 Hz. On the other hand, the plot in figure 4 also indicates that maintaining the size of the Mel-bands gives similar results to keeping the number of bands, but with the advantage of

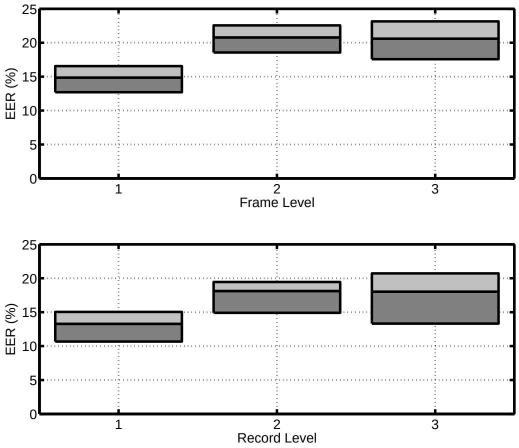

Figure 4: Average EER (central line of each box) and their 95% confidence interval (top and bottom of each box) at frame level (up) and record level (down). Case (1) corresponds to the original records and 31 Mel-band filters, (2) to band-limited signals with 31 Mel bands and (3) to band-limited signals with 16 Mel bands.

lower dimensionality. Consequently, this will be the preferred option for the next experiments.

5.2 Effect of Amplitude Distortion

In (Fraile et al., 2007), it was shown that the amplitude distortion of the speech signal has the effect of performing a quasi-linear transformation in the MFCC values. Taking this into account and recalling (1), the transformed MFCC can be written as:

$$\begin{aligned} \tilde{c}_p[q] &= A + c_p[q] + \\ &+ \frac{1}{M+1}\sum_{k=1}^{M}\log|\xi(k)|\cdot\cos\left(\frac{\pi k}{M+1}\cdot q\right) \end{aligned} \quad (9)$$

where A is a constant that depends on the amplitude response of the filter and $\xi(k)$ is a variable term that depends on the relation between the spectrum of the speech signal and the response of the filter within the k^{th} Mel-frequency band.

Figure 5 shows the plots that illustrate the average EER with the associated confidence intervals when the training stage of the classifier is done with the original speech records, with band limitation and M=16, and the testing is done with the outputs of filtering those records with the filter corresponding to figure 2 (case 3). To ease comparison, plots corresponding to the original records without band limitation (case 1) and the band limited analysis with no distortion (case 2) are plotted in the same graph. It can be noticed that the limited distortion allowed within the 300-3400 Hz band by ITU specifications (ITU, 1998) has the consequence of not affecting greatly the performance of the system.

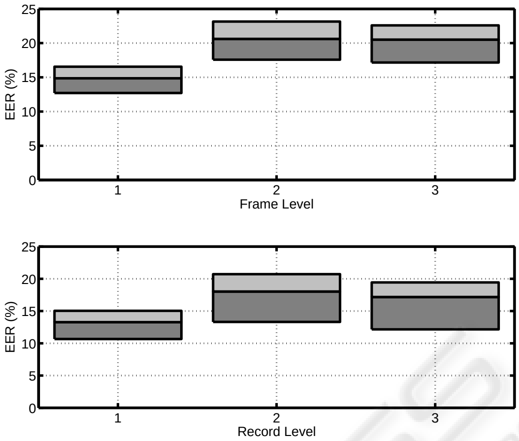

Figure 5: Average EER and 95% confidence intervals at frame level (up) and record level (down). Case (1) corresponds to the original records and 31 Mel-band filters, (2) to band-limited signals with 16 Mel bands and (3) to amplitude-distorted band-limited signals with 16 Mel bands.

5.3 Effect of Phase Distortion

As proven in (Fraile et al., 2007), the computation of MFCC involves calculation of the modulus of the discrete Fourier transform of the signal, as indicated in (1). Consequently, MFCC are insensitive to phase distortions and there is no need to analyse this effect of the channel.

5.4 Effect of Noise Distortion

The last effect of the channel to be analysed is noise distortion. This has been modelled as AWGN with different power levels. The effect of noise was analysed both independently and in conjunction with the band-limiting scheme explained before. As for the independent analysis, the obtained distributions of EER for different levels of signal-to-noise ratio (SNR) are plot in figure 6. In all cases, the training was done with the clean records and the testing with the noisy ones. The plot indicates that for SNR values around 30 dB the overall performance does not degrade greatly. However, if SNR falls below 24 dB, the error rate at record level tends to grow above 15%. While the effect of noise in the case of the telephone channel is not isolated from other distortions, these results are also useful for determining the minimum required quality of speech recordings for pathology assessment. Under the AWGN assumption, SNR val-

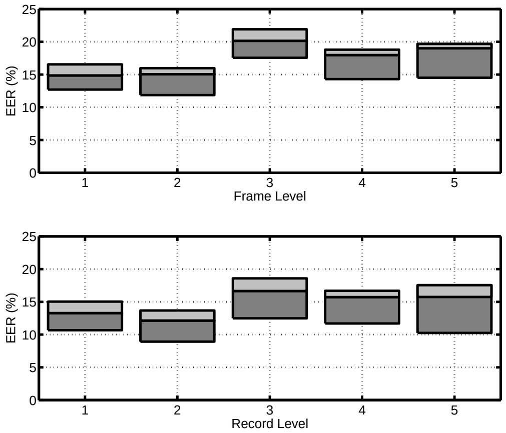

Figure 6: Average EER and 95% confidence intervals at frame level (up) and record level (down). Case (1) corresponds to the original records and cases (2) to (5) to SNR values of 30 dB, 24 dB, 18 dB and 12 dB, respectively.

ues below 24 dB seem not to be acceptable for this application.

The figure of 20 dB has been considered as a reference for the combined analysis of band limitation and amplitude and noise distortions. It has been found that, coherently with above-reported results, there is not any significant difference between adding the noise previously to the amplitude distortion (transmitter side) or after (receiver side). For the subsequent experiment, noise addition has been split in two parts: half of the power prior to amplitude distortion and half of the power after. Figure 7 shows the plots of average EER for the original speech records and those obtained after the three distortions (band limitation, amplitude distortion and noise addition). On the whole, the average EER suffers a degradation of below 10%, yielding a success classification rate over 80% at the record level. A DET plot of the same results is depicted in figure 8.

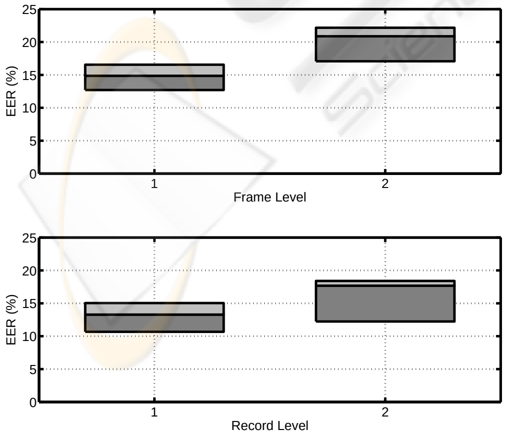

Figure 7: Average EER and 95% confidence intervals at frame level (up) and record level (down). Case (1) corresponds to the original records and case (2) to records undergoing the full modeled channel distortion.

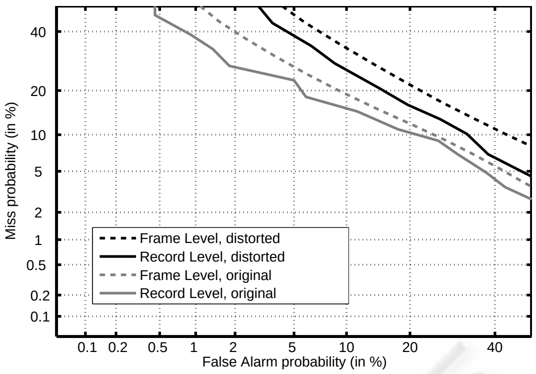

Figure 8: DET plot of the pathology detection system for the original speech records (gray) and those with simulated telephone channel distortion (black).

6 CONCLUSIONS

Within this paper, the performance of a speech pathology detector based on Mel Frequency Cepstral Coefficients when the speech signal has undergone the distortion of an analogue communications channel has been analysed. Namely the telephone channel has been modeled as a concatenation of linear effects: band limitation, amplitude distortion, phase distortion and noise addition. It has been shown that while the overall performance of the system is degraded, success rates over 80% can still be achieved. This study also reveals that the performance degradation is mainly due to band limitation and noise addition. Amplitude distortion, if complying with norm (ITU, 1998), has little impact and phase distortion has no impact at all.

As for the most relevant sources of distortion, it has been shown that the loss of information in the 0-300 Hz band makes performance to decrease significantly. Additionally, the effect of noise degradation becomes very relevant for values of SNR below 24 dB. For SNR equal to 20 dB, and considering bandwidth limitation and amplitude distortion too, success classification rate can reach 80%. This figure is better than the results reported in (Moran et al., 2006).

The whole set of reported results allow to conclude, in the first place, that remote pathology detection on speech transmitted through the analogue telephone channel seems feasible and, in the second place, that MFCC parameterization can provide a robust method for assessing the quality of degraded speech signals.

ACKNOWLEDGEMENTS

This research was carried out within projects funded by the Ministry of Science and Technology of Spain (TEC2006-12887-C02) and the Universidad Politécnica de Madrid (AL06-EX-PID-033). The work has also received support from European COST action 2103.

REFERENCES

(1994). Voice disorders database v.1. CD-ROM. Massachusetts Eye and Ear Infirmary.

(1998). Transmission characteristics of national networks. Series G: Transmission Systems and Media, Digital Systems and Networks Rec. G.120 (12/98), ITU-T.

Baken, R. J. and Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice*. Singular Publishers, San Diego (USA).

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004(4):430–451.

Boyanov, B. and Hadjitodorov, S. (1997). Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. *IEEE Engineering in Medicine and Biology*, 16(4):74–82.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366.

Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (1993). *Discrete-time processing of speech signals*. Macmillan Publishing Company, New York (USA).

Dimolitsas, S. and Gunn, J. E. (1988). Modular, off line, full duplex telephone channel simulator for high speed data transceiver evaluation. *IEE Proceedings*, 135(2):155–160.

Fraile, R., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., and Gomez-Vilda, P. (2007). Analysis of the impact of analogue telephone channel on MFCC parameters for voice pathology detection. In *Proceedings of the 8th INTERSPEECH Conference (INTERSPEECH 2007)*, pages 1218–1221.

Fraile, R., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., and Gómez-Vilda, P. (2008a). Use of cepstrum-based parameters for automatic pathology detection on speech. Analysis of performance and theoretical justification. In *Proceedings of Biosignals 2008*, volume 1, pages 85–91.

Fraile, R., Saenz-Lechon, N., Godino-Llorente, J. I., Osma-Ruiz, V., and Gomez-Vilda, P. (2008b). Use of mel-frequency cepstral coeffcients for automatic pathology detection on sustained vowel phonations: Mathematical and statistical justification. In *Proceedings of the International Symposium on Image/Video Communications over fixed and mobile networks*, volume Accepted.

Godino-Llorente, J. I. and Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384.

Godino-Llorente, J. I., Gomez-Vilda, P., and Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *IEEE Transactions on Biomedical Engineering*, 53(10):1943–1953.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Macmillan, New York.

Jamieson, D. G., Parsa, V., Price, M. C., and Till, J. (2002). Interaction of speech coders and atypical speech ii: Effects on speech quality. *Journal of Speech, Language and Hearing Research*, 45:689–699.

Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M., and Przybocki, M. A. (1997). The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech '97*, volume IV, pages 1895–1898, Rhodes, Crete.

Moran, R. J., Reilly, R. B., de Chazal, P., and Lacy, P. D. (2006). Telephony-based voice pathology assessment using automated speech analysis. *IEEE Transactions on Biomedical Engineering*, 53(3):468–477.

Murphy, P. J. and Akande, O. O. (2005). Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation. In *Proceedings of the 3^{rd} International Conference on Non-Linear Speech Processing (NOLISP'05)*, pages 224–232.

Parsa, V. and Jamieson, D. G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language and Hearing Research*, 43(2):469–485.

Pouchoulin, G., Fredouille, C., Bonastre, J. F., Ghio, A., and Giovanni, A. (2007). Frequency study for the characterization of the dysphonic voices. In *Proceedings of the 8th INTERSPEECH Conference (INTERSPEECH 2007)*, pages 1198–1201.

Reynolds, D. A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C., and Carlson, B. A. (1995). The effects of telephone transmission degradations on speaker recognition performance. In *Proceedings of ICASSP '95*, volume 1, pages 329–332, Detroit, MI, USA.

Sdersten, M. and Lindhe, C. (2007). Voice ergonomics - an overview of recent research. In Berlin, C. and Bligard, L. O., editors, *Proceedings of the 39th Nordic Ergonomics Society Conference*.

TM Alliance Team (2004). Telemedicine 2010: Visions for a personal medical network. Technical Report BR-29, ESA Publications Division.

Umapathy, K., Krishnan, S., Parsa, V., and Jamieson, D. G. (2005). Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering*, 52(3):421–430.