Automatic Detection of Laryngeal Pathology on Sustained Vowels Using Short-Term Cepstral Parameters

Analysis of Performance and Theoretical Justification

Rubén Fraile[1], Juan Ignacio Godino-Llorente[1], Nicolás Sáenz-Lechón[1],
Víctor Osma-Ruiz[1], and Pedro Gómez-Vilda[2]

[1] Department of Circuits & Systems Engineering
Universidad Politécnica de Madrid
Carretera de Valencia Km 7, 28031 Madrid, Spain
rfraile@ics.upm.es, igodino@ics.upm.es, nicolas.saenz@upm.es,
vosma@ics.upm.es
[2] Department of Computer Systems' Architecture and Technology
Universidad Politécnica de Madrid
Campus de Montegancedo s/n, Boadilla del Monte, 28660 Madrid, Spain
pedro@pino.datsi.fi.upm.es

Abstract. The majority of speech signal analysis procedures for automatic detection of laryngeal pathologies on speech mostly rely on parameters extracted from time-domain processing. Moreover, calculation of these parameters often requires prior pitch period estimation; therefore, their validity heavily depends on the robustness of pitch detection. Within this paper, an alternative approach based on cepstral-domain processing is presented which has the advantage of not requiring pitch estimation, thus providing a gain in both simplicity and robustness. While the proposed scheme is similar to solutions based on Mel-frequency cepstral parameters, already present in literature, it has an easier physical interpretation while achieving similar performance standards.

1 Introduction

Analysis of recorded speech is an attractive method for pathology detection since it is a low-cost non-invasive diagnostic procedure [1]. Although there is a wide range of causes for pathological voice (functional, neural, laryngeal, etc.) and a correspondingly wide range of acoustic parameters has been proposed for its detection (see [2] for summarising tables and typical values), these intend to detect speech signal features that may be roughly classified in only three classes [3]:

- *Short-term frequency perturbations*: both in fundamental frequency and in formants.
- *Short-term amplitude perturbations*.
- *Noise* or, more specifically, speech-to-noise ratio.

Calculation of above-mentioned acoustic parameters requires previous and reliable detection of speech fundamental frequency (pitch) [4] [1]. Nevertheless, pitch detection is not an easy task due to its sensitiveness to noise, signal distorsion, speech formants, etc. [5].

An alternative approach to speech signal analysis is doing it in cepstral domain, more specifically in Mel-frequency cepstral domain. Such approach, consisting in classifying patterns of so-called Mel-frequency cepstral coefficients (MFCC), does not require prior pitch estimation and has proven to be fairly robust against different kinds of speech distortion [6], including that of the telephone channel [7], and reasonably independent of the particular way in which computations may be implemented [8]. For these reasons, their application to automatic voice pathology detection has been proposed during the last years [9]. Yet, to authors' knowledge, up to now no physical explanation exists on the meaning of MFCC and their relevance on pathology detection.

Within this paper, a new scheme for automatic voice pathology detection is proposed. This lies half-way between usual cepstral domain and Mel-frequency cepstral domain. Namely, it takes profit from the conceptual interpretation of cepstral processing of speech signals [10], the pattern separation capability of cepstral distances [11] and the smoother spectrum estimation provided by the filter banks in MFCC calculation [11]. The mathematical formulation of both cepstrum and MFCC parameters is revised in Sect. 2, while the newly proposed set of parameters is introduced in Sect. 3. The results from the application of these features to the detection of pathologies on voices belonging to a commercial database are reported in Sect. 4. Last, the conclusions are presented in Sect. 5.

2 Mathematical Formulation

2.1 Short-Time Fourier Transform

As stated in previous section, the variability of speech signal is a key feature for pathology detection. The need for detecting such variability leads to the convenience of employing short-time techniques for speech processing. For this reason, in the following lines the mathematical framework for short-time processing of speech provided in [10] is revised.

Let $x[n]$ be a speech signal composed by N samples ($n = 0 \cdots N-1$) obtained at a sampling frequency equal to f_s; then it can be segmented in frames defined by:

$$f[n;m] = x[n] \cdot w[m-n] \quad , \tag{1}$$

where $w[n]$ is the framing window:

$$w[n] = 0 \;\; \text{if } n < 0 \text{ or } n \geq L \tag{2}$$

and L is the frame length. Consequently, $f[n;m]$ has non-zero values only for $n \in [m-L+1, m]$. If consecutive speech frames are overlapped a number of l_0 samples, then m may have the following values:

$$m = L + p \cdot (L - l_0) - 1 \quad , \tag{3}$$

where p is the frame index and it is an integer such that:

$$0 \leq p \leq \frac{N-L}{L-l_0} \quad . \tag{4}$$

Considering the relation between the frame shift m and the frame index p, frames without time shift reference may be renamed as:

$$\begin{aligned} g_p[n] &= f[n+m-L+1;m] = f[n+p\cdot(L-l_0);m] = \\ &= x[n+p\cdot(L-l_0)]\cdot w[(L-1)-n] \quad , \end{aligned} \tag{5}$$

where $n = 0 \cdots L-1$. From these speech frames, the short-term Discrete Fourier Transform (stDFT) is computed as:

$$S_p(k) = \sum_{n=0}^{N_{\mathrm{DFT}}-1} \widetilde{g}_p[n]\cdot e^{-j\cdot\frac{2\pi}{N_{\mathrm{DFT}}}\cdot kn} \quad , \tag{6}$$

where N_{DFT} is the number of points of the stDFT, $k = 0 \cdots N_{\mathrm{DFT}}-1$ and:

$$\widetilde{g}_p[n] = \begin{cases} g_p[n] & \text{if} \quad 0 \leq n < L \\ 0 & \text{otherwise} \end{cases} \quad . \tag{7}$$

Thus, if $N_{\mathrm{DFT}} \geq L$ then (6) is equal to:

$$S_p(k) = \sum_{n=0}^{L-1} g_p[n]\cdot e^{-j\cdot\frac{2\pi}{N_{\mathrm{DFT}}}\cdot kn} \tag{8}$$

and the frequency values that correspond to each stDFT coefficient are:

$$f_k = \begin{cases} f_s\cdot\frac{k}{N_{\mathrm{DFT}}} & \text{if} \;\; k \leq \frac{N_{\mathrm{DFT}}}{2} \\ f_s\cdot\frac{k-N_{\mathrm{DFT}}}{N_{\mathrm{DFT}}} & \text{if} \;\; k > \frac{N_{\mathrm{DFT}}}{2} \end{cases} \quad . \tag{9}$$

2.2 Short-Time Cepstrum

In [10], an algorithm for computing the short-time cepstrum from the stDFT is given, under the assumption that $N_{\mathrm{DFT}} >> L$:

$$c_p[q] = \frac{1}{N_{\mathrm{DFT}}}\cdot\sum_{k=0}^{N_{\mathrm{DFT}}-1} \log|S_p(k)|\cdot e^{j\cdot\frac{2\pi k}{N_{\mathrm{DFT}}}\cdot q} \quad . \tag{10}$$

A physical interpretation of cepstrum can be derived from the discrete-time model for speech production that can also be found in [10]. This model may be written in frequency domain as:

$$S\left(e^{j\Omega}\right) = E\left(e^{j\Omega}\right)\cdot G\left(e^{j\Omega}\right)\cdot H\left(e^{j\Omega}\right) \quad , \tag{11}$$

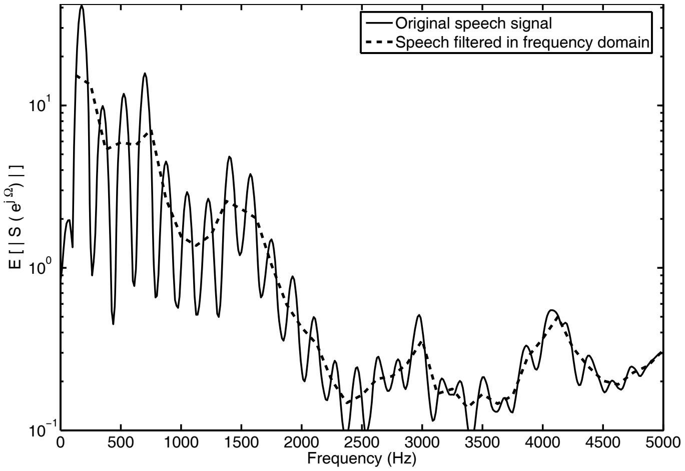

Fig. 1. Average modulus of the short-term DFT for one voice record.

where $S\left(e^{j\Omega}\right)$ is the speech, $E\left(e^{j\Omega}\right)$ is the impulse train corresponding to the pitch and its harmonics, $G\left(e^{j\Omega}\right)$ is the glottal pulse waveform that modulates the impulse train and $H\left(e^{j\Omega}\right)$ is, herein, the combined effect of vocal tract and lip radiation. These components can be appreciated in Fig. 1, which corresponds to the average modulus of the short-term DFT calculated from one of the voice records belonging to the database referred in Sect. 4.1.

The quick impulse-like variations in Fig. 1 correspond to the fundamental frequency and its harmonics $E\left(e^{j\Omega}\right)$ and the evolution of the impulse amplitude envelope is related both to the glottal waveform $G\left(e^{j\Omega}\right)$ and to the formants induced by the vocal tract $H\left(e^{j\Omega}\right)$. These formants correspond to the three envelope peaks with a decreasing level of energy that are centered at 750 Hz, 1375 Hz and 3000 Hz. In fact, these center frequencies are coherent with the range of typical values given in [2].

The logarithm operation in (10) converts the products in (11) into sums. Consequently, it allows the cepstrum to separate fast from slow signal variations in frequency domain. This widely known fact is illustrated in Fig. 2, where the peak around 5.7 ms clearly identifies the fundamental frequency (175 Hz) and the values below 2 ms correspond to the spectrum envelope.

2.3 Short-Time MFCC

Once the stDFT of a speech signal is available, another option for further processing, as mentioned in Sect. 1, is the calculation of short-time MFCC (stMFCC) parameters. For stMFCC computation, only the positive part of the frequency axis is considered [11], that is, $f_k \geq 0$ and, therefore, $k \leq N_{\mathrm{DFT}}/2$. In order to calculate stMFCC coefficients, a transformation is applied to the frequencies so as to convert them to Mel-frequencies

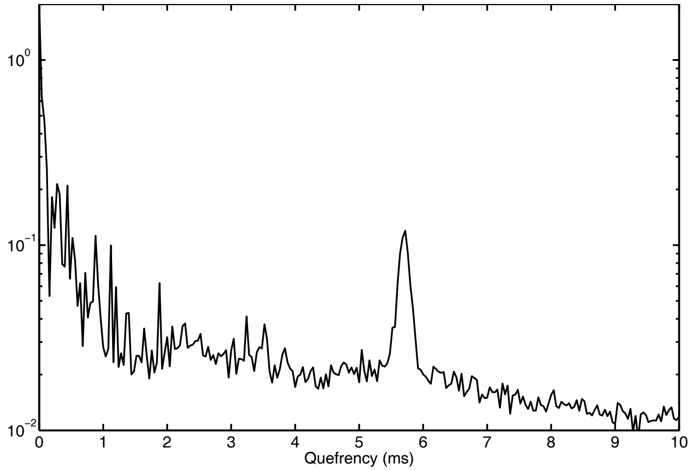

Fig. 2. Short term cepstrum averaged for all frames of the same voice record as used for figure 1.

f_k^m [9]:

$$f_k^m = 2595 \cdot \log_{10}\left(1 + \frac{f_k}{700}\right) \tag{12}$$

and the stDFT is further processed through band-pass integration along M equally long Mel-frequency intervals, being $M = \lfloor 3 \cdot \log_{10} f_s \rfloor$ ($\lfloor\cdot\rfloor$ means rounding to the previous integer). Namely, the i^{th} interval ($i = 1 \cdots M$) in Mel-domain is defined by:

$$I_i^m = \left[F^m \cdot \frac{i-1}{M+1}, F^m \cdot \frac{i+1}{M+1}\right] \quad , \tag{13}$$

where F^m is the maximum Mel-frecuency:

$$F^m = \max_k f_k^m = 2595 \cdot \log_{10}\left(1 + \frac{f_s/2}{700}\right) \tag{14}$$

and the interval length in Mel-domain is given by:

$$L\left(I_i^m\right) = \frac{2}{M+1} \cdot F^m \quad . \tag{15}$$

According to the previous equations, the N_{DFT} stDFT coefficients are transformed to M frequency components as follows:

$$\widetilde{S}_p(i) = \sum_{f_k \in I_i} \left(1 - \frac{\left|f_k^m - F^m \cdot \frac{i}{M+1}\right|}{L\left(I_i^m\right)/2}\right) \cdot \left|S_p(k)\right| \quad . \tag{16}$$

Last, the q^{th} ($q = 1 \cdots Q$) stMFCC of the p^{th} speech frame, where Q is the desired length of the Mel-cepstrum, is given by cosine transform of the logarithm of the smoothed “Mel-spectrum” [11]:

$$\widetilde{c}_p\left[q\right] = \sum_{i=1}^{M} \log\left|\widetilde{S}_p\left(i\right)\right| \cdot \cos\left[q \cdot \left(i - \frac{1}{2}\right) \cdot \frac{\pi}{M}\right] \quad . \tag{17}$$

3 Cepstral Coefficients Based on Smoothed Spectrum

3.1 Justification

As stated in Sect. 1, while MFCC parameters exhibit both good performance and robustness in feature extraction from speech, they lack a clear physical interpretation. On the opposite, cepstrum has a physical meaning (recall Sect. 2.2), yet raw cepstrum coefficients are not as useful for speech parametrisation. In the next paragraphs, the reasons for these facts are exposed.

Cepstrum calculation, as formulated in (10), is based on the spectrum estimate provided by the absolute value of the stDFT. Due to the logarithm, this gives a result that is proportional to the case of periodogram-based spectrum estimation. However, such estimation is very dependent on the specific values of the original speech frame. A more robust spectrum estimate can be obtained by smoothing of the periodogram (Blackman and Tukey method, [12]). In fact, this is what (16) expresses in the calculation of MFCC. Therefore, filtering of the stDFT may be assumed to be one of the sources of MFCC robustness.

In contrast, an explanation for the lack of clear interpretation of MFCC also lies in the meaning of (16). According to that equation, stDFT smoothing for MFCC computation is carried out with a variable-length filter, that is, a Bartlett window whose length decreases for lower frequency bands. Moreover, the smoothed stDFT is downsampled to obtain only M samples in the interval $[0, f_s/2]$ that are not uniformly spaced [11]. While the downsampling is positive in the sense that it reduces the dimensionality of the problem, its non-uniformness, together with the previous variable-length filtering, obscures the interpretation of the output of the cosine transform in (17).

From the previous reasoning, if stDFT is smoothed with a fixed-length filter and its output is uniformly decimated prior to the logarithm computation, the cepstral coefficients in (10) can be transformed to a more robust parameter set. Moreover, this is achieved while keeping the physical meaning of cepstrum, since the output of the first operation gives an improved spectrum estimate and the second only limits the length of cepstrum in quefrency domain.

3.2 Formulation

Starting from (8), if the stDFT modulus is smoothed with a Bartlett window of constant length equal to Δf then the following output is obtained:

$$S'_p\left(i\right) = \sum_{f_k \in I_i} \left(1 - \frac{\left|f_k^m - i \cdot \Delta f/2\right|}{\Delta f/2}\right) \cdot \left|S_p\left(k\right)\right| \quad , \tag{18}$$

where $I_i = [\Delta f \cdot (i-1)/2, \Delta f \cdot (i+1)/2]$ and the Bartlett window has been chosen for similarity with (16). Herein, only the positive part of the frequency axis has been considered, as in Sect. 2.3.

If the filtered stDFT is decimated so as to keep only the outputs of consecutive windows with a 50% overlap, this is equivalent to decimation by a factor:

$$D = \lfloor \Delta f \cdot N_{\mathrm{DFT}} / (2 \cdot f_s) \rfloor \quad . \tag{19}$$

The modified cepstrum then becomes:

$$c'_p[q] = \frac{D}{N_{DFT}} \cdot \sum_{k=0}^{\frac{N_{DFT}}{2 \cdot D}} \log \left| S'_p(k \cdot D) \right| \cdot \cos\left((k-1) \cdot \frac{2\pi D}{N_{DFT}} \cdot q \right) \quad , \tag{20}$$

where only the positive frequencies have been considered, hence computing the inverse DFT as a cosine transform as in (17). $c'_p[q]$ has the twofold advantage over $c_p[q]$ of being based on a smoother spectrum estimate $S'_p(i)$ and having a period length that has been reduced by a factor D, thus providing some dimensionality reduction.

3.3 Cepstral Distances

Differences in cepstrum can be used for speech signal classification. An example of such usage is the definition of the cepstral distance in [11] as the norm of the vector resulting form substraction of the two cepstra to be compared. This, if directly applied to pathology detection, would result in comparing the cepstrum of consecutive speech frames so as to assess the variability of the signal. Mathematically:

$$d_p^2 = \sum_{q=0}^{\frac{N_{\mathrm{DFT}}}{D}-1} \left| c'_{p+1}[q] - c'_p[q] \right|^2 \quad . \tag{21}$$

However, bearing in mind the physical interpretation of cepstrum, this definition has the drawback of mixing pitch variations with formant and glottal pulse variations. To overcome this problem an individual frame-to-frame cepstral parameter variation analysis is proposed:

$$d_p[q] = \left| c'_{p+1}[q] - c'_p[q] \right| \quad . \tag{22}$$

This way, analysis of the distribution of $d_p[q]$ related to speech formant and glottal pulse variability (low values of q) can be isolated from pitch changes associated to values of q around the pitch period.

4 Application and Results

For the purpose of performance analysis, the modified cepstral parameters presented in previous section have been applied to the problem of automatic pathology detection on recorded voice. The results have been compared to those produced by MFCC. Within this section, first the voice database is presented, second the used parameter set is specified, third the classifier is described and, last, the results are shown and commented.

4.1 Database

The voice records used in this investigation are the same as in [13]. They belong to a database distributed by the company Kay Elemetrics [14]. The recorded sounds correspond to sustained phonations (1-3 s long) of the vowel /ah/ from patients with either normal or disordered voice. Such voice disorders belong to a wide variety of organic, neurological, traumatic and psychogenic classes. Sampling rate of speech records has been made uniform for all of them and equal to 25 KHz, while the coding has a resolution of 16 bits. The subset taken from the database contains 53 normal and 173 pathological speakers which are uniformly distributed in age and gender [13].

4.2 Classifier Description

For both classification schemes, a Multilayer Perceptron (MLP) with two hidden layers, each consisting of 4 neurons, and a two-neuron output layer has been used as a classifier. All neurons have logistic activation functions. An MLP with a single hidden layer having 50 neurons was utilised in [9]. The structure herein proposed, in contrast, has less free parameters, thus allowing a faster learning, and the reduced number of neurons is compensanted by the introduction of an additional hidden layer that permits learning of more complex relations [15].

The MLP classifier has been trained with 60% of available speech records in such a way that its output is expected to be “1” for pathological voices and “0” for normal voices. 10% of the records have been used for cross-validation during the training phase as a criterion to stop training. The remaining 30% of records have been used for testing. The experiment has been repeated 200 times, each of them with different, randomly chosen, training and cross-validation sets.

4.3 Results for short-time parameters

Within the previous classification scheme, each feature vector, corresponding to one speech frame, is assigned a pair of likelihoods, one coming from each output neuron:

- likelihood of belonging to the phonation of a healthy person (l^p_{nor}) and
- likelihood of that person having pathology (l^p_{pat});

being p the frame index. From the pair ($l^p_{\mathrm{nor}}, l^p_{\mathrm{pat}}$), the classification decision at frame level is taken based on the value of the log-likelihood ratio (LLR^P) and comparing it to a threshold θ:

$$LLR^p = \log \frac{l^p_{\mathrm{nor}}}{l^p_{\mathrm{pat}}} \gtrless \theta \ . \tag{23}$$

For a record consisting of P frames, decision at record level is taken based on the mean log-likelihood:

$$LLR = \frac{1}{P} \cdot \sum_{p=1}^{N} \log \frac{l^p_{\mathrm{nor}}}{l^p_{\mathrm{pat}}} \gtrless \theta \ . \tag{24}$$

In this first experiment, for each speech record short-term cepstrum-based coefficients, as defined in (20), have been calculated. Namely, a filter length $\Delta f = 200$ Hz

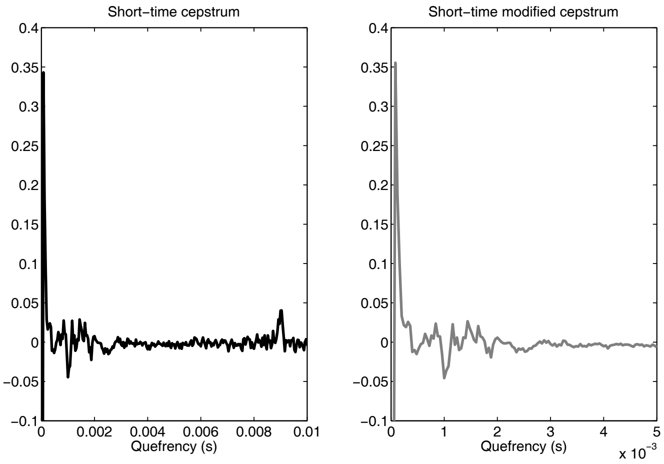

Fig. 3. Short-time cepstrum of a speech frame taken from the database, without spectrum filtering (left) and after spectrum filtering (right). It can be noticed that the limitation on the length of cepstrum has produced a loss of information about pitch (peak on the right part of the left graph).

has been chosen for sfDFT smoothing. Consequently, this results in a cepstrum length equal to $(f_s - \Delta f/2)/(\Delta f/2) = 124$ samples. The choice of Δf is consistent to the approximate length of the low-band filters used for MFCC calculation (recall (16)). At first sight, however, it has the drawback of loosing pitch information of the signal spectrum. This is illustrated in Fig. 1 where the filtered DFT has been plotted with a dashed line and also in Fig. 3 where the cepstrum obtained with and without spectral smoothing is represented. Nevertheless, such filtered spectrum contains information on both harmonic-to-noise ratio (HNR) and glottal pulse waveform [16] and HNR is a useful parameter for pathology detection that is closely related to both frequency and amplitude perturbations of pitch [2].

For the sake of comparison, another classifier based on a parameter vector consisting of 20 stMFCC calculated using 31 Mel-band filters ($M = \lfloor 3 \cdot \log f_s \rfloor = 31$) has also been tested. Figure 4 shows the detection-error-tradeoff (DET) plots [17] for both parametrisation schemes at frame level and at record level. It can be noticed that while the stMFCC-based system provides better performance (14.96 % equal error rate - EER), possibly due to the higher dimensionality reduction, the herein presented scheme has a performance within the same range (17.79 % EER) with a clearer physical interpretation. Another observation that can be drawn from the plot is that results at record level (respectively, 13.36% and 15.40% EER) are better that at frame level. Such fact indicates the presence of a certain degree of variability among speech frames belonging to the same record. This is intended to be confirmed in the second experiment, as reported next.

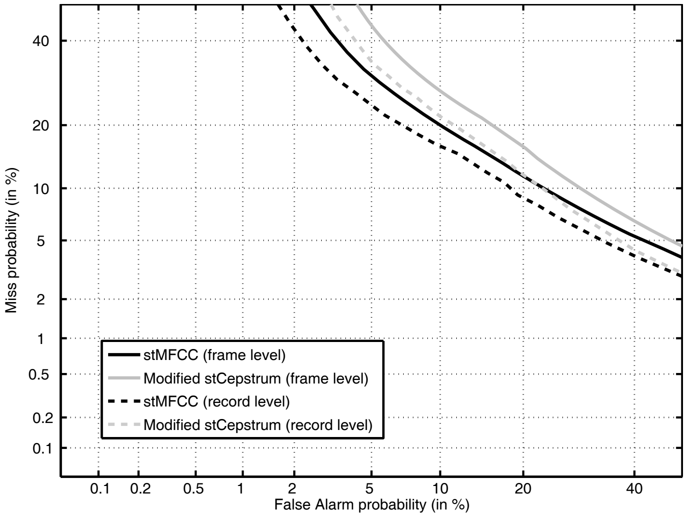

Fig. 4. DET plot for the stMFCC (black) and the short-time modified cepstrum (gray) parametrisation schemes.

4.4 Results for averaged parameters

In order to assess the relevance of cepstral variability, a second experiment has been carried out. In this case, the input vectors for the pathology detector are calculated for each record, instead of working at the frame level as before. The first 124 elements correspond to the average values of the short-time cepstrum, as calculated before. The rest of the input vector contains information about the variability of cepstrum around those average values. More specifically, the mean and variance of $d_p[q]$ for each value of q are used as descriptors of the cepstrum variability. Therefore, on the whole, a parameter vector of 124×3 elements is produced.

Figure 5 shows the results of using the above-mentioned scheme for speech record parametrisation compared to thise obtained only with the first 124 components of the feature vectors, that is, without including information on cepstral variability. It should be noted that the structure of the MLP classifier for this experiment has been simplified by removing one of the hidden layers. The reason for this is that when passing from the frame level to the record level a great reduction on the number of feature vectors is obtained. Still, a similar performance at record-level classification is achieved for the case of the feature vectors including information on cepstral variability (14.70 % EER). However, if this information is removed from the classifier inputs, the performace is degraded (19.17 % EER). This confirms the relevance of cepstral variability for pathology detection.

In order to acquire a deeper understanding of the reasons for these results, an analysis of the relevance of cepstrum-based parameters for speech classification as either pathological or not has been realised. Such analysis is based on the evaluation of the Fisher criterion [18] for each individual parameter of the above-described 372-element feature vectors. The results, differentiated for the three subsets of parameters (modified

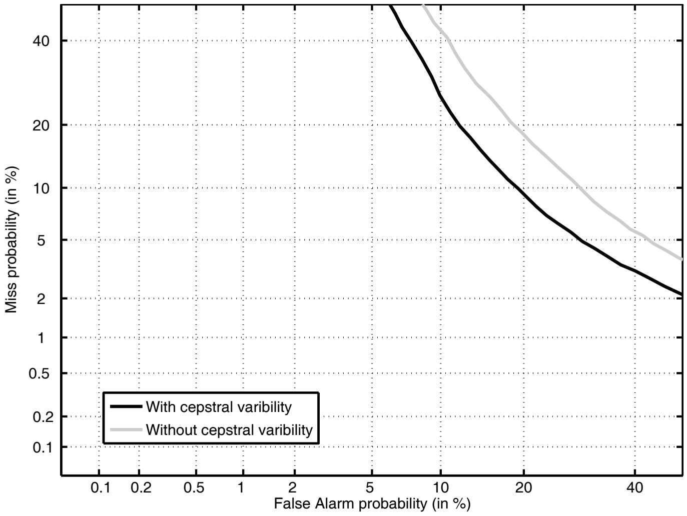

Fig. 5. DET plot for short-time cepstral parameters including information on cepstral variability (black) and lacking that information (gray).

cepstrum, variance of differences and average of absolute differences) are plotted in Fig. 6.

According to this plot, the most relevant cepstral parameters for pathology detection maybe roughly classified into two groups:

- The modified cepstrum values with lowest indices (plot at the bottom of Fig. 6): these are related to the slowest components of the spectrum envelope in Fig. 1, which, on their side, are associated to spectral noise levels and HNR [16].
- The frame-to-frame variations in cepstrum-based coefficients whose quefrecies are within the interval $[0.5, 1.5]$ miliseconds approximately: coefficients within that interval correspond to the short frequency range components of the spectrum envelope. These components, as justified in Sect. 2.2, are related to glottal waveform and speech formants. However, this information itself does not help to discriminate the presence of pathology, as indicated by the low values of the Fisher criterion in the bottom plot of Fig. 6. Instead, frame-to-frame variations of these factors are much more relevant, as depicted in the other two plots of the same Fig..

To be more specific, since the voice records of the database used for this experiment correspond to sustained vowel phonations, it can be assumed that the vocal tract has very little variations, hence formants do not change and the second group of parameters should be more closely related to changes in the glottal waveform. As for the limits of the quefrency interval in which parameters from the second group are relevant, the lower limit of 0.5 ms corresponds to the quefrency band that separates slow components of the spectrum envelope (first group of parameters) from faster components (associated to the second set); on the other hand, the upper limit of 1.5 ms corresponds to the highest quefrency range at which the modified cepstrum $c'_p[q]$ has significant values. This is

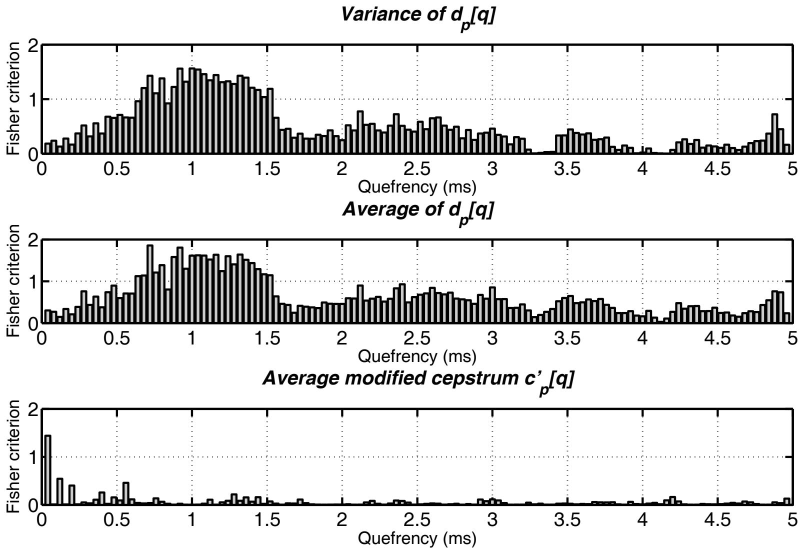

Fig. 6. Value of Fisher criterion for each cepstral parameter.

shown in Fig. 7, where a plot of the frame-averaged modified cepstrum of one voice record is depicted.

5 Conclusions

Speech parametrisation in cepstral domain is a useful technique for automatic pathology detection. Specifically, MFCC have been successfully used for this purpose. While the computation of these parameters has an intrinsic robustness due to its independency from pitch extraction and the spectrum filtering, their physical interpretation is obscure because of the non-linear Mel-frequency transformation.

Within this paper an alternative set of cepstrum-based parameters has been proposed. Such parameters share the robustness of MFCC since they do not require pitch estimation and filtering of the estimated speech spectrum is also performed. In contrast to MFCC, the calculation of these newly proposed parameters does not involve any non-linear frequency transformation and, consequently, their physical interpretation remains clear. Namely, their values have been shown to be related to the amount of noise energy present in speech and the glottal waveform variability. Both factors are directly associated to laringeal pathologies.

Finally, the performance of the proposed cepstral parameters for pathology detection has been tested using a MLP classifier and results have been compared to those of MFCC. The obtained misclassification rates indicate that the performances of both sets of parameters are similar. Moreover, a deeper analysis on the individual impact of each parameter on the classification task has revealed that the most relevant parameters are

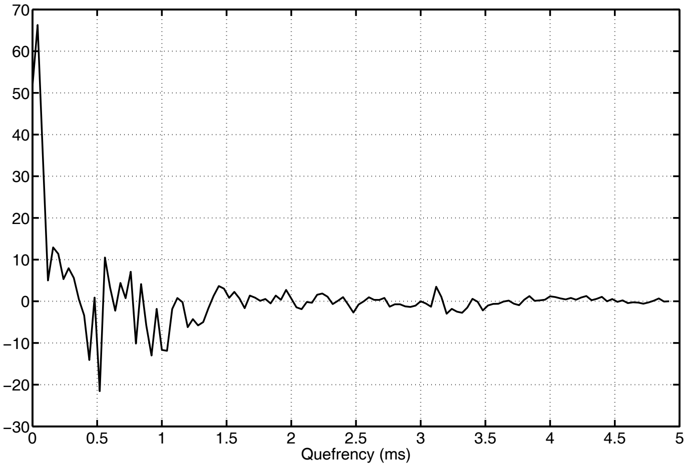

Fig. 7. 124 modified cepstral parameters from one of the database's voice records.

those more closely linked to the above-mentioned two factors: noise energy and glottal wave variations.

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).

References

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

[2] Jackson-Menaldi, M.C.A.: La voz patológica. Editorial Médica Panamericana, Buenos Aires (Argentina) (2002)

[3] Godino-Llorente, J.I., Sáenz-Lechón, N., Osma-Ruiz, V., Aguilera-Navarro, S., Gómez-Vilda, P.: An integrated tool for the diagnosis of voice disorders. Medical Engineering & Physics **28**(3) (March 2006) 276–289

[4] Deliyski, D.D.: Acoustic model and evaluation of pathological voice production. In: Proceedings of the 3^{rd} Conference on Speech Communication and Technology (EUROSPEECH'93), Berlin (Germany) (1993) 1969–1972

[5] Boyanov, B., Ivanov, T., Hadjitodorov, S., Chollet, G.: Robust hybrid pitch detector. IEE Electronics Letters **29**(22) (October 1993) 1924–1926

[6] Bou-Ghazale, S.E., Hansen, J.H.L.: A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Transactions on Speech and Audio Processing **8**(4) (July 2000) 429–442

[7] Fraile, R., Godino-Llorente, J.I., Sáenz-Lechón, N., Osma-Ruiz, V., Gómez-Vilda, P.: Analysis of the impact of analogue telephone channel on mfcc parameters for voice pathology detection. In: 8^{th} INTERSPEECH Conference (INTERSPEECH 2007), Antwerp (Belgium) (2007) 1218–1221

[8] Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative evaluation of various MFCC implementations on the speaker verification task. In: Proceedings of the 10^{th} International Conference on Speech and Computer (SPECOM 2005), Patras (Greece) (2005) 191–194

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

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

[11] Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs (USA) (1993)

[12] Proakis, J.G., Manolakis, D.G.: Digital Signal Processing. Principles, Algorithms and Applications. 3^{rd} edn. Prentice-Hall International, New Jersey (USA) (1996)

[13] Godino-Llorente, J.I., Gómez-Vilda, P., Blanco-Velasco, M.: 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) (October 2006) 1493–1953

[14] Kay Elemetrics Corp.: Disordered voice database.version 1.03 (1994)

[15] Haykin, S.: Neural Networks: a comprehensive foundation. 1^{st} edn. Macmillan College Publishing Company, New York (USA) (1994)

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

[17] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Proceedings of the 5^{th} Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes (Greece) (1997) 1895–1898

[18] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. 2^{nd} edn. John Wiley & sons, New York (USA) (2001)