Evaluation of the Effect of the GSM Full Rate codec on the Automatic Detection of Laryngeal Pathologies Based on Cepstral Analysis

Rubén Fraile, Carmelo Sánchez, Juan I. Godino-Llorente, Nicolás Sáenz-Lechón, Víctor Osma-Ruiz, Juana M. Gutiérrez
Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid, Madrid, Spain
rfraile@ics.upm.es

Abstract

Advances in speech signal analysis during the last decade have allowed the development of automatic algorithms for a non-invasive detection of laryngeal pathologies. Bearing in mind the extension of these automatic methods to remote diagnosis scenarios, this paper analyzes the performance of a pathology detector based on Mel Frequency Cepstral Coefficients when the speech signal has undergone the distortion of a speech codec such as the GSM FR codec, which is used in one of the nowadays most widespread communications networks. It is shown that the overall performance of the automatic detection of pathologies is degraded less than 5 %, and that such degradation is not due to the codec itself, but to the bandwidth limitation needed at its input. These results indicate that the GSM system can be more adequate to implement remote voice assessment than the analogue telephone channel.

Index Terms: speech analysis, speech coding, voice function assessment

1. Introduction

Research in the automatic processing of voice recordings for the detection of laryngeal pathologies in speakers is ongoing for more than one decade now [1]. The performance evaluation of proposed processing algorithms is usually done using databases of recordings made in clinical environments using tools as the ones described in [2] and [3]. In this context, the primarily foreseen application of voice recording is, precisely, its onsite analysis through the use of those tools. However, at present remote speech and voice assessment and treatment is becoming feasible [4] and it has clear potential benefits.

The feasibility of remote voice assessment, from the technical point of view, 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 analogue telephone has been shown to experiment a performance degradation figure around 15% when detection is based on traditional acoustic parameters [5]. In contrast, if analysis is based on cepstral parameters, performance degradation can be kept around 10% and success rates near 80% are achievable [6] [7].

Secondly, the impact of several digital speech codecs on the voice quality of disordered speech has been studied during the last years. Specifically, [8] shows that speech coding with data compression beyond 64 kbps with the MP3 standard leads to degraded performance in pathology detection. Additionally, the parallel studies reported in [9] and [10] analyze the impact of some common codecs used in communication systems on the quality of atypical speech. The comprehensive set of experiments reported therein showed that the codecs achieving higher compression rates tend to distort very differently normal and pathological voices, since their algorithms are based on assumptions whose validity is restricted to normal speech [10]. Those experiments also indicated that, for some codecs, the output speech resembles more a disordered speech than the input, no matter if it was recorded from a healthy speaker [9]. Remarkably, CELP and GSM RPE-LTP codecs yielded the best performance in both cases, the reason for this being that they do not use a parametric approach for describing the residual left after linear prediction (LP) modeling of speech.

Bearing in mind the remote voice assessment framework mentioned before, this paper reports on an analysis of the impact of speech coding and decoding using the GSM 6.10 full rate (FR) codec on the separability between normal and pathological voices. Therefore, this study represents a step forward with respect to [5], [6] and [7] in the sense that it analyzes a digital communication scheme. On the other hand, it evaluates quantitatively the effect of the GSM 6.10 FR codec, thus giving a complementary view to [9] and [10]. The rest of the paper is organized as follows: section 2 contains the specific formulation of the cepstral analysis used for speech parameterization, section 3 provides a brief overview of the GSM 6.10 FR codec, 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. Mel frequency cepstral coefficients

For the herein reported study, speech has been parameterized by means of the short-time Mel frequency cepstral coefficients (MFCC), according to the definition given in [11]. While such definition is slightly different from the original proposal in [12], it has an easier interpretation. This change is also justified by the inherent robustness of MFCC against implementation changes [13]. As for the choice of short-term analysis, this is justified by the relevance of signal variability for the detection of laryngeal pathologies by means of cepstral analysis [14]. The specific formulation is as follows:

\[
c_{p}[q] = \frac{1}{M+1} \sum_{k=1}^{M} \log |\tilde{S}_{p}(k)| \cdot \cos \left( \frac{\pi k}{M+1} \cdot q \right)
\]

(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 $\tilde{S}_p(k)$ is the estimate of the spectral energy in the $k^{th}$ Mel band. Specifically:

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

where $S_p(i)$ is the $i^{th}$ element of the short-time discrete Fourier transform (DFT) of the $p^{th}$ speech frame, $F_m^k$ is its associated Mel frequency, $f_i^m = \left[ F_m^k \cdot \frac{k-1}{M+1}, F_m^k \cdot \frac{k+1}{M+1} \right]$ 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 ($f_s$) of the speech signal. The frequency transformation that allow passing from linear to Mel scale is:

$$f^m = 2505 \cdot \log_{10} \left( 1 + \frac{f}{700} \right) \quad (3)$$

Speech frame duration has been chosen to be 20 ms, as in [15], which allows capturing the spectral envelope of speech for fundamental frequencies above 50 Hz, thus covering the cases of both male and female voices. Overlap between consecutive frames was 50%. Vectors of 21 MFCC, that is $q \in [0, 20]$, have been used as feature vectors for each speech frame.

3. GSM 6.10 FR codec

The full rate GSM speech codec is described in the 3GPP technical specification 46.010 (formerly GSM 6.10) [15]. Its basic structure is depicted in figure 1. The codec is based on a short-term LP block that produces linear prediction coefficients (LPC) for each 20 ms speech frame. The residual prediction error given by the LP block is coded in the second block using long term prediction (LTP) analysis. Both steps are preceded by a preprocessing stage that removes the offset and performs pre-emphasis of the speech signal. Another relevant aspect of the whole system is that it needs the input to be speech digitized at a sampling frequency equal to 8000 Hz and quantized with 13 bits. For the work reported in this paper, the free software SoX [16] has been used for performing GSM coding and decoding.

4. Simulation procedure

4.1. Database

All the herein reported results have been obtained using a database distributed by Kay Elemetrics [17]. Specifically, the utilized speech records correspond to sustained phonations of the vowel /ah/ (1-3 s long) from patients with normal voices and others having diverse voice disorders. The subset taken corresponds to that reported in [18] 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 had sampling rates equal to either 50 or 25 kHz with 16 bits of resolution. Down-sampling with previous half band filtering was realised on 50 kHz 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 with 40 hidden nodes having logistic activation functions (as in [7]) and two outputs with linear activations. The use of two linear outputs allows obtaining two values for each speech frame, characterized by its MFCC vector $c_p$. 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 $L_{nor}(c_p)$ (first output) or pathological $L_{pat}(c_p)$ (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 [19]:

$$\log \left[ L_{nor}(c_p) \right] - \log \left[ L_{pat}(c_p) \right] > \theta \quad (4)$$

If the previous condition is met, then the speech frame is classified as normal, if not, it is considered pathological. Ideally that is, if the likelihoods could be perfectly estimated by the classifier, the value for the threshold $\theta$ should be $\theta = 0$. However, in practice the choice of $\theta$ 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:

$$\frac{1}{N_{frames}} \cdot \sum_{p=1}^{N_{frames}} \log \left[ L_{nor}(c_p) \right] - \log \left[ L_{pat}(c_p) \right] > \theta \quad (5)$$

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

4.3. Testing protocol

The testing of each detection scheme consists of repetitive sequence of experiments process. Within each experiment 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. The remaining (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 have been computed. Forty-five repetitions of the experiment with independently chosen training, validation and testing sets have been carried out.

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 [19]. Among these indicators, the DET plot 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 refers to 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 $\theta$ in (4) and (5) changes. 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 $\theta$ is tuned so that both coincide. In all the experiments results have been computed both at frame and record levels, corresponding to (4) and (5).

5.1. Effect of band limitation

As indicated in section 3, the GSM FR codec needs the input speech to be sampled at 8000 Hz and quantized with 13 bits. Bearing in mind that the available data has been sampled at 25 kHz and quantized with 16 bits, the preprocessing needed before performing the GSM coding involves restricting both bandwidth and the number of quantization levels. This implies a loss of information that should affect the performance of the automatic detection of pathologies. The first step in the herein reported analysis consists in evaluating such performance degradation. For this purpose, original speech records have been parameterized using 21 MFCC calculated with $M_d = 3 \cdot \log f_s = 3 \cdot \log 25000 \approx 31$ Mel filters (as recommended in [11]). In parallel, they have been converted to a sampling frequency of 8000 Hz by means of interpolation, filtering and decimation. Filtering has been done in frequency domain by elimination of the corresponding coefficients of the DFT in order to avoid distortion. The resulting records have been parameterized also with 21 MFCC, but calculated with $M = 3 \cdot \log 8000 \approx 25$ Mel filters. Figure 2 shows the DET plots obtained at frame and record level, for both original and downsampling records. These plots indicate a significant degradation in the performance that is not due to the GSM FR codec itself, but to the limitation of bandwidth needed at its input. Such performance degradation represents approximately a 3\% increase in the median EER, as can be noticed in figure 3.

5.2. Effect of quantization and coding

After assessing the effect of bandwidth reduction, the following steps in the analysis have been the evaluation of the impact of the reduction in the number of quantization bits and the effect of the GSM FR coding and decoding. It has been found that these to processes have a minor effect on the performance of the automatic classification when compared to the effect of bandwidth reduction. In fact, the obtained distributions of EER are not significantly different, as indicated by the overlap of the notches of boxes in figure 4. Another view of the same results is depicted in the DET plot of figure 5. Such plot is almost the same as the one in figure 2, hence highlighting the little effect of GSM FR coding/decoding on the classification performance.

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 produced by the GSM FR codec has been analyzed. The use of such codec for transmitting speech through a GSM wireless network implies a signal distortion intrinsic to the codec design, but also a predistortion needed to adjust the sampling frequency of the signal and the number of quantization bits to the requirements for the codec input. The reported results indicate that the bandwidth reduction is the main factor affecting the performance of the automatic detection of pathologies, while the coding itself and the reduction in the number of quantization bits have not any significant effect.

Quantitatively, the effect of bandwidth reduction is an increase of approximately 3\% in the median EER. This figure contrasts to the 10\% degradation reported for the analogue telephone channel [7]. While in this study the effect of the communications channel has not been accounted for, the specific design of the GSM system should minimize such effect. The reason for this is that speech frames with errors are either discarded or substituted with properly received frames [20]. As the frame length in GSM is 20 ms, the same used for this study, even in...
cases of frame transmission errors, only correct frames would be processed by the pathology detector.

The obtained results are fully coherent with previous studies that indicated, on the one hand, that the GSM RPE-LTP codec was among the ones that yielded best results in terms of voice quality for atypical speech [9] [10]; herein, it has been shown that quantitatively, except for the bandwidth limitation, the effect of the codec on the performance of a MFCC-based automatic pathology detector can be disregarded. On the other hand, the bandwidth limitation has an effect that is much less important than the bandwidth limitation imposed by the analogue telephone channel [7]. The reason for this is that a sampling frequency equal to 8000 Hz allows maintaining the most significant frequency band for pathology detection [21] and the GSM FR codec does not distort the 0-300 Hz band, as the analogue telephone channel does. Therefore, GSM and, in general, digital systems have the potential of serving better remote voice assessment than the traditional analogue telephone.

7. Acknowledgements

This research was carried out within a project funded by the Ministry of Science and Technology of Spain (TEC2006-12887-C02) and in the context of European COST Action 2103.

8. References

[20] Full rate speech; Substitution and muting of lost frames for full rate speech channels, 3GPP TS 46.011, Rev. 8.0.0, Dec. 2008.