Multi-referenced Correction of the Voice Timbre Distortions in Telephone Networks

Gaël Mahé[*] *and André Gilloire*[**]

[*] Université René Descartes - Paris V, 45 rue des Saints-Pères 75270 Paris Cedex 06, France
[**] France Télécom R&D / DIH / IPS 2, avenue Pierre Marzin, 22307 Lannion cedex, France
andre.gilloire@rd.francetelecom.com

Abstract

In a telephone link, the voice timbre is impaired by spectral distortions generated by the analog parts of the link. We first evaluate from a perceptual point of view an equalization method consisting in matching the long term spectrum of the processed signal to a reference spectrum. This evaluation shows a satisfying restoration of the timbre for most speakers. For some speakers however, a noticeable spectral distortion remains. That is why we propose a multi-referenced equalizer, based on a classification of speakers and using a different reference spectrum for each class. This leads to a decrease of the spectral distortion and, as a consequence, to a significant improvement of the timbre correction.

1. Introduction

In this paper, we define the timbre as the subjectively significant long term spectral characteristics of the voice of a given speaker. 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 terminals and line connections to the local exchange) in the transmission and reception paths. For the PSTN, this filtering is described by the *modified Intermediate Reference System* (IRS) [1]. The frequency responses and masks of the sending and receiving parts of the modified IRS, respectively called *sending* and *receiving systems*, are defined in [1].

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 the gain, in dB, is proportional to the square root of the frequency:

$$H_{dB}(f) = H_{dB}(800\text{Hz})\sqrt{\frac{f}{800}} \qquad (1)$$

where the gain (in dB) at 800 Hz, H_{dB}(800Hz), is –3 dB for an average line and –9.5 dB for the longest lines. This low-pass filtering muffles the voice in the case of a long line.

A blind spectral equalizer, placed in the digital part of the network (see Fig. 1), was proposed in [2] to compensate for these distortions and restore a timbre as close as possible to that of the original voice of the speaker. This method consists in matching the spectrum of the processed signal to a given reference spectrum and it provides a significant improvement of the timbre of the received voice; nevertheless, a noticeable spectral distortion remains for some speakers. We propose in this paper to improve the timbre correction by classifying speakers in several classes and using one reference spectrum per class instead of the same reference spectrum for the whole population, *i.e.* we propose a *multi-referenced equalizer*. In Section 2, we evaluate objectively and subjectively the former equalizer [2]. In Section 3, we describe and we evaluate in depth the multi-referenced equalizer .

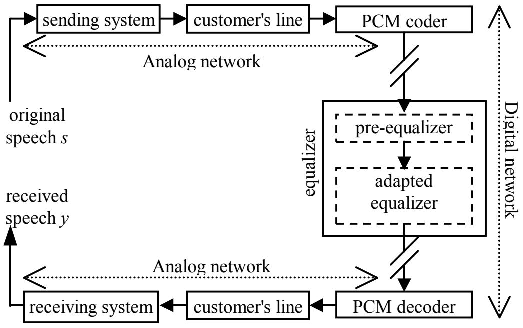

Figure 1: Telephone link with equalizer.

2. Blind spectral equalization

2.1. Principles and implementation

The equalizer presented in [2] combines two filters. The first one is a fixed filter, called *pre-equalizer,* which frequency response is the inverse of the global response of the average analog channel in the band [F_c - 3150 Hz]. The lower cut-off frequency F_c is fixed in order to avoid amplification of components with low SNR. The average analog channel is defined as the combination of average customer analog lines and sending and receiving systems having frequency responses according to the nominal response of the modified IRS [1].

This filter is completed by an *adapted equalizer*, in order to adapt the global correction of the equalizer to various conditions of transmission. The frequency response of this second filter is computed as follow. Denoting G the frequency response of the analog channel in series with the pre-equalizer, $|S(f)|^2$ the short term PSD of the original signal s and $|Y(f)|^2$ the short term PSD of the received signal y,

$$|Y(f)|^2 = |G(f)|^2 |S(f)|^2 . \qquad (2)$$

If the channel is assumed to be time-invariant, the time averages are related by:

$$\overline{|Y(f)|^2} = |G(f)|^2 \overline{|S(f)|^2} , \qquad (3)$$

The frequency response of the adapted equalizer is therefore defined by:

$$|EQ(f)| = 1/|G(f)| = \sqrt{\frac{\gamma_s(f)}{\gamma_y(f)}}\,, \qquad (4)$$

where γ denotes the long term spectrum of a signal, defined as the time average of the short term spectrum.

Since the long term spectrum of the original voice $\gamma_s(f)$ is unknown, it is approximated in [2] by the average spectrum of speech defined by the ITU [3], called reference spectrum and denoted $\gamma_{\text{ref}}(\text{f})$. Moreover, since the equalizer is placed inside the network, $\gamma_y(f)$ is not directly available and therefore it must be derived from the output of the pre-equalizer, x, using:

$$\gamma_y(f) = |L\_RX(f)|^2 |S\_RX(f)|^2 \gamma_x(f)\,, \qquad (5)$$

where L_RX is the frequency response of the receiving line and S_RX is the frequency response of the receiving system.

The frequency response of the adapted equalizer becomes:

$$|\,EQ(f)\,| = \frac{1}{|S\_RX(f).L\_RX(f)|}\sqrt{\frac{\gamma_{\text{ref}}(f)}{\gamma_x(f)}} \qquad (6)$$

Only the band F_c -3150 Hz is equalized; the values of $|EQ|$ out of this band are therefore replaced by a linear extrapolation of $|EQ|_{[Fc\text{ -3150 Hz}]}$, before deriving the impulse response from $|EQ|$ by an IFFT.

Because of the roughness of the approximation $\gamma_s(f) = \gamma_{\text{ref}}(f)$, only the global shape of this frequency response is relevant. That is why the frequency response $|EQ|$ must be smoothed. Since we have implemented the equalizer in the time domain, this smoothing is achieved by a narrow windowing of the impulse response.

2.2. Simulation and results

The above equalizer has been simulated with different speakers uttering a text of approximately 20 s of voice activity. 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 [1]. The original speech, denoted ORI, was compared to the signals at the outputs of three links:

- RX_0: link without equalizer;
- RX_{EQ}: equalized link;
- RX_{ID}: link equalized by the ideal equalizer, defined as an equalizer compensating exactly for the spectral distortions on the band Fc -3150 Hz.

Fc was set to 200 Hz. We asked a group of 24 listeners to compare the timbres of the reception signals to the timbre of the original signal ORI according to the MUSHRA method [4]. We replaced the quality scale of the MUSHRA method by a scale of timbre proximity: a test signal is rated 100 if its timbre is the same as the reference timbre, 0 if it is very different from the reference. Only the second half of the text was presented to the listeners, so that the equalizer was evaluated after it had reached its steady state.

The mean scores and confidence intervals are presented in Fig. 2 for two speakers, facing the corresponding spectral distortion curves. For a majority of speakers, of whom speaker A is representative, the equalizer is very close to the ideal one, which results in a nearly flat distortion curve in the band 200-3150 Hz. As a consequence, no significant subjective timbre difference appears between RX_{EQ} and RX_{ID}. For some speakers, like speaker B, the note of RX_{EQ} is significantly closer to ORI than RX_0, but inferior to the note of RX_{ID}. From the observation of the spectral distortion curves, we deduce that this sub-optimal result is due to the inappropriate approximation of the long term spectrum of the speaker B by the ITU reference spectrum [3].

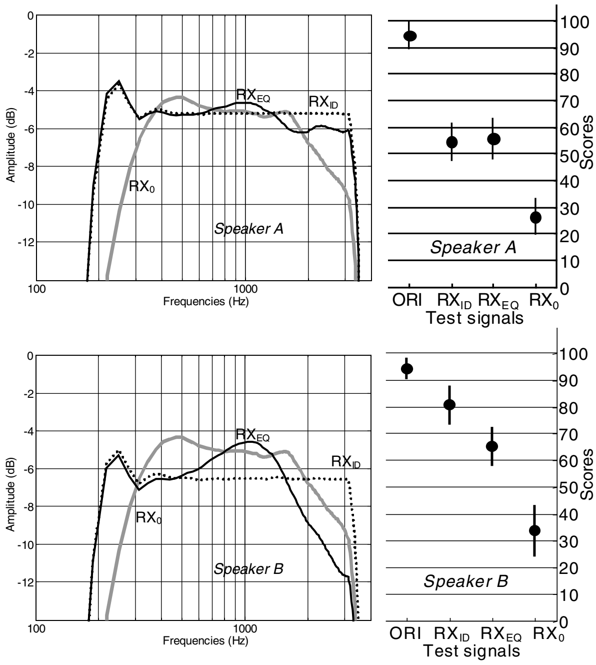

Figure 2: Mean scores and corresponding spectral distortions of outputs of telephone link.

3. Multi-referenced equalization

In order to limit as far as possible the residual distortion 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.

3.1. Definition of classes

The smoothing of $|EQ|$ means that only the spectral envelope of the processed signal is matched to the reference spectrum. Consequently, the classes have to be defined on the basis of the spectral envelope. That is why the classification is performed in the space of the first cepstral coefficients (except C_0), the dimension of the space depending on the desired spectral resolution of the envelope.

Since the equalization algorithm only takes into account the frequencies in a limited band F_1-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-F_2. Denoting k_1 and k_2 the frequency bins corresponding respectively to F_1 and F_2, and γ the long term spectrum, the partial cepstrum is defined by:

$$C^P = \mathrm{TFD}^{-1}\left(10\log\left(\gamma(k_1\ldots k_2)\circ\gamma(k_2-1\ldots k_1+1)\right)\right) \quad (7)$$

where ° 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 [5].

In our experiments we used a database of 63 speakers (33 male and 30 female), each of them pronouncing a text of utterance duration extended from 23 to 52 s. Speech was recorded in quiet environment with high quality microphones and is henceforth assumed to be representative of original speech (as at the input of a telephone link). The resulting classification tree is represented in Fig. 3. This tree, having high index gaps about the value 12, clearly shows four classes. The k-means algorithm [5] initialized with the centers of these classes reduces the intra-class variance and leads to sexually homogeneous classes. Figure 4 represents the spectra, limited to 187-3187 Hz, corresponding to the centers of the classes. These spectra will be used as reference spectra.

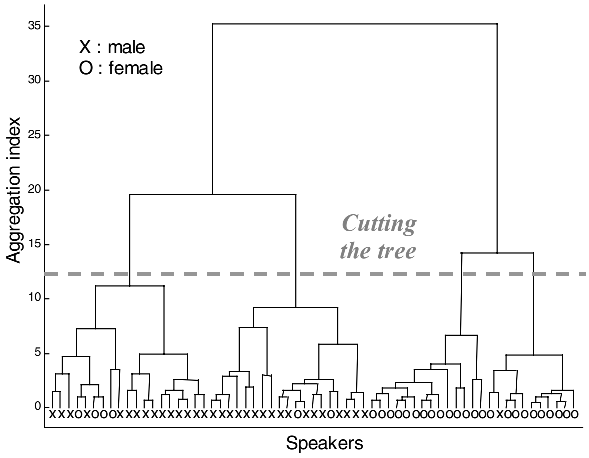

Figure 3: Classification tree of the 63 speakers.

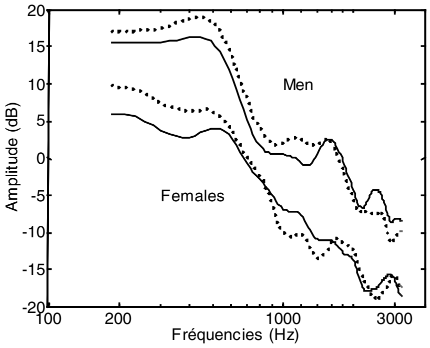

Figure 4: Reference spectra of the 4 classes.

3.2. Improvement of the equalization performance

The equalization method presented in Section 2 was simulated in the same conditions as in Section 2.

Three equalizers are compared for each speaker (reception signals in brackets):

- Equalizer using as reference spectrum the spectrum corresponding to the center of the whole corpus in the space of the partial cepstrum (RX_{EQ1});
- Equalizer using as reference spectrum the spectrum corresponding to the center of the class of that speaker (RX_{EQ4});
- Ideal equalizer (RX_{ID}).

The spectral distortion between the reception signal and the original signal in the equalization band (187-3187 Hz here) can be measured by a cepstral error defined as the cesptral distance between RX_{ID} and the reception signal. The time averages of cepstral errors of RX_{EQ1} and RX_{EQ4}, denoted respectively e_1 and e_4, are compared in Fig. 5: each speaker is represented by a point of coordinates (e_1, e_4). For most of the speakers, e_4 is inferior to e_1, which means that the use of different reference spectra adapted to the classes of speakers leads to a reduction of the spectral distortion in the equalization band.

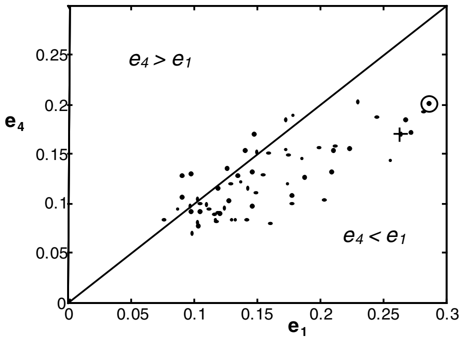

Figure 5: Comparison of cepstral errors when using 1(e_1) or 4 (e_4) reference spectra.

For the speakers marked by a circle and by a cross on Fig. 5, this improvement was evaluated by a formal subjective test using the MUSHRA method [4], modified in the same way as in Section 2. A group of 18 expert-listeners compared RX_{ID}, RX_{EQ1} and RX_{EQ4} (test signals) to RX_{ID} (reference) and noted each test signal according to its timbre proximity with the reference signal.

The mean scores and confidence intervals are presented on Fig. 6, facing the corresponding spectral distortions. These subjective results confirm the objective improvement of the correction of the spectral distortion.

3.3. Classification of speakers in operational conditions

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.

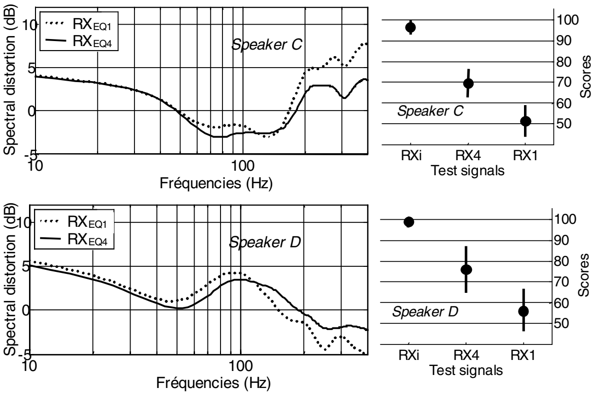

Figure 6: Comparing $\mathrm{RX}_{\mathrm{EQ1}}$ and $\mathrm{RX}_{\mathrm{EQ4}}$ to $\mathrm{RX}_{\mathrm{ID}}$: mean scores and corresponding spectral distortions.

The robustness is obtained first by the choice of classification parameters. The classes defined in Section 3.1 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 represented by a vector $x = [\,\overline{F_0}\,;\, C^p(1)\,;\,\ldots\,;\, C^p(20)\,]$, where $\overline{F_0}$ denotes the time average of pitch. For each class q, the pitch component of the center g_q is the mean of time average pitches of the speakers of class q.

Since the classes are fairly well grouped around their centers, we use the linear discriminant analysis (LDA) technique to classify the speakers. According to the LDA principles, we define a set of linear functions of the vector x minimizing the intra-classes variance. A new observation x is then classified according to a Bayesian criterion, in the class q minimizing the discriminative score s_q:

$$\begin{aligned} s_q(x) = &\left(a(x)-a(g_q)\right)' S_q^{-1}\left(a(x)-a(g_q)\right) \\ &+\log\left(\left|S_q\right|\right)-2\log\left(\mathrm{P}(q)\right) \end{aligned} \qquad (8)$$

where a is the discriminant function with values in $\Re^3$ and S_q is the matrix of covariance of $a(x)$ in the class q.

In addition to the use of pitch, the robustness of these classification functions 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. This training database is derived from the clean speech database of Section 3.1, by adding to each speaker's cepstrum 81 cepstral biases.

The resulting apparent classification error, computed among the training database, is of 0 % for the two female classes, and 4.5 and 11 % for the two male classes.

We tested this criterion with the simulated telephone link presented in section 2.2, using the database of Section 3.1 as input of the chain. At each frame, the vector x is updated and the criterion is applied. The resulting classification errors are illustrated in Fig. 8. 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 of 24 %.

These results show that speakers can be classified "on-line" with reasonable error, which means that the potential reduction of cepstral error illustrated in Fig. 5 is relevant in a practical point of vue. Note that the two speakers that were subjectively tested are correctly classified in this experiment.

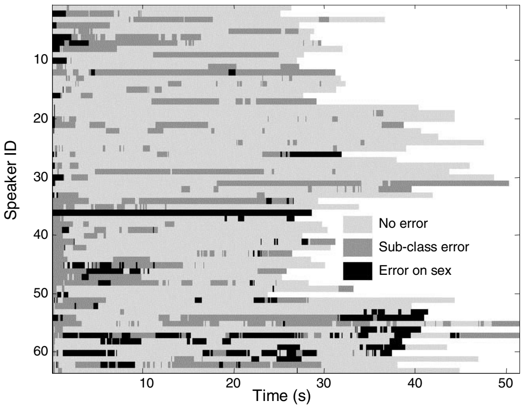

Figure 8: For each speaker and each frame, classification error.

4. Conclusions

The blind spectral equalizer proposed in [2], while performing satisfying restoration of the timbre for most speakers (within the limits of the equalization band), leads to sub-optimal results for some of them. We showed that speakers can be classified on the base of their long term spectra, which led us to replace the single reference spectrum of the previous equalizer by as many reference spectra as classes. The use of multiple references provides a significant objective and subjective improvement of the timbre restoration.

5. Acknowledgements

We thank gratefully Laetitia Gros for her help in subjective evaluations and Delphine Charlet for her advice about classification techniques.

6. References

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

[2] G.Mahé and A. Gilloire, “Correction of the Voice Timbre Distortions on Telephone Network”, in *Proc. Eurospeech*, pp 1867-1870, 2001.

[3] ITU-T, Recommendation P.50, “*Artificial voices,*” 1993.

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

[5] L. Lebart, A. Morineau, M. Piron, "Statistique exploratoire multi-dimensionnelle", Ed. Dunod, 2000.