Analysis of the Impact of Analogue Telephone Channel on MFCC Parameters for Voice Pathology Detection


Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid, Spain
rfraile@ics.upm.es

Abstract

In this paper, the feasibility of a system developed for the remote diagnosis of voice pathologies is analysed. More specifically, the performance of MFCC-based pathology detectors over speech transmitted through an analogue telephone channel is studied. Results indicate that MFCC are voice features fairly robust to amplitude distortion and almost insensitive to phase distortion, but the efficiency of a voice pathology detector based on these features is clearly decreased when the speech samples are transmitted through a telephone channel.

Index Terms: Voice disorders, telephone channel, pattern classification.

1. Introduction

Automatic voice pathology detection has been approached in different ways during the last years. Many implementation proposals rely on speech features extracted after pitch detection (e.g. [1]). While fairly good results have been achieved with these methods, their performance degrades in the presence of signal distortion [2], partially because of their dependence on the performance of pitch detection algorithms under the same circumstances. This dependence can be avoided if an alternative approach is undertaken. Voice classification based on Mel Frequency Cepstral Coefficients (MFCC) is an alternative and its validity for pathological voice detection has already been demonstrated [3].

In parallel, remote diagnostics based upon biological signals transmitted through some communication channel is one of the expected applications of telemedicine technology for the next years [4]. Specifically, laryngeal disease detection on voice transmitted over an analogue telephone channel should be one of the first applications to be developed, since it relies on an already existing and widespread communication technology. However, pathology detection on voice transmitted over the phone must be implemented using methods that are robust against the significant signal distortion that happens on this type of channel [2].

Within this paper, an approach to analyse the impact of the telephone channel on automatic pathology detection on voice is presented. Namely, MFCC parameters have been chosen for speech feature extraction. This kind of parameters have been tested to produce highly separable patterns in pathological voice detection [3] and, from another point of view, they are a priori less prone to errors in the presence of signal distortion than pitch-based methods. Parameters have been extracted for classification from the same data records as reported in [3].

The telephone channel can be modelled as a band-pass system that introduces both amplitude and phase distortion in the speech signal. For this work, it has been modelled by software in discrete-time domain as a cascade association of a linear-phase band-pass FIR filter that models amplitude distortion and an all-pass HR filter that models phase distortion [5]. Such separation allows independent modelling of amplitude and phase distortions, hence simplifying subsequent analysis.

The paper is organised as follows: section 2 provides a brief outline of the implemented analogue telephone channel model; section 3 refers to the speech database that has been used; parameters for these records have been extracted according to MFCC definitions reviewed in section 4; in section 5, the impact of the phone channel over the MFCC features is analytically studied; simulation results are presented in section 6 and, last, section 7 contains the conclusions.

2. Telephone channel model

Signal transmitted through an analogue telephone channel suffers from different types of distortion whose magnitudes are limited by ITU recommendation G.120 [6]. Among all types of distortion, the most significant ones, apart from noise, are amplitude and phase distortion. As for amplitude, the channel behaves as a band-pass filter. Attenuation of high frequencies comes from the low-pass behaviour of the transmission line while attenuation of low frequencies allows the use of out-of-band signalling. Limits recommended by [6] are represented as continuous lines in figure 1.

Regarding phase distortion, [6] 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 2.

![Amplitude distortion of telephone channel](image)

Figure 1: Amplitude distortion of telephone channel.

Simulation of the phone channel with both amplitude and phase distortion can be realised by a two-stage filter, as proposed in [5] (figure 3). Within such a setup, the first stage
models amplitude distortion and the second one, phase distortion. Since voice data utilised for testing comes from a digitised record database, as described in next section, a straightforward way to model a band-pass linear-phase system is to design a symmetric FIR filter. For this purpose, and bearing in mind restrictions of [6], a 176-order filter has been designed that has the frequency response plotted in figure 1 (dashed line).

The obtained frequency-dependent group delay is depicted in figure 2. Maximum phase distortion happens at the limits of the pass band of the FIR filter, as specified by [6].

![Figure 2: Phase distortion of telephone channel.](image)

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 function:

\[
H(z) = H_p(z; f_1)H_m(z; f_m)
\]

where \( r = 1.01 \), \( f_1 = 250 \text{ Hz} \) and \( f_m = 3450 \text{ Hz} \). The obtained frequency-dependent group delay is depicted in figure 2. Maximum phase distortion happens at the limits of the pass band of the FIR filter, as specified by [6].

![Figure 3: Simulated telephone channel model.](image)

### 3. Speech database

The voice records used in this investigation are the same as in [3]. They belong to a database distributed by the company Kay Elemetrics [7]. The recorded sounds correspond to sustained phonation (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 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 [3].

### 4. Computation of MFCC parameters

According to the mathematical framework for short-time processing of speech provided in [8], a speech signal composed by \( N \) samples \((n=0...N-1)\) obtained at a sampling frequency equal to \( f_s \) can be segmented in frames defined by

\[
f[n;m] = x[n]w[n-m]
\]

where \( w[n] \) is the framing window:

\[
w[n] = \begin{cases} 0 & \text{ if } n < 0 \text{ or } n \geq L \\ 1 & \text{ otherwise} \end{cases}
\]

and \( L \) is the frame length. Consequently, \( f[n;m] \) has non-zero values only for \( n \in [m-L+1,m] \). If consecutive windows are overlapped a number of \( L_0 \) samples, then possible values for \( m \) are:

\[
m = L + p(L-L_0)-1\]

being \( p \) an integer such that

\[
0 \leq p \leq \frac{N-L}{L-L_0}
\]

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

\[
g_p[n] = f[n+m-L,L-m] = f[n+p(L-L_0),m] = x[n+p(L-L_0)]w[L(L-1)-n]
\]

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

\[
S_p(k) = \sum_{n=0}^{N-1} g_p[n]e^{-j2\pi nk/N}
\]

where \( k=0...N_{\text{FFT}}-1 \), \( N_{\text{FFT}} \) is the number of points of the stDFT and:

\[
g_p[n] = \begin{cases} g_p[n] & \text{ if } 0 \leq n < L \\ 0 & \text{ if } L \leq n < N_{\text{FFT}} \end{cases}
\]

The frequency values that correspond to each DFT coefficient are:

\[
f_k = \begin{cases} f_s \frac{k}{N_{\text{FFT}}} & \text{ if } k \leq N_{\text{FFT}}/2 \\ f_s \frac{k-N_{\text{FFT}}}{N_{\text{FFT}}} & \text{ if } k > N_{\text{FFT}}/2 \end{cases}
\]

The stDFT is further processed through band-pass integration along \( M \) equally long Mel-frequency intervals, and so on to convert them to Mel-frequencies \( f_p^m \) [3]:

\[
f_p^m = 2595 \log_{10} \left( 1 + \frac{f_k}{700} \right)
\]

For MFCC computation, only the positive part of the spectrum is considered. Consequently, only the positive part of the frequency axis will be considered hereon [9], that is, \( f_1 \geq 0 \) and, therefore, \( k \leq N_{\text{FFT}}/2 \). In order to calculate MFCC coefficients, a transformation is applied to the frequencies so as to convert them to Mel-frequencies \( f_p^m \) [3]:

\[
f_p^m = 2595 \log_{10} \left( 1 + \frac{f_k}{700} \right)
\]
being \( M = \lceil 3 \log f_s \rceil \) (\([\cdot]\) means rounding to the previous integer):

\[
\tilde{S}_p(i) = \sum_{k=0}^{N-1} \left( 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2} \right) \hat{S}_p(k)
\]

where \( I_i \) is the \( i \)-th interval, as transformed into usual frequency-domain, \( f_i^n \) is its centre frequency in Mel-domain and \( L(I_i^n) \) is the interval length, also in Mel-domain. Last, the \( n \)-th MFCC \( (n = I \ldots L) \) is given by [9]:

\[
\tilde{c}_p(n) = \sum_{m=0}^{M} (\log \tilde{S}_p(i)) \cos \left( n - \frac{1}{2} \frac{\pi}{M} \right)
\]

5. Effect of channel distortion on MFCC

The previous section has outlined the extraction of MFCC directly from the original speech signal. If such speech signal passes through a distorting filter, be it either the telephone channel or other, whose frequency response is given by:

\[
H(f)|_{f \leq f_i} = H(k) = H_k e^{j \phi_k}
\]

where \( H_k \) is real, then the stDFT of the speech signal (7) becomes:

\[
S'_p(k) = H_k e^{j \phi_k} \cdot S_p(k) = H_k e^{j \phi_k} \cdot \sum_{n=0}^{N-1} \tilde{S}_p[i] e^{-j \pi \frac{n}{M}}
\]

and integration along Mel-frequency intervals (11):

\[
\tilde{S}_p(i) = \sum_{k=0}^{N-1} \left( 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2} \right) |H_k e^{j \phi_k} \cdot S_p(k)|
\]

5.1. Amplitude-distorting filter

A linear-phase filter that only produces amplitude distortion, such as the one mentioned in section 2 (figure 1) behaves as the combination of a zero-phase filter and a constant delay. Therefore, for this kind of filter it may be assumed \( \theta_k = 0 \) and:

\[
\tilde{S}_p(i) = \sum_{k=0}^{N-1} \left( 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2} \right) |H_k \cdot S_p(k)|
\]

\[
\tilde{S}_p(i) = \sum_{k=0}^{N-1} \left( 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2} \right) \hat{H}_k \cdot \hat{S}_p(k)
\]

(17) can be viewed as a scalar product of two vectors. Specifically, let \( f_{i0}, f_{i1}, \ldots, f_{iM} \) be the set of frequencies \( f_i \) that belong to the interval \( I_i \) and let:

\[
y_k = 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2}
\]

\[
H' = (H_{i0}, H_{i1}, \ldots, H_{iM})
\]

\[
\gamma_k = \hat{y}_k \cdot \hat{S}_p(k)
\]

then (17) becomes:

\[
\tilde{S}_p(i) = \gamma_k \cdot H'
\]

where \( || \cdot || \) is the Euclidean norm and \( \gamma_k \in [-1, 1], \) since it is the output of a cosine function. If the outcome of Mel-frequency integration is now used to calculate MFCC, (12) changes to:

\[
\tilde{c}_p(n) = \sum_{i=0}^{M} (\log ||\gamma_k || \cdot || H' \cdot \xi ||) \cos \left( n - \frac{1}{2} \frac{\pi}{M} \right)
\]

Therefore, each MFCC can be written as the addition of three terms:

- The first term that depends on the signal frequency characteristics and on the Mel-domain integration filter. Thus, it is independent of the type of filtering that has been applied.
- The second term entirely depends on the amplitude frequency response of the filter. It has the effect of an additive component to the MFCC.
- The last term depends on the parameter \( \xi \), which relates the frequency components of the signal to the filter amplitude response within each Mel-frequency band. It only implies variations with respect to the non-filtered signal in the Mel-frequency bands in which the filter response is not constant.

Consequently, the amplitude-distorting filter has the effect on MFCC of adding a constant value that depends solely on the filter, thus it is the same for all signals, and it has the effect of a variable value that depends of the relation between signal and filter frequency characteristics.

5.2. Phase-distorting filter

The all-pass filter defined in (1) produces a response whose DFT is such that \( H = I \) for all \( k \), according to notation in (13). Therefore, (15) can be particularised as follows:

\[
\tilde{S}_p(i) = \sum_{k=0}^{N-1} \left( 1 - \frac{f_k^n - f_i^n}{L(I_i^n)^2} \right) |H_k \cdot S_p(k)|
\]

That is, an all-pass filter, even if it introduces phase distortion does not change the values of MFCC parameters, hence their ability for speech classification remains unaltered.
this paper. The telephone channel has been software-
on automatic voice pathology detection has been presented in

The chosen classifier [10] consists of a 3 layered

time derivatives has been used as a speech frame

descriptor. The chosen classifier [10] consists of a 3 layered

MLP neural-network with 50 hidden nodes having logistic

activation functions and single output that gives scores

For testing purposes, the system depicted in figure 4 has been

implemented in order to compare the efficiency of a voice

pathology detector using high quality speech and speech

quality audio recordings.

By the telephone channel simulator. The

results can be extrapolated for each speaker calculating a

new score averaging the scores given for each frame in the

MFCC features. Preliminary results indicate that the MFCC

features provide a fairly robust scheme for over-the-phone

voice-based diagnostics. Specifically, a degradation of about

10% (from 90% down to 80%) in the classification accuracy

has been observed, which is an improvement over results

previously reported in literature using other parameterization

approaches; for instance in [2] a degradation of 89% down to

74% is reported.

Moreover, and based on the formulation of the MFCC, we

can conclude that these parameters are insensitive to phase
distortion and, thus, the decrease of the efficiency in the

system is more related to the amplitude distortion introduced

by the telephone channel over its pass-band.

Future work will pursue the two-fold objective of

improving these results by a finer selection of both speech

features and pattern classifier and extending them to the case

of more advanced communication systems such as GSM,

UMTS, VoIP, etc.

6. Simulation results

For testing purposes, the system depicted in figure 4 has been

implemented in order to compare the efficiency of a voice

pathology detector using high quality speech and speech

passed through the telephone channel simulator. The

parameter set composed by the first 12 MFCC’s and their fist

temporal derivatives has been used as a speech frame
descriptor. The chosen classifier [10] consists of a 3 layered

MLP neural-network with 50 hidden nodes having logistic

activation functions and single output that gives scores

between 0 (normal voice) and 1 (pathological voice). This can

be interpreted as an estimation of the probability that each

frame belongs to a class.

The MFCC calculation and classification is performed

frame by frame and the final decision is taken for each frame.
The results can be extrapolated for each speaker calculating a

new score averaging the scores given for each frame in the

logarithmic domain. The scores given by the MLP output are

used to calculate a decision threshold that optimally separates

both classes. For both high quality and filtered speech, 70% of

the records available in the database were used for training

the MLP classifier while the rest were used for testing

purposes. A cross validation strategy with 11 folds has been

followed to evaluate the system. The results were obtained

averaging the results for each run.

Figure 5 shows the DET plots that summarize obtained

results. It can be noticed that the performance with filtered

records is degraded about 10% with respect to the high

quality audio recordings.

7. Conclusions and further work

An analysis of the effect of the analogue telephone channel

on automatic voice pathology detection has been presented in

this paper. The telephone channel has been software-
simulated while the pathology detection has been based on

MFCC features. Preliminary results indicate that the MFCC

features provide a fairly robust scheme for over-the-phone

voice-based diagnostics. Specifically, a degradation of about

10% (from 90% down to 80%) in the classification accuracy

has been observed, which is an improvement over results

previously reported in literature using other parameterization

approaches; for instance in [2] a degradation of 89% down to

74% is reported.

Moreover, and based on the formulation of the MFCC, we

can conclude that these parameters are insensitive to phase
distortion and, thus, the decrease of the efficiency in the

system is more related to the amplitude distortion introduced

by the telephone channel over its pass-band.

Future work will pursue the two-fold objective of

improving these results by a finer selection of both speech

features and pattern classifier and extending them to the case

of more advanced communication systems such as GSM,

UMTS, VoIP, etc.

8. Acknowledgements

This research was carried out under grant TEC2006-12887-
C02 from the Ministry of Science and Technology of

Spain; and AL06-EX-PID-033 from the Universidad

Politécnica de Madrid, Spain.

9. References


pathological voice. A voice analysis system for the

screening of laryngeal diseases” IEEE Eng Med Biol


“Telephony-based voice pathology assessment using

automated speech analysis” IEEE Trans Biomed Eng,


Velasco, “Dimensionality reduction of a pathological

voice quality assessment system based on gaussian

mixture models and short-term cepstral parameters”


personal medical network, ESA Publications Division.


duplex telephone channel simulator for high speed data


[6] International Telecommunications Union, Transmission

characteristics of national networks, ITU Rec G.120.

Dec 1998.


Version 1.03. 1994.


detection of voice impairments by means of short-term

cepstral parameters and neural network based detectors”,