Correction of the Voice Timbre Distortions on Telephone Network

Gaël Mahé & André Gilloire

France Télécom R&D / DIH / IPS
2, avenue Pierre Marzin, 22307 Lannion cedex, France
gael.mahe@francetelecom.com

Abstract

In a telephone link, the voice timbre is affected by the loss of low frequencies components and distortions due to the analog lines. We analyze first how the quantization noise limits the restoration of the timbre. Within this limitation, a method of equalization, inspired by the cepstral subtraction, is then proposed to correct the timbre and is validated by experimental results.

1. Introduction

In this paper, we define the timbre as the subjectively significant spectral characteristics of the voice. In a telephone link, the voice timbre is affected 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 lines (telephone terminals and line accesses to the local exchange) in the transmission and reception paths, which particularly attenuates the low frequency (LF) components of the voice. For the PSTN, this filtering is described by the *"Intermediate Reference System"* (IRS) [1]. Frequency responses and masks are defined in [1] for the sending and receiving parts of the IRS. We will call the corresponding parts of the network respectively sending system and receiving system.

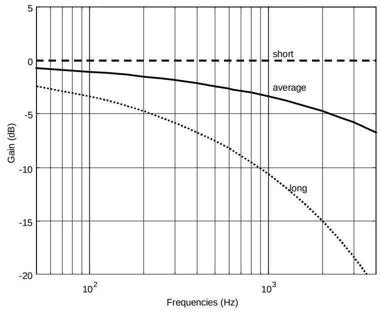

Figure 1: Frequency responses of different customer lines.

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. The frequency responses of different types of lines are shown in Fig. 1. This low-pass filtering muffles the voice in the case of a long line.

In this paper, we present an equalization method to correct these spectral distortions of the voice. The correction is assumed being implemented in the digital part of the PSTN, as shown in Fig. 2, presenting the complete communication chain. We analyze in Section 2 the limits of equalization relatively to quantization noise. Section 3 presents a method to adapt this equalization to different contexts of transmission and experimental results are given to validate our approach.

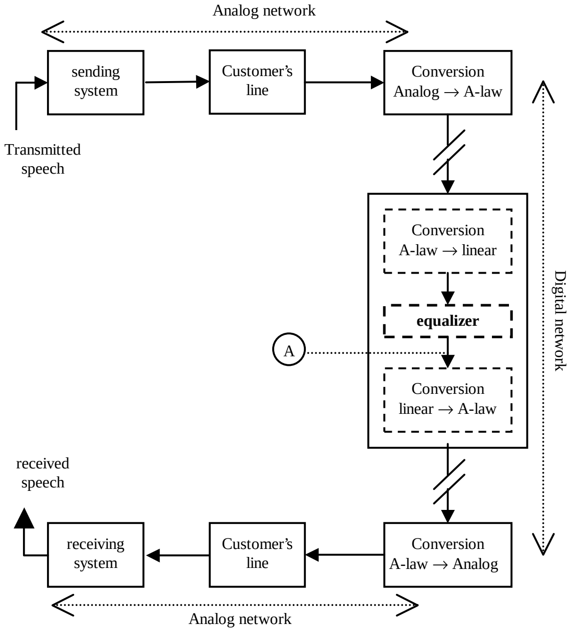

Figure 2: Telephone link with equalizer.

2. Equalization and quantization noise

In this part, we assume the analog part of the link is constituted of the sending and receiving parts of the IRS defined in [1] and average customer analog lines. We consider here a fixed equalizer, which frequency response is the inverse of the total response of the analog channel in the band [F_c - 3150 Hz], where F_c is a chosen lower cut-off frequency.

F_c must be below 300 Hz, in order to restore the LF components of speech. But if F_c is too low, the received signal is affected by a disturbing non-stationary white noise, which can be identified as quantization noise.

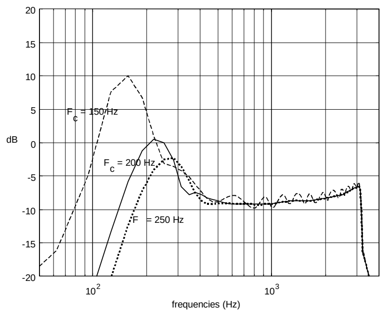

Figure 3: Frequency response of the filtering applied to voice before linear to A-law conversion.

Figure 3 represents, for different values of F_c, the frequency response of the filtering applied to the speech signal between the input of the chain and the linear to A-law conversion (point A in Fig. 2). The low frequency (LF) components are all the more attenuated by the IRS as their frequencies are low. Since the equalizer is placed before the receiving part of the IRS, this anticipated equalization causes differences of level between LF and other components that are all the higher as F_c is low. This is particularly visible on Fig. 3 for F_c = 150 Hz. Because of this level difference, the quantization noise level is close to the level of medium and high frequency (MF and HF) components of speech.

Let us analyze the segmental SNR (Fig. 4) of a sentence filtered by the chain where Fc = 150 Hz. The segmental SNR is the ratio of clean speech energy to noise energy in each 32 ms frame of the received signal, with an overlapping of 50 % between two frames.

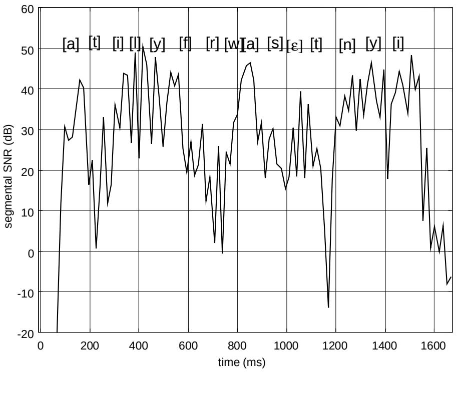

Figure 4: Segmental SNR at the reception.

The SNR is the worst for phonemes with low energy evenly distributed in the band, especially plosive consonants like [t]. In this case, according to Fig. 3, the MF and HF components of the equalized signal are about 20 dB below the LF components. Since the SNR of an A-law quantizer is significantly less than the maximum (38.16 dB) for such low-level phonemes, the level of MF and HF components is close to the noise level. After the LF loss in the receiving system, only these components with the worst SNR remain.

In order to avoid the strong quantization noise induced for some phonemes by the unbalance between LF and the other components, it is necessary to raise F_c, since the gain of the equalizer is all the stronger as the frequency is low. Table 1 represents the effect of the choice of F_c on the segmental SNR for three speakers with various long-term spectra represented in Fig. 5. For each combination (speaker, F_c), the table represents the percentage of frames, in 12 s of speech, for which the SNR is above 30 dB.

F_c Speaker	150 Hz	200 Hz	250 Hz
A	26 %	40 %	42 %
B	31 %	36 %	41 %
C	38 %	48 %	48 %

Table 1: Percentage of frames which SNR is above 30 dB.

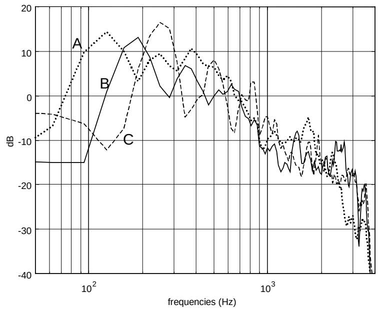

Figure 5: Long-term spectra of 3 speakers.

The raise of SNR when F_c increases is all the more significant as the voice of the speaker is rich in low frequencies: since the voice of speaker C has few components below 250 Hz, the SNR is less affected by the combination of strong amplification below 250 Hz and quantization. The drawback of increasing F_c is that the original timbre of the voice is not so well restored, since the LF are perceptually an important part of it. That is why we must find a trade-off between timbre quality and quantization noise. Choosing F_c = 250 Hz leads to a significant improvement of timbre relatively to the usual telephone quality (without equalization), with an acceptable noise level.

3. Adapted equalization

In the previous part, we assumed that the analog lines are average ones, that the sending and receiving systems comply with the nominal frequency responses defined in [1] and that the equalizer is adapted to them for the band [F_c-3150 Hz]. But if one of the two analog lines is a long line, this equalization is insufficient to avoid the muffling of the voice. On the other hand, the received voice contains too much HF if the line is very short. Other distortions can appear if the sending and receiving systems are too far from the ITU specification. That is why the fixed equalization, that we will call pre-equalization, has to be combined with an equalization adapted to the real analog channel, which will be placed after it.

3.1. Principle

Our method is inspired by the cepstral subtraction [2], used in centralized speech recognition to correct the line effect. In [2], an estimation of the cepstrum of the line is given by the average cepstrum of the signal received by the speech recognizer:

$$\hat{C}_h(\tau) = \mathrm{E}\left[C_r(\tau)\right] \qquad (2)$$

where $\hat{C}_h$ is the estimated cepstrum of the line, and C_r is the cepstrum of the signal *r* received by the speech recognizer. E denotes averaging. In the spectral domain, this result can be translated into:

$$\left|\hat{H}(f)\right|^2 = \mathrm{E}\left[\gamma_r(f)\right] \qquad (3)$$

where $\hat{H}$ is the estimated frequency response of the line and γ_r is the power spectral density (PSD) of *r*.

Since speech recognition systems include a pre-emphasis filter that flattens the long-term spectrum of the voice, $\gamma_r(f)$ must be replaced by $\gamma_r(f)\,/\,\gamma_{\mathrm{ref}}(f)$ in the case no pre-emphasis filter is used, assuming $\gamma_{\mathrm{ref}}(f)$ is the reference average long-term spectrum of the voice. Equation (3) becomes:

$$\left|\hat{H}(f)\right|^2 = \frac{\mathrm{E}\left[\gamma_r(f)\right]}{\gamma_{\mathrm{ref}}(f)}\,. \qquad (4)$$

We can generalize this method to the complete telephone link. Denoting *G* the frequency response of the global analog channel in series with the pre-equalizer and $\hat{G}$ its estimation,

$$\left|\hat{G}(f)\right|^2 = \frac{\mathrm{E}\left[\gamma_y(f)\right]}{\gamma_{\mathrm{ref}}(f)}\,, \qquad (5)$$

where γ_y is the PSD of the received speech *y*.

This result can be derived from the interferences formula:

$$\gamma_y(f) = |G(f)|^2\,\gamma_s(f)\ , \qquad (6)$$

where γ_s is the PSD of the original speech *s*. If the channel is assumed to be invariant,

$$\mathrm{E}[\gamma_y(f)] = |G(f)|^2\ \mathrm{E}[\gamma_s(f)] \qquad (7)$$

Now if we assume the long-term spectrum of the original speech is equal to the known reference $\gamma_{\mathrm{ref}}(f)$, equation (7) leads to equation (5). In other words, the frequency response of the equalizer adapted to the channel is:

$$|EQ(f)| = \sqrt{\frac{\gamma_{\mathrm{ref}}(f)}{\mathrm{E}[\gamma_y(f)]}}\,. \qquad (8)$$

3.2. Application

Since the equalizer is placed inside the network, only the transmission part of the channel can be estimated by this method. $|EQ(f)|$ must be derived from the output of the pre-equalizer, *x*, using:

$$\gamma_y(f) = \left|L\_RX(f)\right|^2\left|S\_RX(f)\right|^2\gamma_x(f)\,, \qquad (9)$$

in the case there is no adapted equalizer, where γ_x is the PSD of *x*, *L_RX* is the frequency response of the reception line and *S_RX* is the frequency response of the receiving system. Since these responses are unknown, we assume the reception line is an average one and the receiving system complies with the ITU specification. We use the corresponding frequency responses L_RX_0 and S_RX_0. The frequency response of the adapted equalizer becomes:

$$|EQ(f)| = \frac{1}{\left|S\_RX_0(f).L\_RX_0(f)\right|}\sqrt{\frac{\gamma_{\mathrm{ref}}(f)}{\mathrm{E}[\gamma_x(f)]}}\,. \qquad (10)$$

We use as long-term spectrum reference $\gamma_{\mathrm{ref}}(f)$ the average spectrum of speech defined by the ITU [3].

3.3. Implementation

The output of the pre-equalizer is analyzed every 32 ms frame, with an overlapping of 50 %. The adapted equalizer is a FIR filter, which coefficients are updated at each frame, according to equation (10), as described below.

According to [2], 2 to 4 seconds of speech are necessary to estimate the channel. So the long-term spectrum of *x*, E[γ_x], is first computed by averaging the PSD over a speech duration increasing from 0 to 4 s. Then it is recursively updated for each frame. Its generic expression is given by:

$$\mathrm{E}\left[\gamma_x(f)\right]_n = \alpha(n)\gamma_x(f,n) + \left(1-\alpha(n)\right)\mathrm{E}\left[\gamma_x(f)\right]_{n-1}\,, \qquad (11)$$

where E[$\gamma_x(f)$]$_n$ is the long-term spectrum at the n^{th} frame, $\gamma_x(f,n)$ is the PSD of the n^{th} frame, and

$$\alpha(n) = \frac{1}{\min(n,N)} \qquad (12)$$

where N is the number of frames in 4 s.

The frequency response of the adapted equalizer is then computed according to (10). As explained in Section 2, only the band 250-3150 Hz must be equalized, so that the values of $|EQ|$ out of this band are replaced by a linear extrapolation of $|EQ|_{[250\text{-}3150\ \mathrm{Hz}]}$, before deriving the impulse response from $|EQ|$ by an IFFT.

3.4. Simulations and results

This method has been simulated for different speakers with the chain considered in Section 2, in which a long line replaces the average transmission line. Figure 6 represents for a given speaker the frequency response of the adapted equalizer after 4 s of speech compared to the ideal response. The difference between these responses leads to a significant distortion of the timbre.

The reason is that the assumption E[$\gamma_s(f)$] = $\gamma_{ref}(f)$ is not fully correct: the global shape of each speaker's long-term spectrum is close to the reference spectrum, but the long-term spectrum itself fluctuates around the reference, even if a longer duration of estimation is used.

That is why $|EQ|$ must be smoothed. Since our goal is to implement the equalizer in the time domain, for low delay and complexity, the simplest way to achieve this smoothing is a narrow windowing of the impulse response.

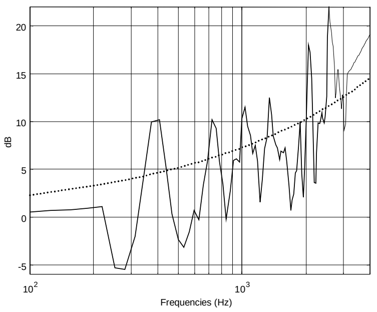

Figure 6: computed (continuous line) and ideal (dotted) frequency response of the adapted equalizer.

Fig. 7 presents, for different speakers, the spectral distortion between the transmitted and the received signals, with pre-equalization and adapted equalization, compared to the distortion with pre-equalization only. For all the speakers we tested, the received voice timbre is significantly closer to the original than without adapted equalization. The improvement is even more significant if we compare it to the timbre without any equalization. Since the principle of the adapted equalization is to bring the long-term spectrum of the voice closer to the reference long-term spectrum, the distortion is the same whichever the lines or the sending and receiving systems are, assuming these systems comply with the masks defined in [1] for the IRS. Consequently, if the pre-equalizer is already adapted to the channel, the adapted equalizer induces a slightly worse distortion for some speakers whose long-term spectrum is too much different from the reference.

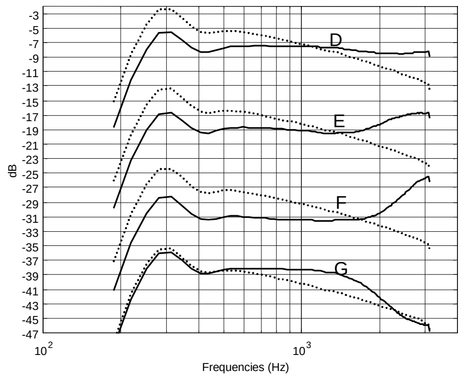

Figure 7: Spectral distortion with (continuous line) and without (dotted line) adapted equalization.

Figure 8 presents, for the four speakers D, E, F and G, the evolution of the cepstral error of the channel estimation over 22 s of active speech, which measures the voice spectral distortion in the equalized band 250-3150 Hz. For each frame, the cepstral error is defined as:

$$e = \sqrt{\sum_{i=1}^{20} (C^i_{\text{eq}} - C^i_{\text{ideal\_eq}})^2} \qquad (13)$$

where C^i_{eq} and $C^i_{\text{ideal\_eq}}$ are the i^{th} cepstral coefficients of the adapted equalizer and the ideal adapted equalizer, respectively. The first cepstral coefficients C^0_{eq} and $C^0_{\text{ideal\_eq}}$ are not taken into account in the computation of *e*, in order that the error does not reflect any possible difference of energy. The threshold value of audible distortion is about 0.2. For nearly all the tested speakers, the error with adapted equalization is less than the error without adapted equalization after 1 s of speech. For most of them, like D and E, the distortion is hardly audible after 2 s, and a minimum value is reached within 4 s.

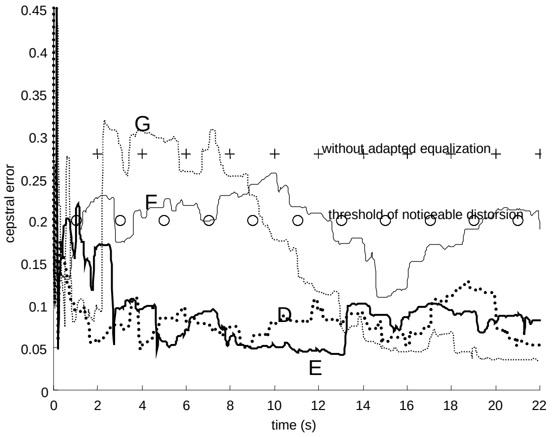

Figure 8: Cepstral error during the first 22 s of speech.

3.5. Further discussion

One could wonder why we choose a two parts structure for the equalizer (pre-equalizer and adapted equalization) instead of a single completely adapted equalizer. The main reason relies on the necessity of smoothing the frequency response of the adapted equalizer. Because of this strong smoothing, the frequency response of the adapted equalizer cannot have the steep slope of the pre-equalizer (typically an IIR filter) in the low frequencies.

4. Conclusions

In the method for correction of the timbre we have presented, the correction is limited to the band 250-3150Hz, since the restoration of the components below 250 Hz would raise the quantization noise in a disturbing manner. Within these limits, the correction is achieved by the combination of a pre-equalizer, which compensates for the average analog channel, and an equalizer that adapts to various conditions of transmission. This adapted equalization, based on the comparison between the long-term spectrum of the received signal and the reference long-term spectrum of the speech, provides a significant improvement of the received voice, which timbre is close to the original in the band 250-3150 Hz.

5. References

[1] ITU, Recommendation P.48, "*Specification for an intermediate reference system,*" 1988.

[2] C. Mokbel, J. Monné and D. Jouvet, "On-line adaptation of a speech recognizer to variations in telephone line conditions," in *Proc. Eurospeech*, pp 1247-1250, sept. 1993.

[3] ITU, Recommendation P.50, "*Artificial voices,*" 1993.