THE SPEECH CONFORMER FOR ACHIEVING GOOD SPEECH INTELLIGIBILITY FOR ALL SPEAKERS

Active Audio
Saint Herblain, France
Xavier Meynial

1 INTRODUCTION

There are now on the market many loudspeaker systems that enable achieving high STI values even in adverse situations caused by high reverberation and/or high background noise. Audio signals sent to the loudspeakers often go through audio processing, with compressors, gains, delays, equalizers and so on.

However, the quality of the audio signal perceived by the listeners also depends of the source, i.e. the voice of the speaker. We have all experienced that: in a given room, a message spoken by speaker A is clear and intelligible, while the same message spoken by speaker B is unclear and poorly intelligible.

For instance, speaker A would be a woman, having a timbre of voice rich in high-mids (1.5 - 6kHz), speaking in the microphone with fairly constant sound level, clear articulation of the phonemes and low phoneme flow rate. And speaker B would be a man with a voice timbre rich in low frequencies, large sound level variations, poor articulation of phonemes and high phoneme flow rate.

Variations of the speech level (for example due to movements of the speaker relative to the microphone) can easily be corrected using a compressor or some kind of automatic gain control (AGC) device. The question of the phoneme flow rate cannot be addressed with a real-time algorithm. The question of the articulation is rather complex and remains to be addressed.

In this paper, we focus on the problem of the voice timbre: we present an adaptive real-time algorithm that corrects the timbre of voices in order to achieve a given target spectrum.

The idea of a dynamic EQ for voice processing is not new. In 1969, Allen *et al.* published a paper [1] describing an analog processor which boosts the “presence band” when necessary. More recently, several manufacturers offered digital processors for automatic EQ of voices, but we haven’t found publications on the principles and algorithms involved. The algorithm we wish to discuss in the present paper aims at achieving faster and more robust adaptation.

2 DESCRIPTION OF THE ALGORITHM

Figure 1 shows the third-octave spectra of male and female voices as specified in standard CEI 60268-16 [2]. The curves have a bell shape with a max at low frequencies. Figure 1 also shows an example of poorly intelligible male voice with a timbre balance toward low frequency level, and a highly intelligible female voice with timbre balance toward mid-high frequencies.

Figure 2 presents the general block diagram of the algorithm we propose in this paper. The general principle is simple:

- the input signal goes into a 4-band filter bank,
- the level in each band is estimated and compared to a target spectrum chosen by the user,
- gains in each band are computed so that the current spectrum matches the target spectrum,
- gains are applied to the 4 bands, and the 4 signals are summed, yielding the output signal.

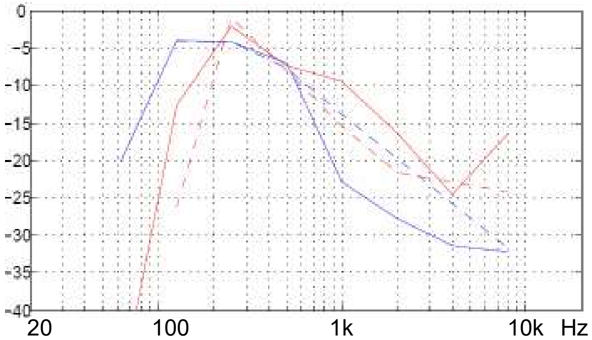

Figure 1: Octave spectra of female (dashed red) and male (dashed blue) voices according to standard CEI 60268-16, and example of a highly intelligible female voice (plain red), and a highly unintelligible male voice (plain blue). All curves in dB referenced to their wide band level.

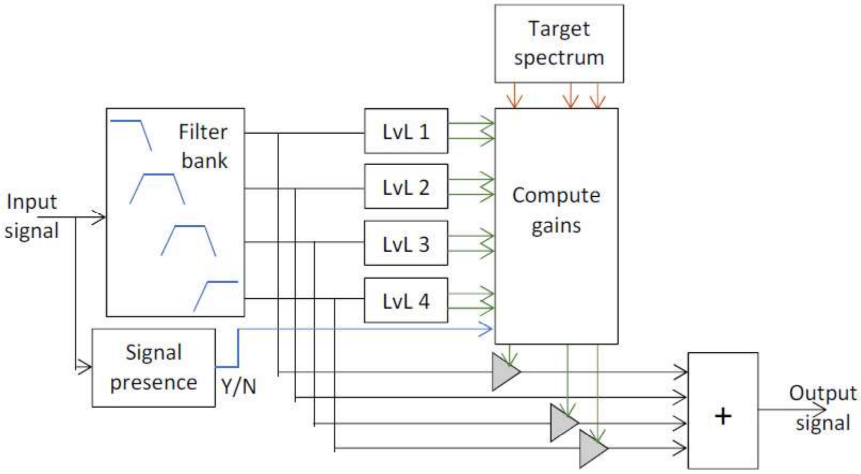

Figure 2: General block diagram of the Conformer algorithm.
Black arrows : signals - green arrows : levels and gains - orange arrow : input parameters.

The 4 bands of the filter bank are:

- Band 1, called “warmth” : 4th order high-pass at 200Hz.
- Band 2, called “reference” : 4th order band-pass at 200-1500Hz.
- Band 3, called “presence” : 4th order band-pass at 1500-6000Hz.
- Band 4, called “brightness” : 4th order high-pass at 6000Hz.

All filters are of type Linkwitz-Riley for achieving perfect subsequent reconstruction.

The target spectrum consists of 3 values which are the desired levels in band 1, 3, 4 relative to the level in band 2 (the reference band).

In each frequency band, the level is estimated using 2 time constants: 1.4 and 9.2 sec. From these 8 levels, and considering the 3 values of the target spectrum, gains G_1, G_3 and G_4 for bands 1, 3, and 4 respectively are computed. No gain is applied to band 2 (G_2=1), which is the reference band.

After gains are applied, the broadband signal is reconstructed by summing the 4 signals, yielding the output signal. As can be seen on figure 1, levels in band 2 are usually much higher than other bands. Consequently, applying gains in bands 1, 3 and 4 does not change the total signal level significantly: output signal level is close to input signal level.

The input signal is also applied to a signal presence detection block, which detects whether a signal is applied to the audio input based on a threshold scheme (a hold of 1 sec is used in order to allow for speaker breathing pauses). When the presence flag is low, the computation of the gains G_n is paused.

The key issue is how fast the algorithm should adapt to a new voice: fast variations of levels due to the phonetic content of speech should not induce large changes in the gains, but an abrupt change of levels due to a change of speaker should induce a quick adaptation of gains, so the algorithm quickly adapts to a new voice. When an abrupt change in the spectrum is detected (corresponding to a change of speaker), the algorithms allow gains G_1, G_2, G_3 to adapt faster than otherwise. We have found that a minimum of 5 seconds of speech is needed to obtain a correct estimate of the levels in each band, this duration corresponding to approximately 10-20 phonemes. Therefore, 5 sec the time the algorithm typically needs to adapt to a new voice.

One of the problems the user might encounter in case the Conformer is used for live events is the risk of howling. Indeed, as the algorithm can amplify certain bands (G_i>1) depending on the input spectrum, a howling effect can occur. Therefore, a test button has been added which sets the 3 gains G_1, G_3, G_4 at the maximum allowed values G_{1max}, G_{3max}, G_{4max} specified by the user. If howling occurs, the user must reduce G_{1max} and/or G_{3max} and/or G_{4max}.

Upstream this algorithm is a compressor (or AGC) which ensures the level of the voice is normalized. Also, a pre-emphasis / de-emphasis scheme (not shown on figure 2) is used to flatten the input spectrum, so that the band filtering is more efficient (less spill-over).

The speech Conformer algorithm being a feedforward algorithm, there is no risk of instability. It uses very little space in RAM (a few hundred words), and moderate CPU (approx. 1000 MIPS).

3 RESULTS

Figure 3 shows the evolution of gains G_1, G_2 and G_3 when the input signal switches from a clear female voice to a dull male voice. In this example, the target spectrum corresponds to the female spectrum of CEI 60268-16 [2]. It is clear that shortly after the change of speaker, the gains jump by steps of 3dB (for quick adaptation) and then evolves slower as the current spectrum approaches the target spectrum.

Figure 4 shows the octave spectra of the input and output signals (same as for fig 3) after the Conformer gains are stabilized (after t=10s on fig 3). One can see that the spectrum correction corresponds to the gains G_1, G_2, G_3 of figure 3.

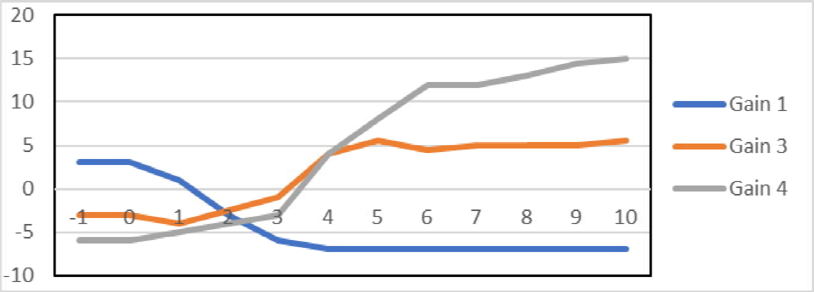

Figure 3: Evolution of gains G_1, G_2, G_3 (in dB vs. sec) when the signal switches (at t=0) from a clear female voice to a male dull voice. Target spectrum: female of CEI 60268-16.

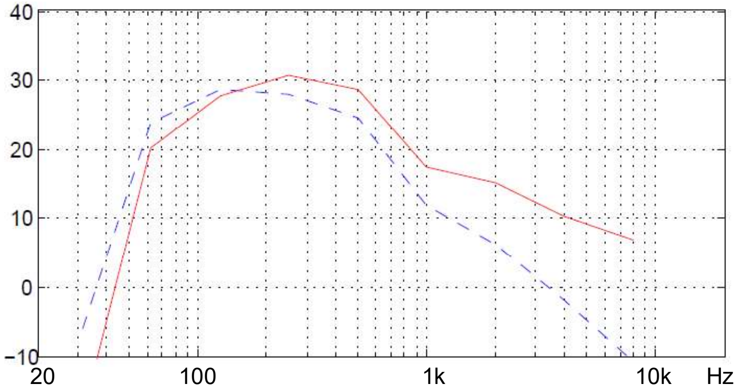

Figure 4: Octave spectra (in dB with arbitrary 0dB reference) of the male voice of fig 3 without (blue dashed) and with (red plain) conformer correction.

The perceptive effect of the Conformer algorithm is striking: when the voice spectrum is close to the target spectrum then almost no correction is applied (G_1=G_3=G_4=1), but when the voice spectrum is far from the target spectrum then large gains values of G_1,G_3,G_4 change the timbre dramatically.

When the target spectrum is high in the “presence” band (1.5-6kHz) and low in the “warmth” band (<200Hz), the algorithm increases noticeably the speech intelligibility, especially if the input speech is poor in the “presence” band and if the acoustics of the room is adverse (high reverberation and/or high background noise).

4 CONCLUSION

We have presented an algorithm which automatically corrects the timbre of voices, based on a given target spectrum. It is an adaptive algorithm which only takes about 5 seconds to adapt to a new voice. It is safe, and requires only little RAM and CPU.

The speech conformer might for example be used:
- in public transports to increase speech intelligibility of vocal announcements,
- in broadcast to give voices more warmth,
- in classrooms to make the teacher more understandable by pupils, …

5 REFERENCES

1. A dynamic presence equalizer, Richard G. Allen; Emil L. Torick, and Benjamin B. Bauer, 37th AES Convention, October 1969, preprint 702.
2. Standard CEI 60268-16, Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index