Correction of the Voice Timbre Distortions in Telephone Networks: Method and Evaluation

Gaël Mahé *

Université René Descartes-Paris V / CRIP5 / InfoCom,
45 rue des Saints-Pères, 75270 Paris cedex 06, France

André Gilloire

France Télécom R&D / DIH / IPS,
2 avenue Pierre Marzin, 22307 Lannion cedex, France

Lætitia Gros

France Télécom R&D / DIH / EQS,
2 avenue Pierre Marzin, 22307 Lannion cedex, France

Abstract

In a telephone link, the voice timbre is impaired by spectral distortions generated by the analog parts of the link. Our purpose is to restore a timbre as close as possible to that of the original voice of the speaker, using a blind equalizer centralized in the network, which compensates for the spectral distortions.

We propose a spectral equalization algorithm, which consists in matching the long-term spectrum of the processed signal to a reference spectrum within a limited frequency bandwidth (200-3150 Hz). Subjective evaluations show a satisfactory restoration of the timbre of the speakers, within the limits of the chosen equalization band.

The A-law quantization of the output samples of the equalizer induces however a disturbing noise at the reception end. A subjective evaluation shows that speakers' voices with corrected timbre, even with quantization noise, are preferred to the same voices at the output of a link without timbre correction (and without noise).

In order to make the reference spectrum more appropriate to the various speakers' voices, we classify them according to their long-term spectra and use a specific reference spectrum for each class. This leads to a decrease of the spectral distortion induced by the equalizer, significantly perceived as an improvement of the timbre correction, as a subjective test shows.

Key words: spectral equalization, speaker classification, timbre and noise perception

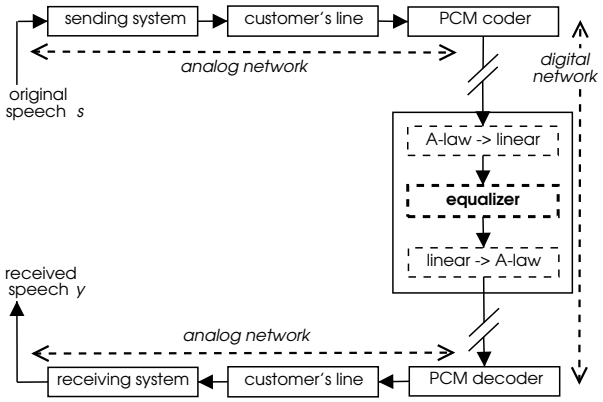

Fig. 1. Telephone link with equalizer.

PACS: 43.72.-p, 43.71.Gv, 43.66.Jh, 43.50.Fe

1 Introduction

In this paper, we consider the voice timbre as an implicitly commonly accepted notion and therefore do not define it. We assume that spectral distortions of a voice signal are perceived as timbre distortions and will validate this asumption by subjective evaluations.

In a PSTN telephone link, as schematized in Fig. 1, the voice timbre is impaired by two kinds of distortions in the analog part of the network.

The first one is the band-pass filtering (300-3400 Hz) at the end-points of the customer line (telephone terminal and line connections to the local exchange) in the transmission and reception paths. For the PSTN, this filtering is described on the average by the *modified Intermediate Reference System* (IRS) (ITU-T, 1996). The nominal frequency responses and masks of the sending and receiving parts of the modified IRS, respectively called sending and receiving systems, are represented in Fig. 2 and 3 respectively, as defined in (ITU-T, 1996). Another band-pass filtering is due to the anti-aliasing filter placed before the PCM coder. We consider in this paper a filter with a frequency response within the mask illustrated in Fig. 4 (National Semiconductor, 1994). This response is nearly flat in the pass-band 200-3400 Hz and has a steep decay in the transition bands.

* Corresponding author. Tel.: +33 1 44 55 35 54; fax: +33 1 44 55 35 35.
Email addresses: `mahe@math-info.univ-paris5.fr` (Gaël Mahé),
`andre.gilloire@rd.francetelecom.com` (André Gilloire),
`laetitia.gros@rd.francetelecom.com` (Lætitia Gros).

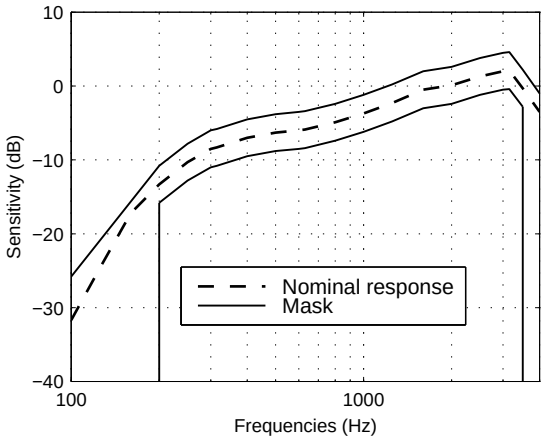

Fig. 2. Nominal frequency response and mask of the sending part of the modified IRS.

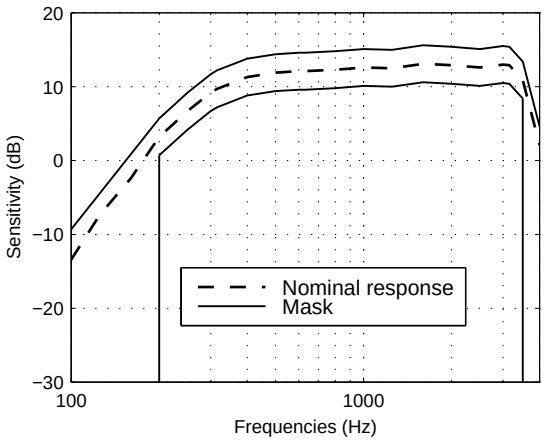

Fig. 3. Nominal frequency response of the receiving part of the modified IRS.

The second distortion is due to the customer's analog line, equivalent to a smooth low-pass filter. The slope of its frequency response is all the steeper as the line is long. In a simple model of the analog line, we assume that the loss, in dB, is proportional to the square root of the frequency:

$$H_{\mathrm{dB}}(f) = H_{\mathrm{dB}}(800\mathrm{Hz})\sqrt{\frac{f}{800}} \tag{1}$$

where the loss (in dB) at 800 Hz, $H_{\mathrm{dB}}(800\ \mathrm{Hz})$, is 3 dB for an average line and 9.5 dB for the longest lines. The frequency responses of different types of lines are shown in Fig. 5. This low-pass filtering muffles the voice in the case of a long line.

In the ISDN and in the GSM network, the signal is digitized in the terminal: the only analog parts are the transmission and reception transducers and their respective amplification and conditioning chains. The ITU-T defined

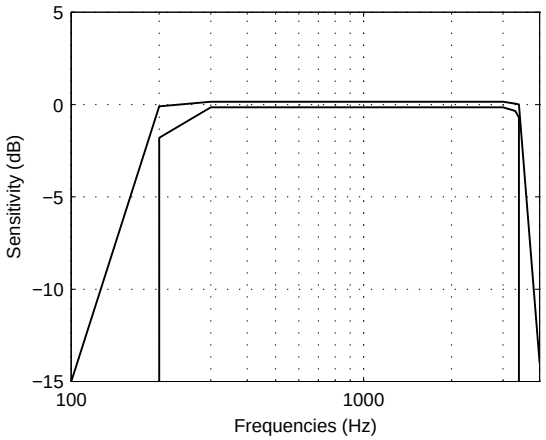

Fig. 4. Mask of the anti-aliasing filter in PCM coder.

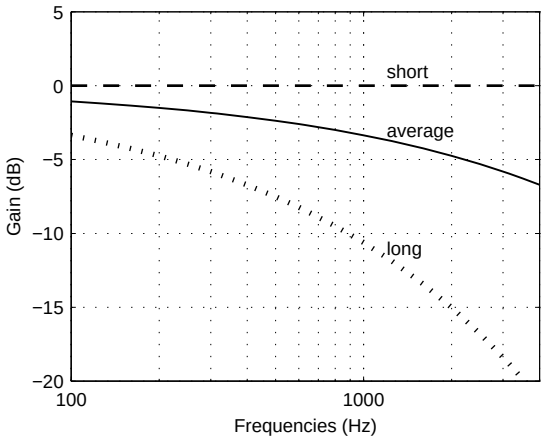

Fig. 5. Frequency responses of different customer lines.

the corresponding transmission and reception sensitivity masks (ITU-T, 1999, 2000). The resulting band-pass filtering is smoother than the modified IRS.

We present in this paper a blind equalization method to compensate for these spectral distortions of voice. Our goal is to make the timbre of the received voice as close as possible to that of the original voice of the speaker. We assume that, when doing this, we can achieve higher naturalness of speech than in not spectrally compensated telephone links. We will consider here only links in the PSTN, in which the spectral distortions are the strongest and therefore the equalization is the most attractive. The correction is assumed to be implemented in the digital part of the PSTN, as shown in Fig. 1, which presents the complete communication chain.

The previously existing spectral equalization methods are presented in Section 2. In Section 3, we propose and evaluate a new blind spectral equalizer. The perceptual effect of quantization noise induced by the equalizer is studied in Section 4. Since noticeable spectral distortions remain for some speakers

in spite of this equalization, we propose in Section 5 to improve the timbre correction by classifying speakers in several classes and differentiating the equalization according to the class of the speaker.

2 State of the art

Fixed equalizers were proposed by Bowker et al. (1993) and by Ho et al. (1993), to compensate for the loss of low frequencies. This compensation is achieved for example (Bowker et al., 1993) by a gain of 10 to 15 dB in the band 100-300 Hz. The drawback of these methods is that the equalization is not adapted to the variability of the conditions of transmission, which can lead to an insufficient or an exaggerated amplification of the low frequencies components. Moreover, it does not compensate for other spectral distortions in middle and high frequencies. The following adaptive equalizers, proposed in other works, perform more accurate corrections.

De Jaco et al. (1997) proposed a device to compensate for non-ideal frequency responses of mobile phone transducers, using two methods. The first one uses two cascaded filters:

- a whitening filter, which coefficients, called "*long term LPC coefficients*" are derived from "*long term autocorrelation coefficients*" of the processed signal;
- a fixed filter which reshapes the spectrum of the signal according to a reference spectrum.

An interest of this method lies in the use of autocorrelation coefficients already computed in the GSM coder. In the PSTN, this characteristic is not so advantageous, since the autocorrelation coefficients would have to be specially computed for the equalization.

The second method presented by De Jaco et al. (1997) consists in dividing the signal in sub-bands (in the time-domain, using a filterbank) and multiplying each of them by a gain, so that the final long-term energy of the sub-band is equal to a target energy. As it will appear in the sequel, such a subband approach coud have been chosen for our purpose. However, in the context of this study, the equalizer was planned to be integrated in a global speech enhancement system, including noise reduction, which frequency response is computed in the frequency domain. That is why it was more advantageous to compute the equalizer in the frequency domain as well.

A similar method was proposed by Mokbel et al. (1996) to compensate for the filtering due to terminals and customer lines, in order to improve speech recog-

nition performance: this filtering results in a deviation of the Mel Frequency Cesptrum Coefficients (MFCC) used in recognition systems, which reduces the recognition performance.

The equalization is performed in the spectral domain, before the computation of the MFCC. The spectrum is divided in 24 sub-bands, and each sub-band energy is multiplied by an adaptive gain. The gains are adapted according to the gradient algorithm, using the criterion of minimization of the mean squared error. For each sub-band, the error is defined as the difference between the sub-band energy and a reference energy. This reference is modulated by the energy of the current frame, to take into account the natural short-term level variations in speech.

This method could be adapted in the purpose of timbre restoration. However, as Mokbel et al. (1996) wrote, the definition of the adaptation step and the modulation of the reference by the energy of the current frame are quite delicate: without a fine empirical tuning, the performances of the equalizer are noticeably impaired.

A simpler method, the *cepstral subtraction*, was formerly proposed by Mokbel et al. (1993), leading to a similar improvement of the recognition performance. Mokbel et al. (1993) showed that a good estimation of the transmission channel is given by the average cepstrum of the pre-emphasized received signal. So the equalization is performed by subtracting the averaged cepstrum of the signal from the current cepstrum. Denoting by C_x and C_y the cepstra of the processed and the equalized speech respectively,

$$C_y = C_x - \overline{C_x} \tag{2}$$

The cepstral subtraction can be transposed in the spectral domain for our purpose. With the usual notations,

$$|Y(f)| = \frac{|X(f)|}{\overline{|X(f)|}^{\mathrm{g}}} \tag{3}$$

where $\overline{a}^{\mathrm{g}}$ denotes, for a variable a, its geometric mean. So, the cepstral subtraction consists in making the average spectrum of the signal uniformly equal to 1. This value leads to satisfactory results in speech recognition, but it is not necessarily optimal for the purpose of timbre correction.

All these methods globally aim at making the spectrum of the processed signal closer to a reference spectrum. We will now explain more in detail this principle and use it in a way appropriate to the context presented in Section 1. (Mahé and Gilloire, 2001)

3 Blind spectral equalization

3.1 Principles

Denoting G the frequency response of the analog channel, $|S(f)|^2$ the short-term power spectral density (PSD) of the original signal s and $|Y(f)|^2$ the short-term PSD of the received signal y, we have:

$$|Y(f)|^2 = |G(f)|^2 |S(f)|^2 \tag{4}$$

If we assume the channel to be time-invariant, the time-averages of $|Y(f)|^2$ and $|S(f)|^2$ are related by:

$$\overline{|Y(f)|^2} = |G(f)|^2 \overline{|S(f)|^2} \tag{5}$$

The frequency response of the equalizer compensating for the spectral distortions of the analog channel is therefore defined by:

$$|EQ(f)| = \frac{1}{|G(f)|} = \sqrt{\frac{\gamma_s(f)}{\gamma_y(f)}} \tag{6}$$

where γ denotes the *long-term spectrum* of a signal, defined as the time average of the short-term PSD (for example, $\gamma_s(f) = \overline{|S(f)|^2}$.

Since the long term spectrum of the original voice $\gamma_s(f)$ is unknown, we propose to approximate it by the average spectrum of speech defined by the ITU-T (1993). This spectrum was derived from speech samples from a large panel of speakers and languages. We will call this spectrum *reference spectrum* and denote it by $\gamma_{\text{ref}}(f)$. So, with this approximation, the equalization consists in matching the spectrum of the processed signal to the defined reference spectrum.

Since the equalizer is placed inside the network, $\gamma_y(f)$ is not directly available and therefore it must be derived from the input of the equalizer, x, using:

$$\gamma_y(f) = |L\_RX(f)|^2 |S\_RX(f)|^2 \gamma_x(f) \tag{7}$$

where L_RX is the frequency response of the receiving line and S_RX is the frequency response of the receiving system. These values are assumed to be known, otherwise they are replaced by average values.

So the frequency response of the adapted equalizer becomes :

$$|EQ(f)| = \frac{1}{|L\_RX(f)||S\_RX(f)|}\sqrt{\frac{\gamma_{\text{ref}}(f)}{\gamma_x(f)}} \tag{8}$$

We will call this equalizer *adapted equalizer*, because it is automatically adapted to the channel.

Since the IRS filtering and the anti-aliasing filter strongly attenuate the frequency components lower than 200 Hz and higher than 3400 Hz, compensating for the spectral distortions outside the band 200-3400 Hz would lead to an unacceptable amplification of noise. Moreover, because of the steep decay of the IRS frequency response in the bands 200-300 Hz and 3150-4000 Hz, the frequency response of the equalizer should have a steep slope in these bands, which would imply a large number of coefficients in a time-domain implementation. In order to reduce the complexity of the equalizer, we will not equalize the band 3150-4000 Hz, which restoration is less perceptible than the restoration of low frequency components. That is why only the band 200-3150Hz is equalized according to Equation (8).

3.2 Need of a pre-equalizer

The assumption $\gamma_s(f) = \gamma_{\text{ref}}(f)$ is correct as far as the global shapes of these spectra are similar, but the long-term spectrum of each speaker fluctuates around the reference, as illustrated in Fig. 4 for a particular speaker. This approximation leads therefore to a residual spectral error equal to the difference between these spectra in dB (see Fig. 6), which is perceived as a disturbing timbre impairment. That is why $|EQ|$ computed according to (8) must be strongly smoothed, to flatten the local differences between $\gamma_{\text{ref}}(f)$ and $\gamma_s(f)$. Since the global shape of the error $\gamma_{\text{ref}}(f)/\gamma_s(f)$ is flat, this smoothing may lead to a nearly flat frequency response of the equalized link.

This smoothing however is incompatible with the need of a frequency response having a steep slope in the band 200-300 Hz. This contradiction is overpassed as follows, using *a priori* knowledge of standard telephone links.

The ITU defined for the IRS a mask which width does not exceed 5 dB in the band 200-3150 Hz. Moreover, frequency responses of analog lines with various lengths spread around the frequency response of an average line. For these reasons, the analog channel does not need to be entirely blindly equalized. We propose to perform first a fixed pre-equalization, compensating for an average analog channel (average analog lines and sending and receiving systems having the nominal characteristics of the modified IRS) and to complete the

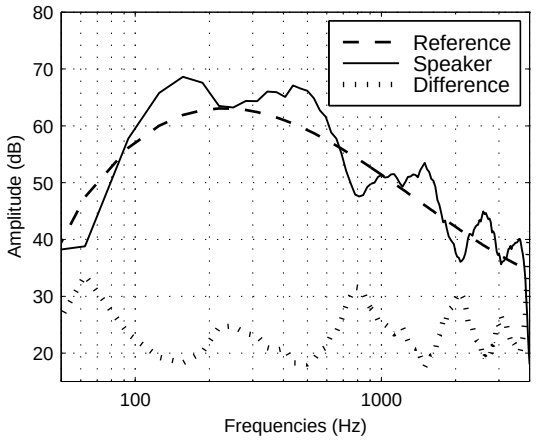

Fig. 6. Long term spectrum of a speaker vs reference spectrum; difference between the two spectra.

equalization by an adapted equalizer as defined in the previous section. The task of the adapted equalizer consists thus in correcting the mismatch between the pre-equalizer characteristics and the real transmission conditions, which allows a smooth frequency response. Note that the equations of Section 3.1 remain valid, G denoting now the analog channel in series with the pre-equalizer and x the output of the pre-equalizer.

3.3 Implementation

The output of the pre-equalizer is analyzed on frames of length 32 ms, with an overlapping of 50%. The adapted equalizer is a FIR filter, which coefficients are updated at each frame, according to Equation (8), as described below. According to Mokbel et al. (1993), 2 to 4 seconds of voice activity are necessary to estimate the channel. So the long-term spectrum of x, γ_x, is first computed by averaging the short-term spectrum over a voice activity duration increasing from 0 to 4 s. Then it is recursively updated for each frame of voice activity. Its generic expression is given by:

$$\gamma_x(f,n) = \alpha(n)|X(f,n)|^2 + (1-\alpha(n))\gamma_x(f,n-1) \tag{9}$$

where $\gamma_x(f,n)$ is the long-term spectrum at the n^{th} frame of voice activity, initialized with $\gamma_x(f,0) = 0$, $|X(f,n)|^2$ the short term spectrum of the n^{th} frame of voice activity, and

$$\alpha(n) = \frac{1}{\min(n,N)} \tag{10}$$

where N is the number of frames in 4 s.

The frequency response of the adapted equalizer is then computed according to (8). As explained in Section 3.1, only the band 200-3150 Hz must be equalized, so that the values of $|EQ|$ out of this band are replaced by a linear extrapolation of $|EQ|_{[200-3150\text{Hz}]}$, before deriving the impulse response from $|EQ|$ by an IFFT. The smoothing of $|EQ|$ is achieved by a narrow windowing of the impulse response, using typically a Hamming window of length 15.

Finally, the adapted equalizer response is multiplied by a gain factor, so that the speech level at the output of the equalized link is equal to the level of the non-equalized output speech signal.

3.4 Simulations and objective results

The above equalizer has been simulated for various telephone links with 34 speakers uttering a text of approximately 25 s of voice activity. The speech signals of our database, taken as inputs of the links, were recorded in the same conditions as these followed by the ITU-T (1993) to define the average spectrum of speech, *i.e.* in quiet environnement with a high quality microphone. They were additionnaly low-pass filtered, so that their frequency range is 0-4000 Hz. In the following example, the analog part of the simulated telephone link is composed of a long transmission line, an average reception line and sending and receiving systems having frequency responses according to the nominal response of the modified IRS (ITU-T, 1996).

The spectral distortion of the received equalized signal in the equalized band 200-3150 Hz is measured by the cepstral error of the channel estimation. For each frame, the cepstral error e is defined as the cepstral distance (Faucon et al., 1993) between the adapted equalizer and the equalizer corresponding to a perfect channel estimation in the equalized band, that we will call *ideal adapted equalizer*:

$$e = \sqrt{\sum_{i=1}^{20} (C_{\text{eq}}^i - C_{\text{ideal\_eq}}^i)^2} \tag{11}$$

where C_{eq}^i and $C_{\text{ideal\_eq}}^i$ are the i^{th} cepstral coefficients of the adapted equalizer and the ideal adapted equalizer, respectively. The first cepstral coefficients C_{eq}^0 and $C_{\text{ideal\_eq}}^0$ are not taken into account in the computation of e, so that the error does not reflect any possible difference of energy. Moreover, because of the smoothing of the frequency responses, the coefficients of orders higher than 20 are very small and therefore can been discarded from the error calculation.

Figure 7 presents the evolution of the cepstral error for the 34 speakers, separated according to the behavior of the error. Since the equalizer is not updated

during the frames of speech inactivity, these latter ones are not represented. Indeed, during these phases, the cepstral error is constant.

For 22 speakers (Fig. 7a), the convergence of the equalizer, corresponding to the stabilization of the cepstral error around a minimum value, is achieved within less than 5 s of speech activity. For these speakers, the evolution of the cepstral errors is represented by a series of histograms. Each n^{th} vertical line represents the histogram of the cepstral errors at the n^{th} frame of voice activity, where the darkness of a pixel of coordinates (n, e) is proportional to the number of speakers having a cepstral error around e.

For 8 speakers, the convergence is slower and is achieved within about 10 s. Three of them (Fig. 7b) have a final cepstral error around 0.1, like in the first group, while the cesptral error converges towards a higher values (about 0.2) for five of them (Fig. 7c).

At last, for four speakers (Fig. 7d), the cepstral error has a special evolution, with irregular variations after a quick decay at the beginning. This phenomenon can be explained by sudden variations of the long term spectrum.

The frequency response of the adapted equalizer after convergence is presented for four speakers in Fig. 8. These speakers were selected as follows:

- M1 : male speaker with high cepstral error;
- M2 : male speaker with low cepstral error;
- F1 : female speaker with low cepstral error;
- F2 : female speaker with high cepstral error;

For each speaker, this frequency response is compared to the frequency response of the ideal adapted equalizer. The difference between these characteristics in dB corresponds to the spectral distortion between the received equalized speech and the original speech in the band 200-3150 Hz. For M2 and F1, which cepstral error after convergence is representative of 75 % of the speakers, the amplitude of this distortion does not exceed 3 dB.

3.5 Subjective evaluation

The subjective evaluation aims first at validating the proposed equalizer in terms of timbre restoration, by comparing the timbres of the original signal, the equalized output signal and the non-equalized output signal, for various simulated telephone links. The evaluation shall also take into account the unavoidable limitation of the equalization band. So, for a given link, the timbres of those signals shall be compared to those of the output signal of the link

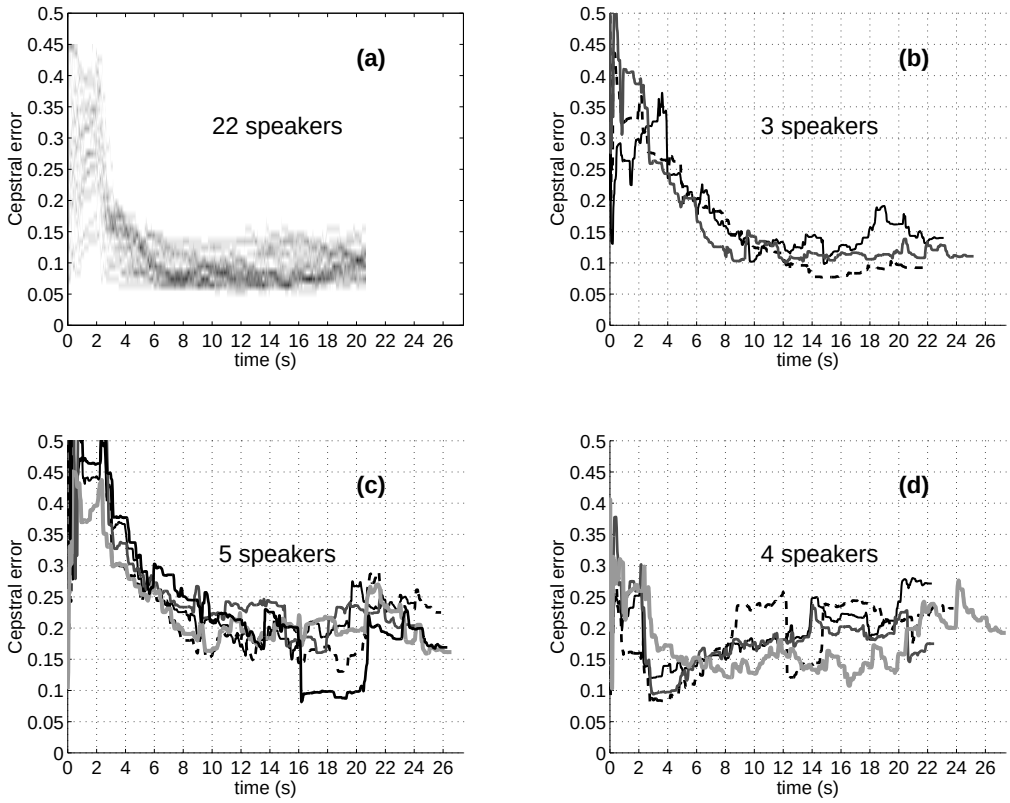

Fig. 7. Evolution of cepstral error for 34 speakers.

perfectly equalized in the band 200-3150 Hz (by the ideal equalizer). Finally, all those signals shall be compared to the output of the link equalized by the pre-equalizer only, in order to evaluate the relevance of the adapted part of the equalization.

A complete pairs comparison test would however lead to an unrealistic test duration if several speakers and transmission conditions are tested. For this reason, we used the *Multi Stimuli test with Hidden Reference and Anchor (MUSHRA)* (ITU-R, 2001) as follows.

3.5.1 Test plan

The equalizer has to be evaluated for several speakers and transmission conditions. For each speaker and for each link, we used only one test sentence (the same), for the following reason. Because of the long term smoothing effect in the spectra estimates (Eq. 9), the frequency response of the equalizer depends on the long term spectrum of speech uttered by each speaker. Therefore, after the equalizer has reached its steady state (convergence of the cepstral error),

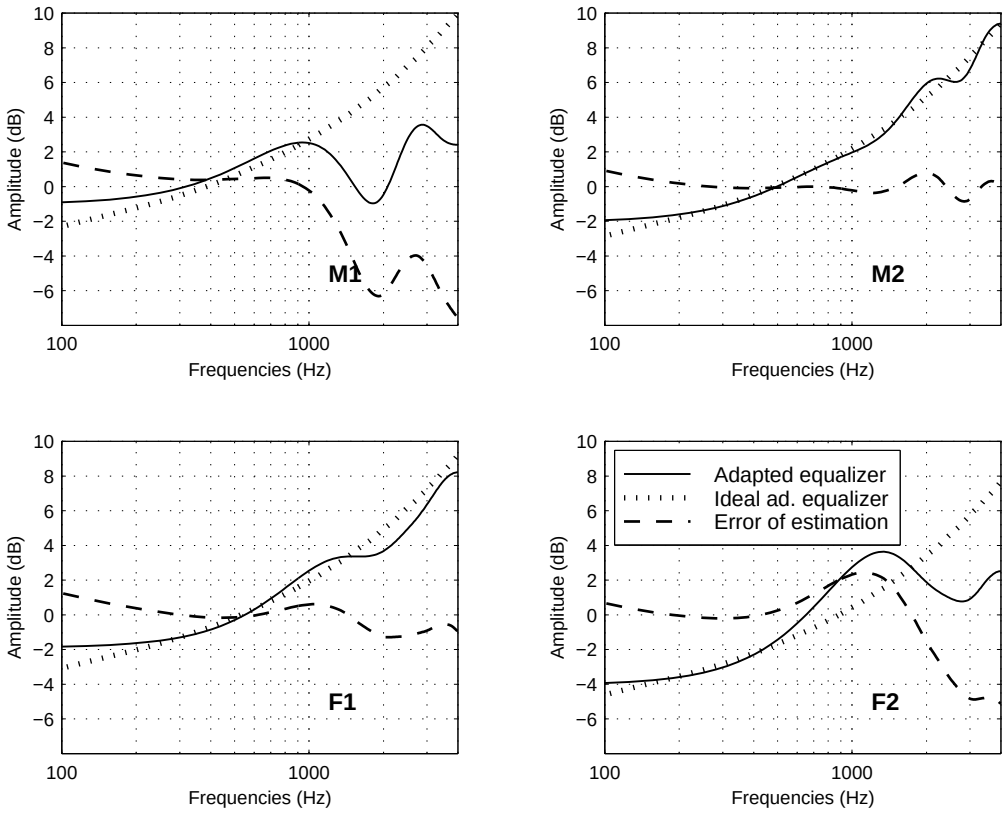

Fig. 8. Frequency response of the adapted equalizer vs ideal adapted equalizer.

the phonemes of the last part of the speech utterance have a fairly weak influence on the frequency response of the equalizer. The links were simulated with the same input utterance as in the previous section (about 25 s of speech activity), but only the last sentence, of duration 10 s and pronounced after convergence of the equalizer, was presented to the listeners.

The four previous speakers M1, M2, F1 and F2 were used in the test. The cepstral error observed for M2 and F1 after convergence is representative of the majority of the speakers, whereas M1 and F2 corresponds to the lowest objective performances of the equalizer.The evolution of the cepstral error of the equalizer during the test sentence is represented in Fig. 9 for the four tested speakers. One could wonder at the choice of F2 as representative of the speakers with high cepstral error, since the latter is around 0.1 during the first 4 seconds of speech. However, we chose F2 on the one hand because we thought that the worst part of the tested sentence, in terms of timbre degradation, woud be determining in the judgement of the listeners (this hypothesis was *a posteriori* validated by the results of the tests), on the other hand because among the rare female speakers with high cepstral error, F2 was the only with a non-disturbing voice, in terms of elocution and clarity.

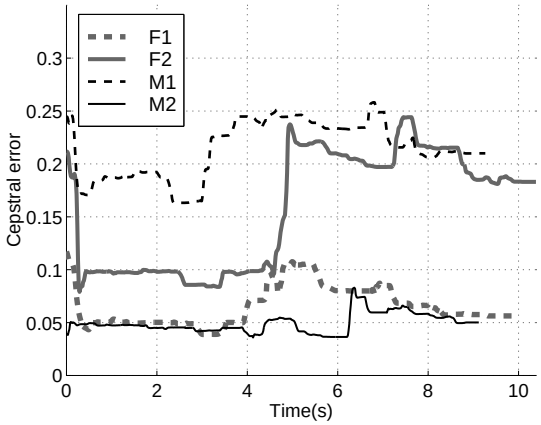

Fig. 9. Cepstral error of the equalizer during the test sentence for the four tested speakers.

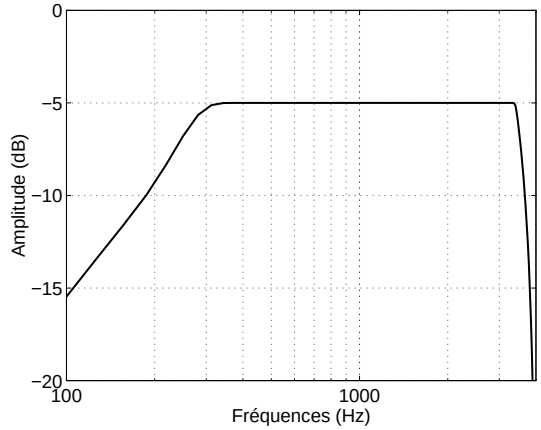

Fig. 10. Frequency response of the simulated ISDN terminal in reception.

Informal listening pre-tests were first conducted to select telephone link configurations supposed to produce different and perceptible timbre impairments. Two telephone links were retained:

- L_1, that seems to muffle the voice: a PSTN link with sending and receiving systems according to the nominal frequency response defined in (ITU-T, 1996), a long transmission analog line and an average reception analog line;
- L_2, that seems to put too much emphasis on high frequencies: a PSTN-ISDN link with a sending system according to the nominal frequency response defined in (ITU-T, 1996), a very short transmission analog line and a reception link in the ISDN, which terminal has the frequency response represented in Fig. 10, complying with the mask defined in (ITU-T, 2000).

For each condition (speaker, link), listeners were asked to compare the timbres of the output signals:

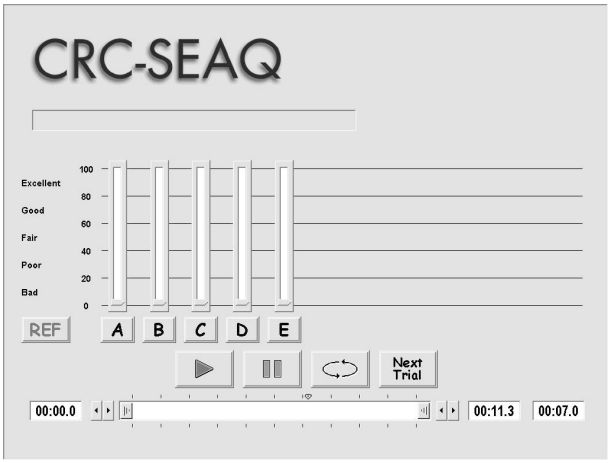

Fig. 11. Interface used for the MUSHRA test.

- RX_0: link without equalizer;
- RX_{EQ}: equalized link;
- RX_{PRE}: link equalized by the pre-equalizer only;
- RX_{ID}: link equalized by the ideal equalizer

to the timbre of the original speech (input signal of the link), denoted by ORI and taken as reference. Each signal was rated according to a scale of timbre proximity : 100 if its timbre is perceived as the same as the reference timbre, 0 if it is perceived as very different from the reference.

For this test, we used the graphical interface represented in Fig 11. For a given condition (speaker, link), the REF button corresponds to the reference (ORI), the buttons A...E to the test signals, composed of ORI, RX_0, RX_{EQ}, RX_{PRE} and RX_{ID}, according to a random order. Subjects can listen to and switch between all the signals, as much as they want, and rate each test signal using the slider above the corresponding button. This presentation allows listeners to adjust accurately the marks.

24 adult subjects of various levels of expertise participated in the test.

3.5.2 Results

Figures 12 and 13 show the mean scores and the associated confidence intervals (graphs on the right) facing the spectral distortions (graphs on the left), for each speaker and each link (Fig. 12 for L_1 and Fig. 13 for L_2). One can see that for all the speakers and both links, the mean score obtained for RX_{EQ} is systematically above the mean score obtained for RX_0, in accordance with the objective results.

Besides, for speakers M1, M2 and F1, the mean scores obtained for RX_{ID}, RX_{EQ} and RX_{PRE} are roughly similar, contrary to what one could expect

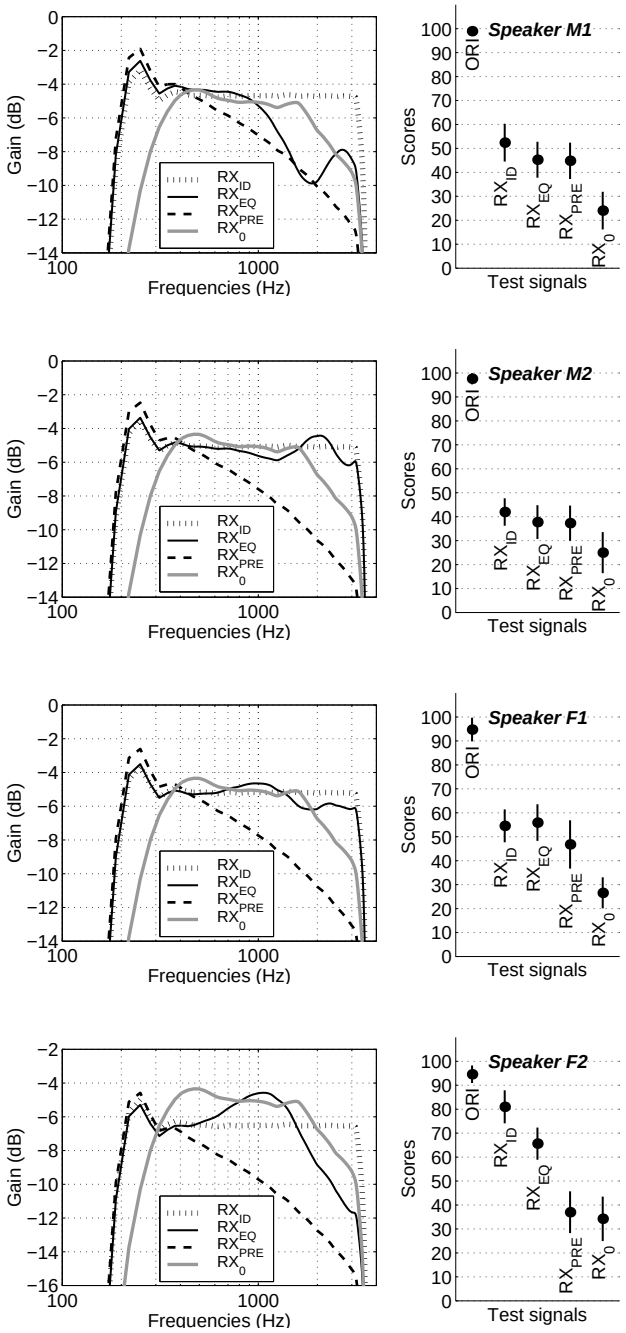

Fig. 12. Mean scores, confidence intervals (right) and corresponding spectral distortions (left) of outputs of telephone link L_1.

considering the spectral distortion curves. Particularly, for M2 and F1, we could expect the score of RX_{EQ} to be more clearly above that of RX_{PRE}. For speaker F2 only, the scores hierarchy matches clearly the spectral distortions.

These unexpected individual results can be explained by the major influence of the loss of low-frequencies on the perception of the timbre distortion. The

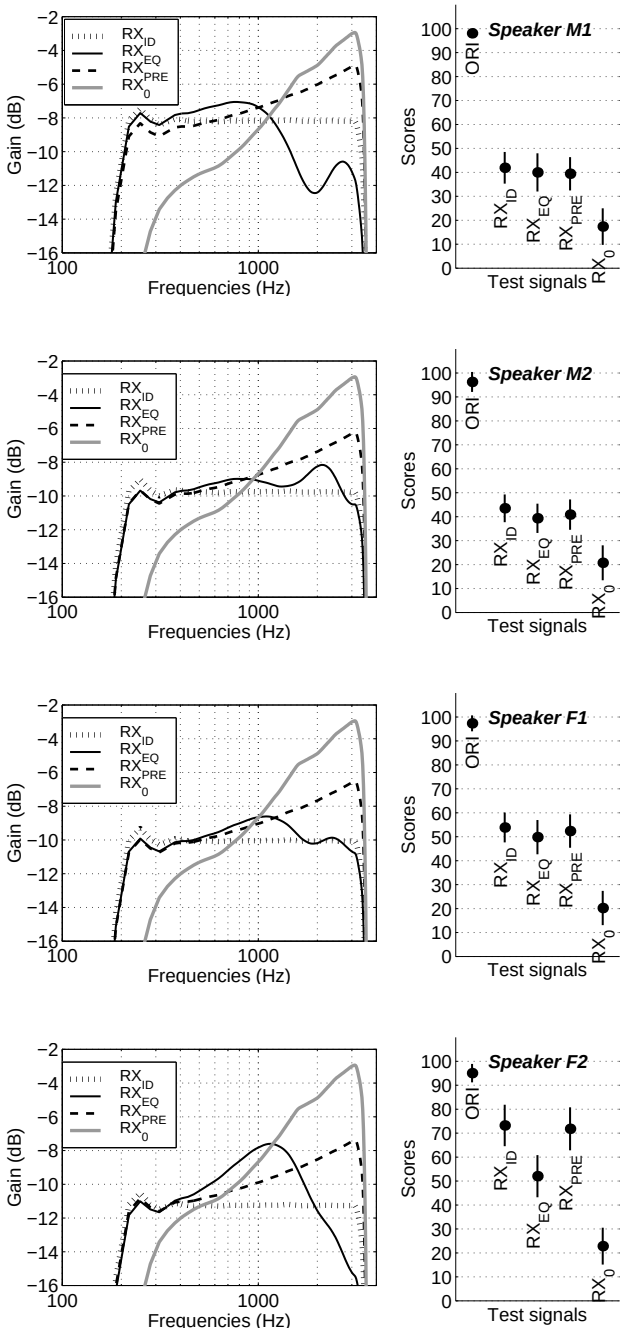

Fig. 13. Mean scores, confidence intervals (right) and corresponding spectral distortions (left) of outputs of telephone link L_2.

timbre of speaker F2, which mean pitch is 240 Hz, is not much affected by the band-pass filtering 200-3400 Hz. On the contrary, for speakers M1, M2 and F1, whose mean pitches are respectively 120, 160 and 210 Hz, the non-restoration of frequency components below 200 Hz modifies so much the timbre that the perceived differences between RX_{ID}, RX_{EQ} and RX_{PRE} are compressed when these signals are compared to ORI and RX_0 in the MUSHRA test. Therefore,

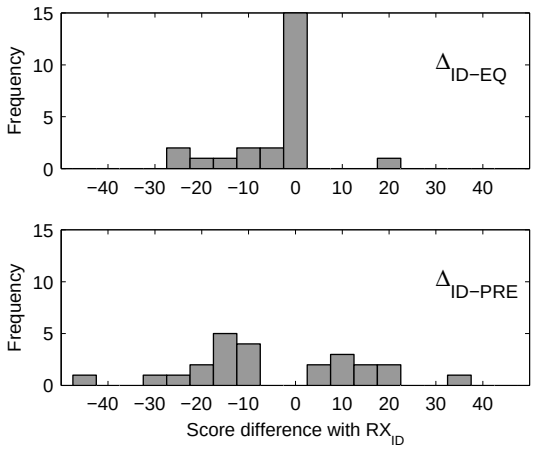

Fig. 14. Distribution of $\Delta_{\mathrm{ID-EQ}}$ and $\Delta_{\mathrm{ID-PRE}}$ for speaker M2 and link L_1.

due to the resulting lack of discrimination of this scale, it cannot be concluded that a fixed filter approach is as good as the adapted equalization. It can only be concluded that, *for some speakers observed individually and compared to* RX_0 *and ORI*, a fixed filter is as good as the adapted one.

In order to study more accurately the significance of the mean score differences considering the *whole set* of tested speakers, Tukey tests (Tukey, 1953) are conducted on the scores considering the factor Link at two levels and the factor Speaker at four levels. It appears that for the link L_1, all speakers being considered:

- $\mathrm{RX_{ID}}$ and $\mathrm{RX_{EQ}}$ are significantly above $\mathrm{RX_{PRE}}$;
- there is no significant difference between $\mathrm{RX_{ID}}$ and $\mathrm{RX_{EQ}}$.

To observe individually differences, it would be necessary to conduct another test without RX_0, which would stretch the score scale. Assuming that $\mathrm{RX_{ID}}$ is the best achievable result (taking into account the limitation of the equalization band), we propose to compare more finely $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{PRE}}$ by observing on the one hand the difference $\Delta_{\mathrm{ID-EQ}}$ between the scores of $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{ID}}$ and on the other hand the difference $\Delta_{\mathrm{ID-PRE}}$ between the scores of $\mathrm{RX_{PRE}}$ and $\mathrm{RX_{ID}}$. Fig. 14 represents for speaker M2 and link L_1 the distribution of these differences: the two distributions are clearly separate, which could not be stressed by a variance analysis, since these distributions are concentric. So, the mean values of $\Delta_{\mathrm{ID-EQ}}$ and $\Delta_{\mathrm{ID-PRE}}$ are nearly the same, but most listeners perceived $\mathrm{RX_{EQ}}$ closer to $\mathrm{RX_{ID}}$ than $\mathrm{RX_{PRE}}$, which matches the objective results. The difference between these signals and the original voice is such that listeners are not able to decide surely whether $\mathrm{RX_{PRE}}$ is closer to ORI than $\mathrm{RX_{ID}}$ and $\mathrm{RX_{EQ}}$.

Thus, both mean values and standard deviations of $\Delta_{\mathrm{ID-EQ}}$ and $\Delta_{\mathrm{ID-PRE}}$ should be compared to get a thorough comparison of $\mathrm{RX_{ID}}$, $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{PRE}}$.

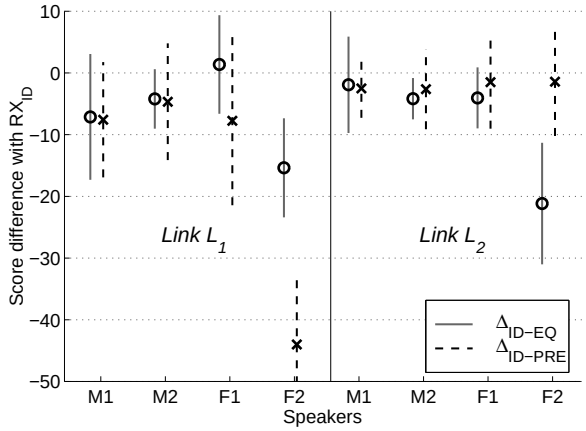

Fig. 15. Mean values and standard deviations of $\Delta_{\mathrm{ID-EQ}}$ and $\Delta_{\mathrm{ID-PRE}}$.

These values are illustrated on Fig. 15. For speakers M1, M2 and F1, the mean values of $\Delta_{\mathrm{ID-EQ}}$ and $\Delta_{\mathrm{ID-PRE}}$ are very close, whatever the respective spectral distortions of $\mathrm{RX_{ID}}$, $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{PRE}}$ may be. But the standard deviation of $\Delta_{\mathrm{ID-EQ}}$ is approximately two times smaller than that of $\Delta_{\mathrm{ID-PRE}}$ for speakers M2 and F1, for whom the objective results indicate that $\mathrm{RX_{EQ}}$ is closer than $\mathrm{RX_{PRE}}$ to $\mathrm{RX_{ID}}$. This means that the proximity observed on an average between $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{ID}}$ is more probable than that between $\mathrm{RX_{PRE}}$ and $\mathrm{RX_{ID}}$. For speaker M1, the variance of $\Delta_{\mathrm{ID-EQ}}$ is similar to the variance of $\Delta_{\mathrm{ID-PRE}}$ for link $\mathrm{L_1}$, bigger for link $\mathrm{L_2}$, which reflects the distorsions curves.

The difficulty of obtaining a stronger confirmation of the objective results for each speaker individually lies mainly in the task listeners were asked to do in the test: since only one reference could be used, listeners were asked to evaluate $\mathrm{RX_{EQ}}$ and $\mathrm{RX_{PRE}}$ not relatively to $\mathrm{RX_{ID}}$, but relatively to ORI.

Finally, these subjective results confirm on the one hand that the limitation of the equalization band (200-3150 Hz) reduces the ability of the equalizer to restore the original timbre of a voice (particularly for male voices). On the other hand, they show that within these limits, for a large majority of speakers, the adapted equalization reaches the objective of a non-perceptible spectral distortion in the equalized band, with better performances than a fixed filter like the pre-equalizer.

4 Equalization and quantization noise

4.1 How equalization induces quantization noise

We have observed experimentally that the output signal of the equalized link is impaired by a white noise with fluctuating level, which is most often higher than the quantization noise level of the output of the same link without equalization. This added noise is caused by the combination of equalization and linear to A-law conversion. Since the equalizer is placed before the receiving system, it must over-amplify the low frequency components of the voice, in order to compensate for their attenuation by the receiving system. Consequently, if the original signal is energetic in the band 200-300 Hz, the spectrum of the signal before linear to A-law conversion is very much unbalanced, so that the level of the quantization noise induced by this conversion is close to the level of the medium and high frequencies components.

Let us observe this phenomenon for a frame of the phoneme [Y] pronounced by a female speaker. Figure 16 represents the spectral envelopes of signal and noises in three successive points of the link: after PCM coding (i); after equalization (ii); at the output of the link in equalized and non-equalized cases ((iii) and (iv) respectively). In this figure, q_0 and q_1 denote respectively the quantization noise of the PCM coder and the quantization noise of the linear to A-law conversion. The phoneme [Y] has a first formant around 200 Hz, which is attenuated by the sending system (i). Thus, after over-amplification of the band 200-300 Hz by the equalizer, this band is very energetic and determines the level of quantization noise q_1, which overpasses the second formant, contrary to q_0 (ii). Consequently, the global quantization noise is more audible at the output of the equalized link (iii) than at the output of the non-equalized link (iv).

4.2 Perceptual approach of noise and timbre correction

The subjective evaluation of Subsection 3.5 was actually performed with simulated links ignoring the linear to A-law conversion. The timbre comparisons were thus not disturbed by the quantization noise q_1, which should not be considered as a component of the timbre. Since the effect of this noise on perception of the equalized signal cannot be neglected however, we propose now:

- to control the audibility of the quantization noise;
- to complete the previous evaluation by a global evaluation of joint perception of timbre and noise.

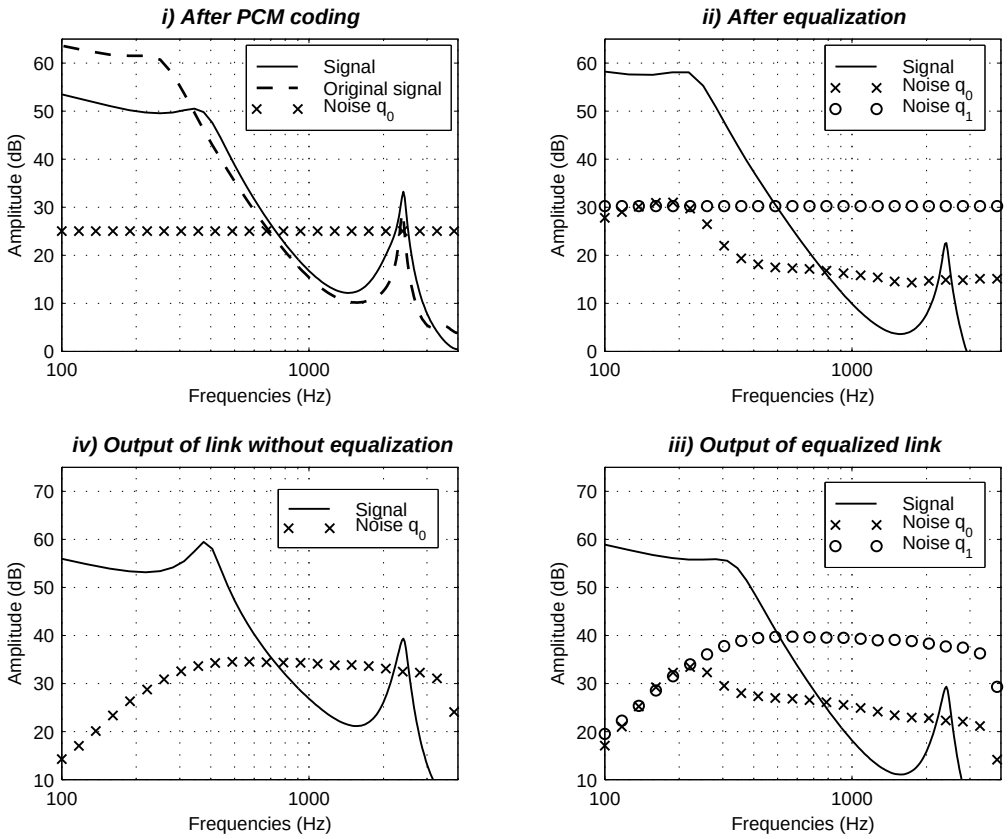

Fig. 16. Spectral envelope of [Y] and corresponding quantization noises in different parts of the link.

4.2.1 Noise shaping and masking

We have proposed (Mahé and Gilloire, 2002) two methods to spectrally reshape the noise q_1, aiming at perceptually masking it. This goal implies that the noise spectrum be below the masking threshold of the voice signal:

$$\gamma_n(f) = \lambda^2 \gamma_{\text{Mask}}(f) \tag{12}$$

where γ_n is the reshaped noise power spectral density, γ_{Mask} the masking threshold and λ a factor inferior to 1.

Usual spectral shaping methods for instantaneaous coders (like PCM coders) (Makhoul and Berouti, 1979; Boite et al., 2000) globally consist in whitening the signal being quantized, and reshaping its spectrum at the reception, which is not possible here, since the entire processing is assumed to be performed in a node of the network. We propose to reshape the noise without any additional filter in reception, simply by modifying the A-law quantization.

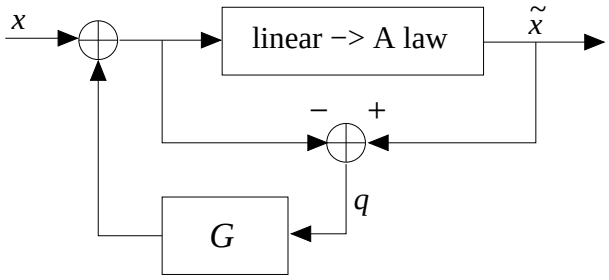

Fig. 17. Noise shaping using a feedback loop.

The first method, inspired from (Makhoul and Berouti, 1979), consists in adding to the input of the quantizer the filtered quantization error, as represented in Fig. 17. According to this structure, the Z-transform of the reshaped noise n is defined by:

$$N(z) = \tilde{X}(z) - X(z) = (1 + G(z))Q(z) \tag{13}$$

where q is the quantization error and G the noise shaping filter. According to (12), the noise shaping implies:

$$N(z) = \lambda H(z)W(z) = \lambda\sigma(1 + G_0(z))W(z) \tag{14}$$

where $H(z) = \sigma(1 + G_0(z))$ is an ARMA model of the masking threshold and $W(z)$ the Z-transform of a white noise of variance unity. From (13) and (14) we deduce the loop filter G and the value of λ:

$$G(z) = G_0(z), \;\; \lambda = \frac{\sigma_q}{\sigma} \tag{15}$$

where σ_q is the standard deviation of the quantization error q.

According to (15), the noise masking implies the minimization of λ, and consequently the maximization of σ. But this problem is strongly constrained by the stability of the loop, so that the level of the noise is not well controled. In (Mahé and Gilloire, 2002), we showed that using for H a MA model which coefficients are identical to those of the AR model of the inverse of the mask guarantees the stability of the loop.

This method was simulated using the Johnston method (Johnston , 1988) to compute the mask, which was updated every 16ms and accurately approximated by a MA model of order 20. The resulting quantization noise is spectrally reshaped according to the model as expected, but its level is not controlled, so that it is occasionaly above the masking threshold. The evolution of λ is represented in4 Fig. 18 for 4 speakers. Referring to (12), effective noise masking corresponds to λ below 0 dB. As can been deduced from this

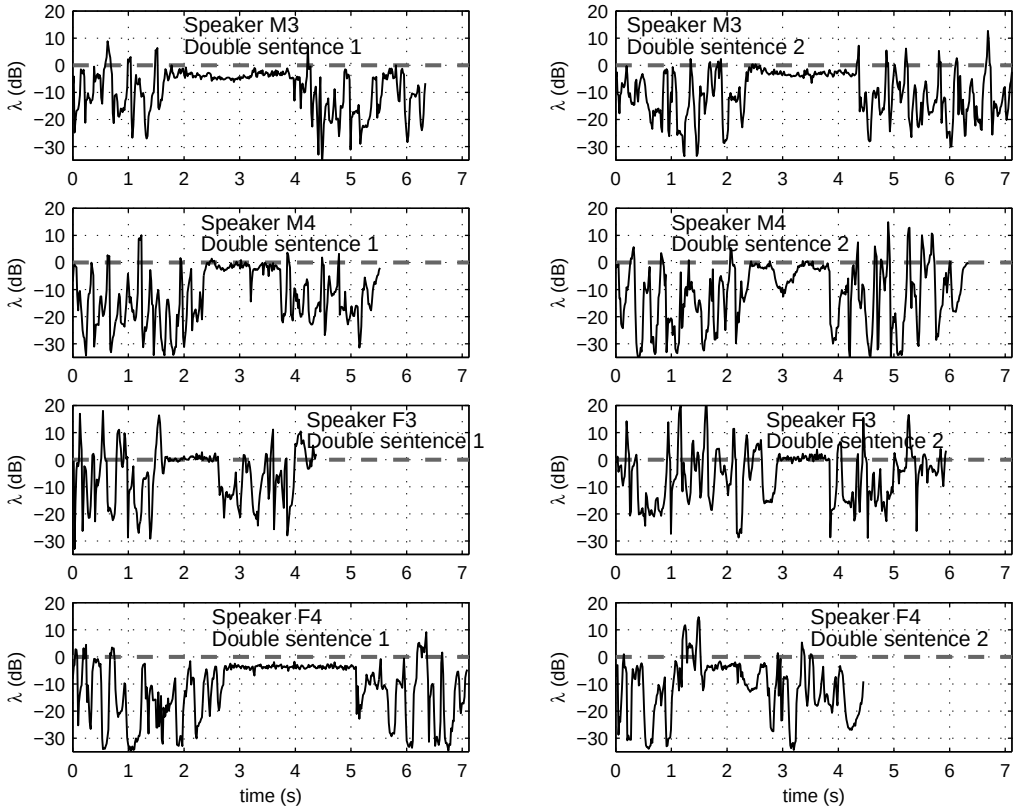

Fig. 18. Level offset λ between the reshaped noise and the masking threshold.

figure, the reshaped noise is sporadically audible and this audibility depends on the speakers and the pronounced phonemes.

The second shaping method explores the possible temporal sequences of quantization levels, in order to maximize a probabilistic criterion of noise masking, using a Viterbi-like algorithm. This method, presented in details in (Mahé and Gilloire, 2002), is theoretically optimal since it is not constrained like the previous method. But it is so complex that a simpler and sub-optimal version must be run, which limits in practice its performance. So, the simulations did not lead to better objective results.

Consequently, only the first method was subjectively evaluated, in order to compare the perception of a permanent white quantization noise to the perception of a sporadically audible reshaped noise.

4.2.2 Test methodology

The tests aim at knowing if a user prefers a restored signal (close to the original voice of his/her interlocutor) even if noisy (with or without noise shaping), to a non-restored signal, but not noisy. For a given link and a given signal, we denote by:

- O: the original signal;
- A: the output signal of the non-equalized link;
- B: the output signal of the equalized link, without noise shaping;
- C: the output signal of the equalized link, with noise shaping.

In this experiment, we have used the previously defined ideal equalizer, which inverts without error the analog channel in the band 200-3150 Hz.

Since the knowledge of the original voice of the speaker may influence the results of the test, two cases must be considered according to the knowledge - or not- of the original voice by the listener.

The case where the listener does not know the original voice of the speaker can be simulated by evaluating A, B and C, independently of O. The evaluation consists then in a pair comparison test (Bonnet, 1986) where listeners give their preferences for each pair {X,Y} of {A, B, C}.

In the case where the original voice is known, the second test should determine which signal, among A, B and C, is preferred, knowing the original signal O. Since the spectral properties of B and C are similar, their timbres are the same, therefore we assume that knowing O does not change the preference between B and C. After the preferred noise (reshaped or not) has been determined in the first test, the signal impaired by the other noise does not need to be evaluated in the second test: only the preferred noise should be kept.

Now, considering that a test where each listener knows the original voice of each speaker is unrealistic, the case where the original voice of the interlocutor is known is simulated as follows: for each pair {A, X} where X = B or C (depending on the result of the first test), listeners have to compare the degradation of A relatively to O to the degradation of X relatively to O.

4.2.3 First test: preferences without knowing the original voice

Since the quantization noise mainly depends on the speaker and the pronounced phonemes, one telephone link was tested, with:

- 4 speakers (2 male: M3 and M4; 2 female: F3 and F4);
- two 8 s double-sentences for each speaker (same sentences as those used for

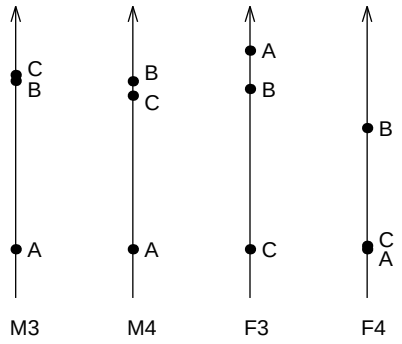

Fig. 19. Thurstone preferences scale for each speaker.

Fig. 18.

For each condition (speaker, sentence), each of the 3 pairs of {A,B,C} was presented in both orders. After each pair, listeners were asked to indicate within 5 s a binary preference between the first and the second sample. 24 listeners participated in the evaluation.

For each speaker and each pair {X,Y} of {A,B,C} we computed the percentage of preferences for X over Y. From these percentages we derived the relative positions of A, B and C on Thurstone preference scales (Bonnet, 1986), which are represented in Fig. 19. The construction of these scales is detailed in Appendix A.

For male speakers, listeners do not exhibit a clear preference between reshaped and not reshaped noises, and clearly prefer (with a score of 80%) B or C to A. In contrast, the output signal C (equalized, with reshaped noise) is the most disliked for female speakers. For speaker F4, B is preferred to A with the same amplitude as for male speakers. F3 is the only speaker for whom B is not preferred to A (A preferred with 57%, which is however not a significant majority).

The objective results of noise shaping have shown that for some speakers, the noise shaping methods result in sporadically non-masked noise. According to the subjective results shown here, it seems that the permanent white noise is preferred to the sporadic colored noise resulting from the spectral shaping, so that B is retained for the second test. The second observation is that for most of the speakers, even without knowing the original voice, listeners prefer the equalized voice to the non-equalized, in spite of the quantization noise induced by the equalizer. Then the second test aims at evaluating to what extent the knowledge of the original voice modifies the preferences between A and B.

4.2.4 Second test: preferences knowing the original voice

For each condition (speaker, sentence), we presented to listeners the series of pairs OA OB OA and OB OA OB. Listeners knew the structure pair 1 - pair 2 - pair 1 of the series. They indicated for each series in which pair (1 or 2), the degradation of the second sample relatively to the first one is the least annoying. 24 listeners participated in the evaluation.

Table 1 presents the preferences for B over A, compared to those of the first test. For all speakers, the knowledge of the original voice increases the preference for B, and particularly for speaker F3. So, for all the tested speakers, listeners clearly prefer a noisy voice with a timbre close to the original one than the classical telephone voice without noise.

Table 1
Preferences for B over A according to the speaker and the test.

Speaker	1^{st} test	2^{nd} test
M3	79.2%	91.2%
M4	75.0%	75.0%
F3	42.7%	68.7%
F4	71.9%	79.2%

5 Multi-referenced equalization

In order to reduce as far as possible the residual distortion, after equalization, due to inappropriate reference spectrum, we propose to classify speakers in different classes according to their long term spectra and to use one reference spectrum for each class (the center of the class) instead of the same reference spectrum for the whole population. (Mahé and Gilloire, 2003)

5.1 Definition of classes

The smoothing of $|EQ|$ means that the long term spectrum of the processed signal is matched to the reference spectrum with a low spectral resolution. Consequently, the classes have to be defined on the basis of the global shapes of the long term spectra. That is why the classification is performed in the space of the first cepstral coefficients (excluding C_0), the dimension of the space depending on the desired spectral resolution.

Since the equalization algorithm only takes into account the frequencies in a

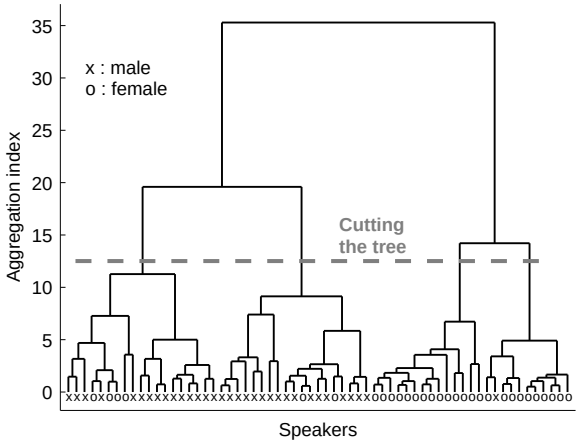

Fig. 20. Classification tree of the 63 speakers.

limited band $[F_1 \dots F_2]$, we define the *long term partial cepstrum* as the cepstral representation of the long term spectrum limited to a frequency band $[F_1 \dots F_2]$. Denoting by k_1 and k_2 the frequency bins corresponding respectively to F_1 and F_2, and by γ the long term spectrum, the partial cepstrum is then defined by:

$$C^{\mathrm{p}} = \mathrm{DFT}^{-1}\Big(10 \log_{10}\left(\gamma(k_1 \dots k_2) \circ \gamma(k_2 - 1 \dots k_1 + 1)\right)\Big) \tag{16}$$

where $\circ$ denotes concatenation in the spectral domain. Speakers are represented by the coefficients 1 to 20 of their partial cepstra, computed with $F_1 = 187$ Hz and $F_2 = 3187$ Hz. Classes are built according to a clustering algorithm, using the generalized Ward criterion of aggregation (see App. B), which consists in minimizing the intra-class variance.

In our experiments we have used a database of 63 speakers (33 male and 30 female), each of them pronouncing a text corresponding to utterances with duration from 23 to 52 s. Speech was recorded in quiet environment with a high quality microphone and is henceforth assumed to be representative of original speech (as at the input of a telephone link). The resulting classification tree of this database is represented in Fig. 20. This tree, having high index gaps about the value 12, clearly shows four classes. These classes are sexually quite homogeneous and a two-classes scission of the tree matches approximately a male/female classification.

The k-means algorithm (Lebart et al., 2000a) initialized with the centers of the four classes reduces the intra-class variance and leads to sexually more homogeneous classes. Figure 21 represents the spectra, limited to 187-3187 Hz, corresponding to the centers of the four sub-classes. These spectra will be used as reference spectra.

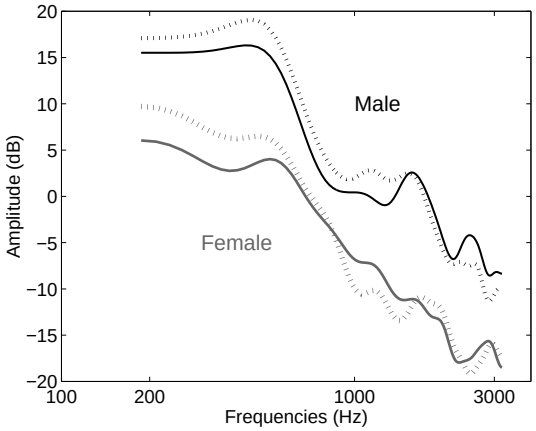

Fig. 21. Reference spectra of the four subclasses.

5.2 Multi-referenced vs mono-referenced equalization

The equalization method presented in Section 3 was simulated for various telephone links, using the database of Subsection 5.1 as input signals. For each speaker, three equalizers were compared to the ideal equalizer, using as references the spectra corresponding respectively to:

- the center of the whole corpus in the space of the partial cepstrum (1 class);
- the center of the sex-class of the speaker (2 classes);
- the center of the subclass of the speaker (4 classes).

The reception signals are denoted respectively by $\mathrm{RX_{EQ1}}$, $\mathrm{RX_{EQ2}}$ and $\mathrm{RX_{EQ4}}$. The output signal of the link equalized by the ideal equalizer is denoted by $\mathrm{RX_{ID}}$.

The spectral distortion between each reception signal $\mathrm{RX_{EQ}}_i$ ($i \in \{1, 2, 4\}$) and the original signal in the equalization band is measured as in Section 3 by the cepstral error e_i, defined as the cesptral distance between $\mathrm{RX_{ID}}$ and $\mathrm{RX_{EQ}}_i$. Fig. 22 and 23 compare the time averages of these errors. For each speaker, the time average of e_i, denoted $\overline{e_i}$, is computed from 10 s of speech activity, to ensure that it reflects the spectral distortion after the equalizer has reached its steady state. In these figures, each speaker is represented by a point of coordinates $(\overline{e_i}, \overline{e_j})$. For most of the speakers, $\overline{e_2}$ and $\overline{e_4}$ are inferior to $\overline{e_1}$, which means that the use of different reference spectra adapted to the classes of speakers reduces the spectral distortion in the equalization band. Using four classes instead of two amplifies this reduction. Table 2 compares the mean values and the standard deviations of the cepstral errors according to the number of classes. The improvement is marginal if two classes are used, but noticeable for four classes. The main interest is that the improvement is the greatest for the speakers with the highest values of e_1, as shown by Fig. 23

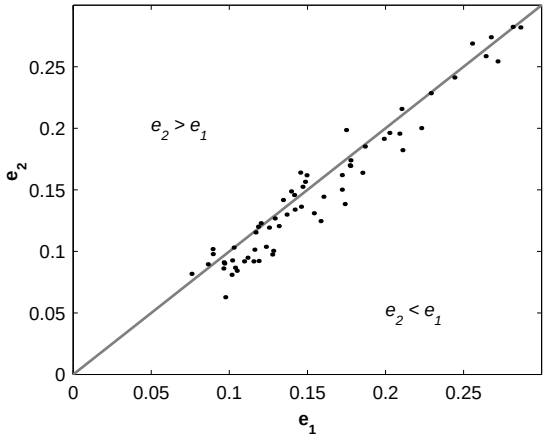

Fig. 22. Comparison of cepstral errors when using 1 or 2 reference spectra.

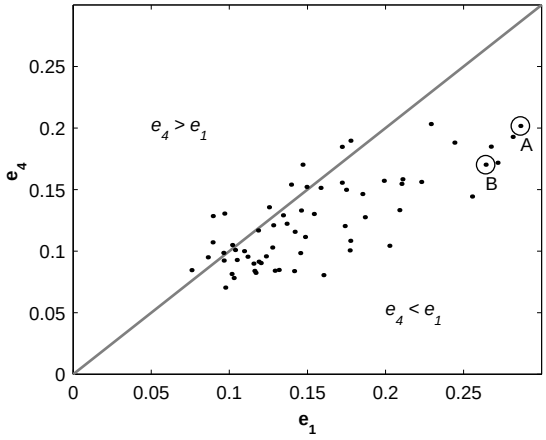

Fig. 23. Comparison of cepstral errors when using 1 or 4 reference spectra.

Table 2
Mean values and standard deviations of the cepstral error according to the number of classes.

Number of classes	$\overline{\overline{e}}$	$\sigma_{\overline{e}}$
1	0.1553	0.0543
2	0.1477	0.0570
4	0.1250	0.0360

Since our goal is to improve the timbre correction, the interest of this objective error reduction lies in its subjective impact. In the case of a male/female classification, informal expert-listening did not reveal timbre difference between $\mathrm{RX}_{\mathrm{EQ2}}$ and $\mathrm{RX}_{\mathrm{EQ1}}$, even for the highest error reductions. Consequently, only the interest of the four class classification was formally evaluated.

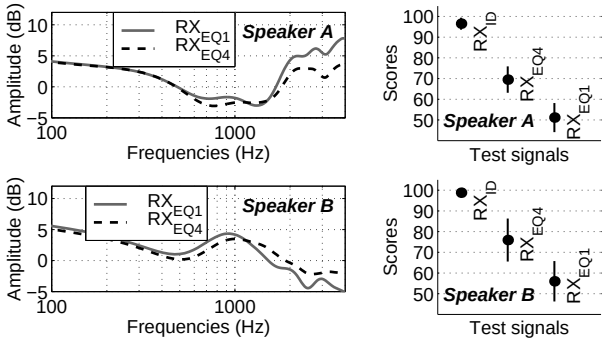

Fig. 24. Comparing $\mathrm{RX}_{\mathrm{EQ1}}$ and $\mathrm{RX}_{\mathrm{EQ4}}$ to $\mathrm{RX}_{\mathrm{ID}}$: mean scores and corresponding spectral distortions.

For the speakers marked by circles in Fig. 23, and denoted by A and B in the sequel, the improvement was evaluated by a subjective test, using the timbre comparison method used in Section 3. The demonstrative interest of these speakers lies in the high values of e_1, which make the error reduction particularly appreciable if it means an improvement of the timbre correction.

For each speaker, a sequence of 7 s of the processed speech, taken after convergence of the equalizer, was presented to a group of 18 expert-listeners. The listeners compared $\mathrm{RX}_{\mathrm{ID}}$, $\mathrm{RX}_{\mathrm{EQ1}}$ and $\mathrm{RX}_{\mathrm{EQ4}}$ (test signals) to $\mathrm{RX}_{\mathrm{ID}}$ (reference) and marked each test signal according to its timbre proximity with the reference signal. The mean scores and confidence intervals are presented in Fig. 24 (right column), facing the corresponding spectral distortions (left column). These subjective results confirm the improvement of the correction of the spectral distortion when using the multi-referenced equalization: $\mathrm{RX}_{\mathrm{EQ4}}$ is perceived as closer to $\mathrm{RX}_{\mathrm{ID}}$ than $\mathrm{RX}_{\mathrm{EQ1}}$. While not very large, the timbre difference between $\mathrm{RX}_{\mathrm{EQ4}}$ and $\mathrm{RX}_{\mathrm{EQ1}}$ is significant.

5.3 Classification of speakers in operational conditions (on-line)

5.3.1 Classification method

The evaluation of the new equalization method presented in the previous section assumes a perfect knowledge of the classes of the speakers. In the network, a speaker cannot be classified simply according to the distances between its partial cepstrum and the centers of the different classes, since this cepstrum is modified by the part of the telephone link preceding the equalizer. According to this deviation, robust classification criteria must therefore be defined, in order to ensure the practical relevance of the multi-referenced equalization.

The robustness is obtained first by the choice of the classification parameters. The classes defined in subsection 5.1 are sexually homogeneous. Since the

mean pitch value is both sexually discriminating and robust to the spectral distortions caused by a telephone link, we use it as a classification parameter, in addition to the partial cepstrum. So, each speaker is now represented by a vector $x = [\overline{F_0}; C^{\mathrm{p}}(1); \ldots; C^{\mathrm{p}}(20)]$, where $\overline{F_0}$ denotes the time average of pitch. For each class q, the pitch component of the center $\overline{x}^q$ is the mean of time average pitches of the speakers of class q.

Since the classes turn out to be fairly well grouped around their centers, we use the linear discriminant analysis (LDA) to classify the speakers. Considering a population of a p-dimensional space, classified *a priori* in K classes, the LDA consists in:

- first, in a training database, finding the $K-1$ independant linear functions that lead to the best separation of the K classes in the $(K-1)$-dimensionnal space they generate, by minimizing the intra-classes variance. This step is detailled in Appendix C.
- then, in a test database, denoting by a the discriminant function with values in $\mathbb{R}^{K-1}$, classifying a new observation x in the class q maximizing the conditional probability of q knowing $a(x)$, denoted by $\Pr(q|a(x))$.

  According to Bayes theorem, $\Pr(q|a(x))$ is proportional to $\Pr(a(x)|q)\Pr(q)$. Assuming a gaussian distribution of $a(x)$ in each class, the probability density of $a(x)$ in the class q is defined by:

$$f_q(x) = \frac{1}{(2\pi)^{\frac{K-1}{2}}\sqrt{|S_q|}} \exp\Big(-\frac{1}{2}(a(x)-a(\overline{x}^q))' S_q^{-1}(a(x)-a(\overline{x}^q))\Big) \quad (17)$$

  where S_q is the covariance matrix of a in the class q, which generic element σ_{ij}^q is defined by:

$$\sigma_{ij}^q = \mathrm{E}[(a_i(x)-a_i(\overline{x}^q))(a_j(x)-a_j(\overline{x}^q))] \quad (18)$$

  The class q of speaker x is the one maximizing $f_q(x)\Pr(q)$, *i.e.* minimizing the discriminative score $s_q(x)$ defined by:

$$\begin{aligned} s_q(x) = {} & (a(x)-a(\overline{x}^q))' S_q^{-1}(a(x)-a(\overline{x}^q)) \\ & + \log(|S_q|) - 2\log(\Pr(q)) \end{aligned} \quad (19)$$

In addition to the use of pitch, the robustness of the classification function a is insured by choosing a training database of speakers which voices are affected by a large variety of spectral distortions, representative of the distortions caused by telephone links.

5.3.2 Training- and test-databases

For the validation of this method, we need a training database and a test database, in which the classes of the speakers are *a priori* known. Since the

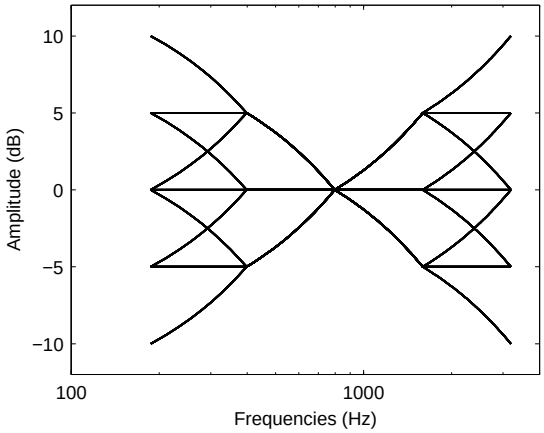

Fig. 25. Set of spectral distortions added to the clean speech database to build the training database.

class of a speaker can be known without error according to its partial cepstrum computed from clean speech recorded in the same conditions as the database used to define the classes, the databases have to be constituted of clean speech samples recorded in these conditions. Moreover, the samples must be long enough to allow the computation of the long term parameters. Because of the difficulty of finding two large databases complying with these conditions, or one database large enough to be split, we use the database of Subsection 5.1 (63 speakers) as follows. The classification functions are

- learned with this database modified by various filterings;
- tested with the samples of this database as inputs of simulated telephone links.

The learning database is built by adding to the partial cepstrum of each speaker of the original database a set of M cepstral biases, which results in a new database of $63M$ samples. We have derived these cepstral biases from the set of $M = 81$ spectral distortions represented in Fig. 25 in the band 187-3187 Hz: each frequency response corresponds to a path from left to right in the lattice. This set covers approximately the continuum of frequency responses of the analog transmission channels allowed by the models of line and sending system used.

Note that in the space of the partial cepstrum, the training-database and the test-database are really different. On one hand, the added cepstral biases are of the same order as the differences between the partial cepstra of the speakers in the original database. On the other hand, the cepstral biases used to build the training database do not include the transfer functions of the links simulated with the test database, although they are in the same range.

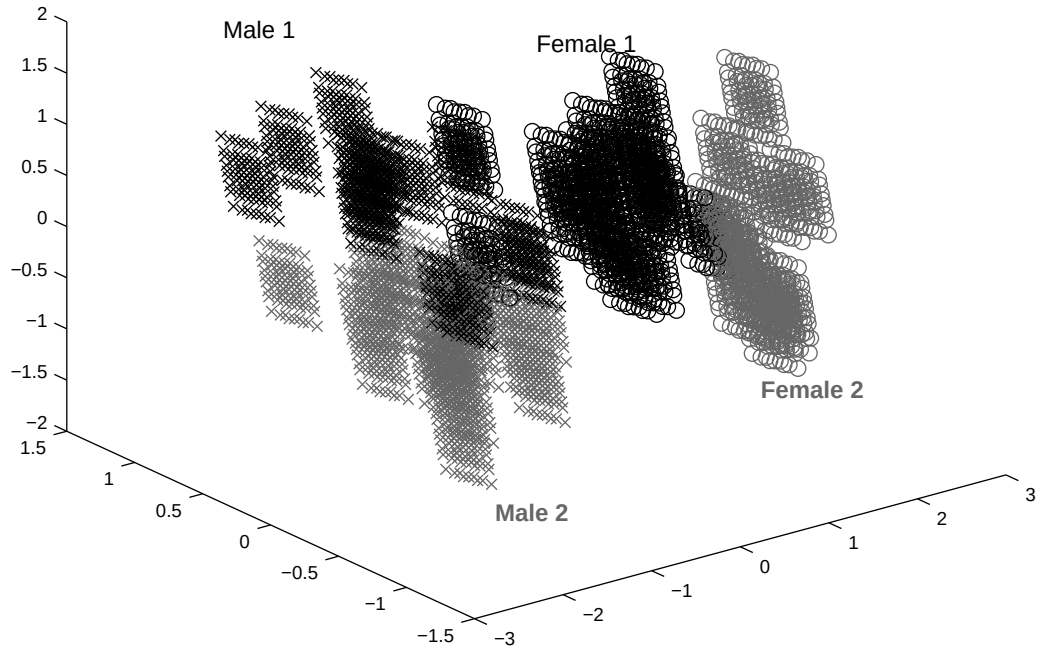

Fig. 26. Projection of the training database in the 3D-space generated by the discriminant functions.

5.3.3 Definition of classification criteria

The linear discriminant analysis with the training database leads to three linear functions of discriminating capacities 0.95, 0.50 and 0.43. The pitch mean value plays a major role in the discrimination, since its normalized multiplicative coefficient is the highest in each function.

The population of the training database is represented in Fig. 26, in the 3D-space generated by the discriminant functions. The apparent classification error, computed among this population, is of 0 % for the two female sub-classes, and 4.5 and 11 % for the two male sub-classes respectively, which partially overlap each other. Male and female classes do not overlap each other.

5.3.4 Classification of the test-database

The classification criteria were tested with the samples of the database of Subsection 5.1 as inputs of various simulated telephone links.

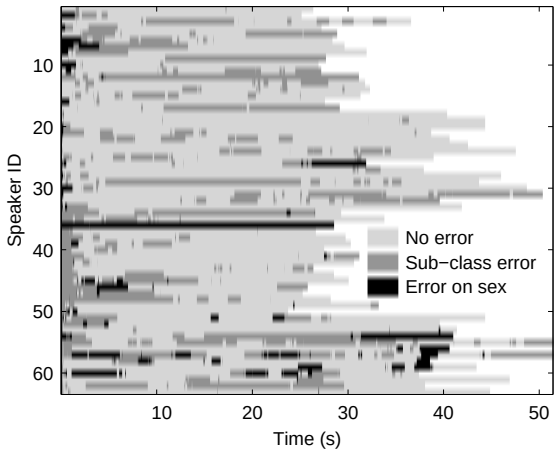

Fig. 27. For each speaker and each frame, classification error.

At each frame, the partial cepstrum is derived from the long-term spectrum of the speaker, updated according to (9). The mean pitch value is updated at each voiced frame, according to:

$$\overline{F_0}(m) = \alpha(m)F_0(m) + (1 - \alpha(m))\overline{F_0}(m-1) \tag{20}$$

where $F_0(m)$ is the pitch of the m^{th} voiced frame. So, at each frame, the vector x is updated and the discriminative scores are computed to update the classification of the speaker.

The resulting classification errors are illustrated in Fig. 27 for the same simulated link as in Subsection 3.4. Each horizontal line corresponds to a speaker. Each pixel in a line corresponds to a frame: the pixel is light gray if the frame is well classified, black if the frame is classified in a class of the wrong sex, dark gray if only the decision on the sub-class is wrong. The percentage of classification error after 10 s of speech (*i.e.* after stabilization of x) is 24 %. Note that the two speakers who were subjectively tested are correctly classified in this experiment (speakers 4 and 18 in Fig. 27).

5.3.5 Influence of classification error on equalization results

If the classification error results only from the distortion of the original long term spectrum by the link, then it leads to a wrong approximation of this spectrum by the reference spectrum. Consequently, the classification error induces here a higher cepstral error than the mono-referenced equalization. In contrast, if the original partial cepstrum itself of a wrongly classified speaker is close to the center of the chosen class, then the classification error does *a priori* not severely decrease the performance of the multi-referenced equalization.

That is why the on-line classification should be evaluated according to its

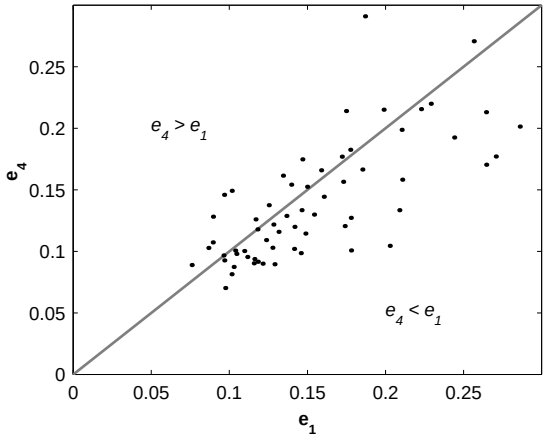

Fig. 28. Comparison of cepstral errors when using 1 or 4 reference spectra, in the case of on-line classification.

influence on cepstral errors rather than to the classification error rate. The simulations of Subsection 5.2 were rerun, using at each frame the reference spectrum given by the above on-line classification. Fig. 28 compares the time averages of cepstral errors e_1 (mono-referenced equalization) and e_4 (multi-referenced equalization using four classes). Comparing figures 23 and 28 shows that the classification errors has finally little effects on the performance, which confirms the interest of multi-referenced equalization.

6 Conclusions

For most speakers, the proposed blind spectral equalizer leads to a low spectral distortion in the equalization band, between transmitted and received signals. From a perceptual point of view, this means a satisfactory timbre restoration, within the limits of the equalization band.

Although the combination of equalization and A-law quantization induces a noticeable quantization noise, the formal evaluation of the global perception of timbre and noise showed that listeners prefer a noisy voice with a timbre close to the original one than the usual telephone voice without noise.

We have shown that speakers can relevantly be classified on the basis of their long-term spectrum, which allows us to use a specific reference spectrum for each class in the equalization algorithm, instead of the same reference for the whole population. This refinement of the equalizer reduces the spectral distortion between the equalized received speech and the transmitted speech, even if classification errors occur, which perceptually means a slightly but significantly better timbre correction for some speakers, as it was demonstrated in a subjective test.

So, we have proposed a blind equalizer that approximates with a good accuracy the original long term spectrum of a speaker across a telephone link and, thus, restores the timbre of this speaker at the output of the link, with reasonable noise, within a chosen equalization band. We have presented here the case of the PSTN; note that experiments done with the GSM led to similar results. This equalizer was successfully implemented in real-time in an experimental PABX. Moreover, its practical interest came to two patents (Mahé and Gilloire, 2001b, 2002b).

A Thurstone preference scales

We consider a set of stimuli which have been compared in pairs. Given for each pair of stimuli XY the percentage of preference for X over Y, we want to represent these results by positioning the whole set of stimuli on a preference scale.

Thurstone assumes that each magnitude of a stimulus S_i results in a judgement s_i, called *discriminative process*, considered as a random value. A subject prefers S_i to S_j if $s_i > s_j$. In terms of probabilities,

$$\Pr(S_i > S_j) = \Pr(s_i - s_j > 0) \tag{A.1}$$

Thurstone assumes the discriminative processes have a gaussian distribution. Considering a pair of stimuli S_i and S_j, which discriminative processes are denoted by s_i and s_j respectively, the difference $s_i - s_j$ is a gaussian random value too. Its mean value is $\mu(s_i) - \mu(s_j)$, where $\mu(s_i)$ and $\mu(s_i)$ denote the mean values of s_i and s_j respectively. Knowing the probability of preference for S_i over S_j, $\Pr(S_i > S_j)$, $\mu(s_i)$ and $\mu(s_j)$ are related by:

$$z(S_i > S_j) = \frac{\mu(s_i) - \mu(s_j)}{\sigma(s_i - s_j)} \tag{A.2}$$

where $z(S_i > S_j)$ is the standard normal random variable corresponding to $\Pr(S_i > S_j)$ and $\sigma(s_i - s_j)$ denotes the standard deviation of $s_i - s_j$. Assuming the discriminative processes are independents and have the same variance σ^2:

$$\sigma^2(s_i - s_j) = 2\sigma^2 \tag{A.3}$$

The relation between $\mu(s_i)$ and $\mu(s_j)$ becomes:

$$\mu(s_i) - \mu(s_j) = z(S_i > S_j)\sigma\sqrt{2} \tag{A.4}$$

Consequently, if we represent each stimulus S_k on a scale by the mean value of its discriminative process s_k, Eq. (A.4) gives the relative positions of S_i and S_j. So, for a set of N stimulus $\{S_i\}_{1\leq i\leq N}$, we can represent each stimulus S_i on the preference scale by the average value of $(\mu(s_i) - \mu(s_k))_{k\neq i}$.

B Clustering using the Ward criterion of aggregation

Given a population of N units, clustering consists in creating a hierarchy of partitions according to the following process: at each step the two closest elements are aggregated, where an element is either an isolated unit or an aggregation of units created at a previous step. The proximity between two elements is defined by a dissimilarity measure called *distance*. The processed goes on until the whole population is aggregated.

The resulting hierarchy of $N-1$ partitions can be represented by a tree like that of Fig. 20, where the height of an horizontal segment aggregating two elements is proportional to the distance between these two elements, called "*aggregation index*". This representation helps to choose a relevant partition: the tree must be cutted above the aggregations of close elements and below the aggregations of distants elements. Consequently, a relevant partition results from cutting the tree at a high gap of aggregation indexes.

The behavior of a clustering algorithm relies on the definition of the *distance*. The *generalized Ward criterion of aggregation*, or *variance criterion* (Lebart et al., 2000a), is based on the idea that the classes have to be homogeneous, which implies to minimize the intra-classes variance.

Considering a population of points x_i with respective weights m_i, classified in classes q with respective barycenters g_q, the intra-classes inertia is defined by:

$$I_{\text{intra}} = \sum_q \sum_{i\in q} m_i \|x_i - g_q\|^2 \tag{B.1}$$

The intra-classes inertia is equal to zero at the initialization of the algorithm and increases unavoidably at each step of the algorithm. The *variance criterion* consists in aggregating at each step the two elements which aggregation leads to the lowest increase of the intra-classes inertia. This increase is therefore taken as distance.

The aggregation of two elements x_i and x_j, of respective weights m_i and m_j, results in a element x of weight m which is their barycenter, defined by:

$$x = \frac{m_i x_i + m_j x_j}{m_i + mj},\ m = m_i + m_j \tag{B.2}$$

Lebart et al. (2000a) show that the resulting increase of intra-classes inertia is:

$$\Delta I_{ij} = \frac{m_i m_j}{m_i + m_j} \|x_i - x_j\|^2 \tag{B.3}$$

This increase defines the *distance* between two elements. So, at each step of the algorithm, we aggregate the two elements x_i and x_j minimizing ΔI_{ij}.

C Linear discriminant analysis: finding the discriminant functions

Considering a population of N units in a p-dimensionnal space, classified *a priori* in K classes, we search the linear combination of their components that maximize the inter-classes variance and minimize the intra-classes variance.

For a unit x^k, let $a(x^k)$ be a linear combination of the components of x^k preliminarily centered:

$$a(x^k) = \sum_{i=1}^{p} a_i.(x_i^k - \overline{x}_i) \tag{C.1}$$

where $\overline{x}$ is the center of gravity of the population. The variance of $a(x)$ is:

$$\sigma_a^2 = a'Ta \tag{C.2}$$

where $T = [(t_{ij})_{1 \leq i,j \leq p}]$ is the covariance matrix of x, which elements can be estimated by:

$$t_{ij} = \frac{1}{N} \sum_{k=1}^{N} (x_i^k - \overline{x}_i)(x_j^k - \overline{x}_j) \tag{C.3}$$

This matrix can be written as $T = D + E$, where $D = [(d_{ij})_{1 \leq i,j \leq p}]$ is the intra-classes inertia matrix and $E = [(e_{ij})_{1 \leq i,j \leq p}]$ the inter-classes inertia matrix, which elements are estimated respectively by:

$$\begin{aligned} d_{ij} &= \frac{1}{N} \sum_{q=1}^{K} \sum_{k \in q} (x_i^k - \overline{x}_i^q)(x_j^k - \overline{x}_j^q) \\ e_{ij} &= \sum_{q=1}^{K} \frac{N_q}{N} (\overline{x}_i^q - \overline{x}_i)(\overline{x}_j^q - \overline{x}_j) \end{aligned} \tag{C.4}$$

where $\overline{x}^q$ and N_q denote the center of gravity and the cardinal of the q^{th} class, respectively. So, the variance of $a(x)$ can be written as the sum of internal and external variances:

$$\sigma_a^2 = a'Ta = a'Da + a'Ea \tag{C.5}$$

Minimizing the intra-classes variance and maximizing the inter-classes variance consists in maximizing $f(a)$ defined as:

$$f(a) = \frac{a'Ea}{a'Ta} \tag{C.6}$$

Since f is invariant when a is multiplied by any scalar value, this problem is equivalent to the maximization of $a'Ea$ under the constraint $a'Ta = 1$. The resolution of this problem leads to the following relationship:

$$Ea = \lambda Ta \tag{C.7}$$

If T is invertible, we get:

$$\begin{aligned} &T^{-1}Ea = \lambda a \\ &a'Ea = \lambda a'Ta = \lambda \end{aligned} \tag{C.8}$$

which implies that a is the eigenvector of $T^{-1}E$ corresponding to the higher eigenvalue. This eigenvalue is called the *discriminating capacity* of the linear function a.

For a classification in K classes, the $K-1$ discriminant functions correspond to the $K-1$ higher eigenvalues of $T^{-1}E$.

Acknowledgements

We thank gratefully Delphine Charlet (France Télécom R&D) for her advice in classification techniques and Alain Le Guyader (France Télécom R&D) for his precious help about noise shaping. Thanks to Mamadou Mboup (University of Paris V / CRIP5) for his careful re-reading of this paper. We also thank the reviewers for their careful analysis and useful comments.

References

Boite, R., Bourlard, H., Dutoit, T., Hancq, J., Leich, H., 2000. Codage, in: Traitement de la parole, Presses polytechniques et universitaires romandes, Lausanne, pp 99–174.

Bonnet, C., 1986. Les échelles psychophysiques, in: Manuel Pratique de Psychophysique, Armand Colin, Paris, pp. 136–142.

Bowker, D.O., Ganley, J.T., James, J.H., 1993. Telephone network speech signal enhancement. AT&T Bell Laboratories, US patent 5333195.

De Jaco, A.P., Miller, J.A., 1997. Adaptive equalizer preprocessor for mobile telephone speech coder to modify non ideal frequency response of acoustic transducer. US patent 5915235.

Faucon, G., Le Bouquin, R., Abkari Azirani, A., 1993. Mesures objectives de la réduction de bruit. In: Proc. GRETSI'93, Juan-les-Pins, France, 587–590.

Ho, H.S., Pratt, M.K., Lim, P.C., Oshidari, T.T., 1993. Voice enhancement system and method. DSC Communications Corporation, US patent 5471527.

ITU-R, 2001. Recommendation BS.1534: Method for the subjective assessment of intermediate quality level of coding systems.

ITU-T, 1993. Recommendation P.50: Artificial voices.

ITU-T, 1996. Recommendation P.830: Subjective performance assessment of telephone-band and wideband digital codecs, annex D.

ITU-T, 1999. Recommendation P.313: Transmission characteristics for cordless and mobile digital terminals.

ITU-T, 2000. Recommendation P.310: Transmission characteristics for telephone-band (300-3400 Hz) digital telephones.

Johnston, J.D., 1988. Transform coding of audio signals using perceptual noise criteria. IEEE Journal on selected areas in communications, 6(2), 314–323.

Lebart, L., Morineau, A., Piron, M., 2000. Quelques méthodes de classification, in: Statistique Exploratoire Multi-dimensionnelle, Dunod, Paris, pp 145–185.

Lebart, L., Morineau, A., Piron, M., 2000. Liens avec les méthodes explicatives usuelles, in: Statistique Exploratoire Multi-dimensionnelle, Dunod, Paris, pp 251–268.

Mahé, G., Gilloire, A., 2001. Correction of the voice timbre distortions on telephone network. In: Proc. Eurospeech'2001, Aalborg, Denmark, 1867–1870.

Mahé, G., Gilloire, A., 2001. Procédé et dispositif de correction centralisée du timbre de la parole sur un réseau de communications téléphoniques. France Telecom R&D, French patent 0104194

Mahé, G., Gilloire, A., 2002. Quantization noise spectral shaping in instantaneous coding of spectrally unbalanced speech signals. In: Proc. IEEE Workshop on Speech Coding, Tsukuba, Japan, 56–58.

Mahé, G., Gilloire, A., 2002. Procédé et système de correction multi-références des déformations spectrales de la voix introduites par un réseau de communication France Telecom R&D, French patent 0215618

Mahé, G., Gilloire, A., 2003. Multi-referenced correction of the voice timbre distortions in telephone networks. In: Proc. Eurospeech'2003, Genève, Suisse, pp. 1381-1384.

Makhoul, J., Berouti, M., 1979. Adaptative noise spectral shaping and entropy coding in predictive coding of speech. IEEE Transactions on acoustics, speech, and signal processing, ASSP-27(1), 63–73.

Mokbel, C., Monné, J., Jouvet, D., 1993. On-line adaptation of a speech recognizer to variations in telephone line conditions. In: Proc. Eurospeech'93, Berlin, 1247–1250.

Mokbel, C., Jouvet, D., Monné, J., 1996. Deconvolution of telephone line effects for speech recognition. Speech Communication, 19(3),185–196.

National Semiconductor, 1994. Technical documentation: TP3054, TP3057 - Enhanced Serial Interface - CODEC/Filter COMBO Family.

Tukey, J.W., 1953. The problem of multiple comparisons, Ditto, Princeton University Press, Princeton.