Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99),NTNU, Trondheim, December 9-11, 1999

A wavelet based method for audio-video synchronization in broadcasting applications

G. Pallone, P. Boussard[*]*, L. Daudet, P. Guillemain, R. Kronland-Martinet*

Laboratoire de Mécanique et d'Acoustique, CNRS
31 chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France

[*]GENESIS, c/o CEEI Provence
Domaine du Petit Arbois BP88
13545 Aix en Provence CEDEX 4, France

gregory@mond.net

ABSTRACT

The difference between standards used for films and for video generates problems when a conversion from one format to another is required : Since all the images are displayed, the change of frame rate induces a pitch change on the sound. To avoid this problem, the whole soundtrack has to be processed during the duplication. In this paper, we address the corresponding sound transformation problem, namely the dilation of the sound spectrum without changing its duration. For broadcasting applications, the ratio of transposition is within the range 24/25-25/24. The wide variety of sounds (music, speech, noise…) used in movies led us to first construct a database of representative sounds containing both transient, noisy and quasi-periodic sounds. This database has been used to compare the performances of different approaches. The reviewing of the most well known methods clearly shows significant disparities between them according to the class of the signal. This led us to reconsider the problem and to propose methods based on wavelet transforms.

1. INTRODUCTION

Due to the coexistence of different standards in film (twenty four frames per second) and video (twenty five frames per second), conversions between these formats are often necessary [1]. Generally speaking, since the number of images per unit of time is different, the duration of the whole sequence varies, according to the type of format used to play the same tape. This causes a dramatic change in sound due to its dilation in both time and frequency domains, just like if you play a 45 RPM at the speed of a 33 RPM.

That is why the sound engineers have to compensate the sound modification due to the slowing down of the film thanks to an appropriate frequency stretching.

Thus, we have to face a scientific problem of frequency stretching also called pitch shifting. The aim is to preserve the quality of the whole soundtrack according to broadcasting standards.

For our purpose, we shall consider that the frequency stretching ratio is above the unity. In other words, the pitch change is a transposition towards higher frequencies, but the opposite is a problem of post-production too.

Up to now, in post-production studios, this sound transformation is calculated by a professional machine, Lexicon 2400. The transcription can be performed simultaneously only on two stereo channels and the technology used is more than fifteen years old. However, the generalization of the digital multichannel sound leads professionals to consider other machines, using new technologies and based on scientific research in digital signal processing in the field of pitch shifting algorithms.

Our work started by recollecting the available software and systems and testing them on a sound database. Auditory tests lead us to conclude that the Lexicon 2400 does not give perfect results but remains the most suitable machine for broadcasting applications.

2. BANK OF SOUNDS

To estimate the quality of existing and developed algorithms, we constituted a bank of sounds. Our aim is to preserve the quality of the elements of the whole soundtrack and we have collected a bank of sounds using the widest variety of sounds. From the cinematographic point of view, we will talk about speech, music, sound effects and surroundings. We have classified these sounds from a signal processing point of view. We shall then talk about quasi-periodic, transient, noise and inharmonic signals. Our database has been made from :

- Female and male voices, characteristic of quasi-periodic and noisy signals.

- "Cocktail party", a sequence with voices and music recorded simultaneously with a dummy head to evaluate the ability of the algorithm to preserve the phase relations, essential to the accuracy of the stereo sound image.
- Piano theme and single note, lightly inharmonic.
- Castanets, characteristic of transient signal.
- Accordion, typically harmonic with wideband spectrum.

These last three sounds has been chosen because they gave bad results with the Lexicon.

3. TIME/FREQUENCY DUALITY

A simple mathematical demonstration can show that time/frequency duality has a significant property. Let's consider a time-signal s(t), and its Fourier Transform S(ω). A pitch shifting of a ratio α (i.e. all frequency are multiplied by this factor) inevitably implies a temporal contraction of a factor 1/α :

$$s'(t) = \int S(\alpha\omega) e^{j\omega t} d\omega = \frac{1}{\alpha} s\left(\frac{t}{\alpha}\right) \qquad (1)$$

That's what occurs when the film is projected at a different speed.

The time/frequency duality shows that no mathematically perfect transformation can be performed. Thus we will be induced to use compromises and many tricks. We will not forget that ear is the ultimate judge of the quality of the algorithms and we can take advantage of this property.

Actually, a pitch shifting is equivalent to a time stretching without change of pitch, followed by a downsampling. So, using this resampling operation, it is equivalent to consider a pitch shifting or a time stretching problem [2]. Moreover the majority of the studied methods calls on the time stretching.

4. CLASSES OF METHODS

We can classify all the existing algorithms into two families :

- Methods using the time-domain description, which manipulate short-duration time-segments extracted from the original signal, usually called splicing methods.
- Methods using the frequency-domain description of the signal, such as the Short Time Fourier Transform.

We shall not describe here the splicing methods although they give good results on a large variety of sounds. Lexicon 2400 uses one of this method. Readers interested in these kind of methods can refer to [2]-[5].

Frequency methods operates in three steps :

An analysis process : The original signal is decomposed by linear filtering into a given number of frequential sub-bands. Parameters are extracted from each sub-band.

A transformation process : A modification of the parameters resulting from each sub-band is performed.

A synthesis process, inverse of the analysis process : It recomputes the output signal from the transformed parameters.

The synoptic is shown in figure 1.

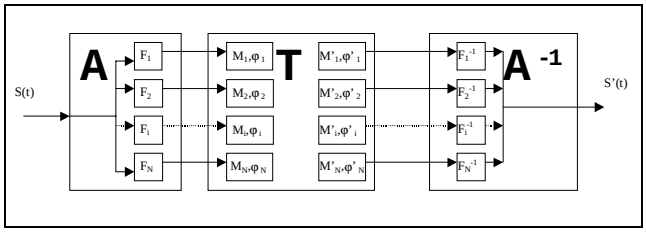

Figure 1. *Detailed synoptic of the analysis / transformations / synthesis process.*

These methods are chosen to give exactly the original signal when no transformation is performed: They are invertible.

We shall now describe briefly a well known example of this type of method : The classical Phase Vocoder. The analysis is performed using a fixed bandwidth and the transformation modifies the phases of each sub-band.

5. CLASSICAL PHASE VOCODER

The principle of this technique is used, for example, in " Super Vocodeur de Phase " of the IRCAM in Paris, and in the phase vocoder of Daniel Arfib. It has been described in a lot of papers [6]-[12].

A " local " spectrum is first calculated at different moments separated by an analysis constant step, and then, the output signal is resynthesized with a step according to the time-modification factor (which is the same as the pitch-modification factor).

Figure 2 shows an example of the slowing down of the spectrum.

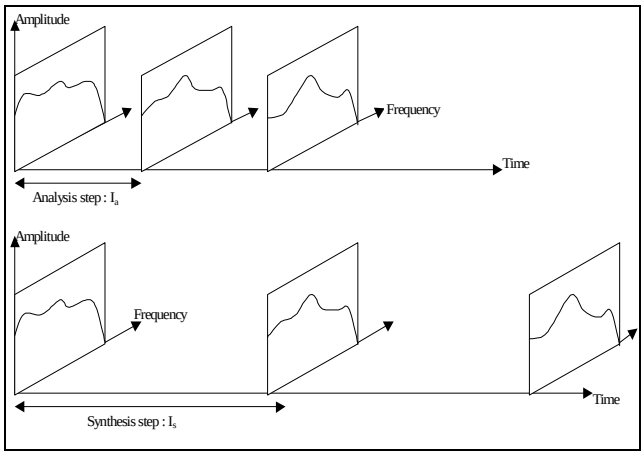

Figure 2. *Phase Vocoder analysis/synthesis principle.*

The resulting signal is longer than the original one, but contains the same frequencies. We just have to resample this signal to obtain the desired transposed signal.

This method gives good results as long as the following assumption made on the input signal is respected : It must be a linear combination of a fixed number of sinusoids, slowly modulated in amplitude and in frequency.

The timbre of the modified sounds can be altered when the number of distinct sinusoids in the original signal becomes too significant, which generally implies more than one sinusoid in each sub-band. There is also a transient smearing, because the representation in a sum of sinusoids is not well adapted to transients.

The four main parameters that can be adjusted for a given time scaling ratio are : the FFT order, the analysis stride, the type and the length of the analysis window. The last parameter is directly linked to the time resolution of the analysis. The time/frequency duality states that it is impossible to have precision both in time and frequency. So, for a given set of parameters, the algorithm cannot be efficient for both orchestra sounds and castanets.

Although such algorithms can be implemented very efficiently [13], this kind of analysis with constant bandwidth does not well reflect how ear achieves the analysis of sounds. Starting from the phase vocoder, we were aware of its underlying limits in both time and frequency, due to the analysis window (too wide for the transients and too narrow for a good frequency resolution). So we paid particular attention to realize acceptable compromises.

6. ANALYSIS-TRANSFORMATION-RESYNTHESIS

We could use elementary waveforms [14] as it is done in the Matching Pursuit algorithm but it can't be done in real-time. We chose an analysis which is no more computed with a constant bandwidth as it is the case in the classical phase vocoders, but based on the Bark scale [15]: the filter bandwidth is constant from 20 Hz until approximately 500 Hz (similar to the Short-Time Fourier Transform [11]), and then its bandwidth is proportional to the frequency (similar to the wavelet analysis [16]). This kind of analysis is closer to the ear functioning [17].

Figure 3 shows the characteristics of the filter bank.

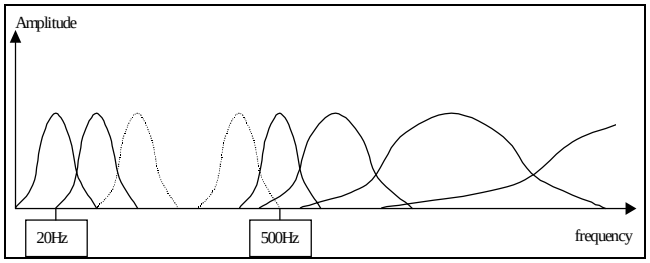

Figure 3 *Filter bank characteristics*

The original signal is filtered through a given number of sub-bands using gaussian filters. This number is chosen according to the bandwidth of the filters to insure a perceptively perfect reconstruction when no transformation is performed. The bandwidth of the analysis below 500 Hz is about 20 Hz and the bandwidth of the analysis above 500 Hz is about 1/10 octave. On one hand, the good frequency resolution of the phase vocoder is preserved below 500 Hz thanks to large time windows (to the detriment of a good time resolution at these frequencies). On the other hand, the time resolution is better preserved because the time windows are much narrower at high frequencies (to the detriment of a good frequency resolution at these frequencies). Such a feature is important for a good representation of transient signals. However, the information about transients does not seem to be preponderant at low frequencies, and ear is not really selective at high frequencies. That's why we choose this kind of analysis.

In this way, artifacts observed with the phase vocoder either on quasi-periodic or on transients using a unique set of parameters are reduced, as it has been observed through psychoacoustic tests.

The use of analytic functions at the step of decomposition directly gives an analytic signal in each sub-band from which the modulus and the phase are easily extracted.

Using the notations in figure 1, we note the output of each sub-band.

$$s_i(t) = M_i(t) e^{j\varphi_i(t)} \tag{2}$$

The pitch shifting is obtained by letting identically the modulus and by multiplying the phase of each sub-band by the ratio of transposition.

$$M_i'(t) = M_i(t) \tag{3}$$

$$\varphi_i'(t) = \alpha\varphi_i(t) \tag{4}$$

The transposed signal is obtained by summing the real parts of the output of each sub-band.

$$s'(t) = \sum_i M_i'(t)\cos\left(\varphi_i'(t)\right) = \sum_i M_i(t)\cos\left(\alpha\varphi_i(t)\right) \tag{5}$$

7. CONCLUSION

In this paper, we have addressed a sound transformation problem, namely the dilation of the sound spectrum without changing its duration. For broadcasting applications, the ratio of transposition is in the range: 24/25-25/24. The wide variety of sounds (music, speech, noise…) used in the movies has led us to first construct a database of representative sounds containing both transient, noisy and quasi-periodic sounds. This database has been used to compare the performances of different approaches. The reviewing of the most well known methods clearly shows significant disparities between them according to the class of the signal. This led us to reconsider the problem and propose a method based on both time-frequency and time-scale representations. The results obtained with this method are rather encouraging : The general quality of the transformations of the sounds from our database is better than the one obtained with the classical phase vocoder. Nevertheless, the quality is not yet superior to recent time domain algorithms in the range of +/- 4% [5]. We still have to work to optimize the parameters and find a

less crude way to achieve the pitch shifting within each sub-band.

Since the transformation is made independently of the signal, the algorithm can be used on more than one channel, without generating any shifting between them. Moreover, this algorithm can be used to do dynamic variations of speed without altering the timbre, and could be used for example to synchronize the sound to the movements of the lips, thanks to an appropriate pitch shifting curve.

8. REFERENCES

[1] Amyes, "Audio Post-production in Video and Film", Focal Press, 1998.

[2] J. Laroche, "Time and Pitch scale modifications of audio signals", *Applications of DSP to Audio and Acoustics*, Kluwer Academic Publishers 1998.

[3] J. Laroche, "Autocorrelation method for high quality time/pitch scaling", *Proc. IEEE Workshop Appl. Of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Platz, NY, 1993.

[4] G. Fairbanks, W. Everitt and R. Jaeger, " Method for time or frequency compression-expansion of speech ", IEEE Trans. Audio and Electroacoustics, AU-2 : pp. 7-12., 1954.

[5] G. Pallone, "Transposition fréquentielle pour des applications de post-production audiovisuelle", Mémoire de DEA ATIAM, 1999.

[6] J.A. Moorer "The Use of the Phase Vocoder in Computer Music Applications", Journal of the Audio Engineering Society, Jan/Feb 1978, vol 26, n°1/2, 1978.

[7] M.R. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis ", IEEE Transactions on ASSP, vol.ASSP-28(1): pp. 55-69, Feb 1980.

[8] R.E. Crochiere, "A Weighted Overlap-Add Method of Fourier Analysis-Synthesis", IEEE Transactions on ASSP, vol.ASSP-28(1): pp. 99-102, Feb 1980.

[9] M.R. Portnoff, "Short-Time Fourier Analysis of Sampled Speech", IEEE Transactions on ASSP, vol.ASSP-29(3): pp. 364-373, June 1981.

[10] M.R. Portnoff, "Time-Scale modification of Speech Based on Short-Time Fourier Analysis", IEEE Transactions on ASSP, vol.ASSP-29(3): pp. 374-390, June 1981.

[11] D. W. Griffin, J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", IEEE Transactions on ASSP, vol.ASSP-32(2): pp. 236-243, April 1984.

[12] M. B. Dolson, "The Phase Vocoder: A Tutorial", *Computer Music J.*, vol.10(4), pp. 14-27, Win 1986.

[13] X. Rodet and P. Depalle, "Spectral Envelopes and Inverse FFT Synthesis", AES Preprint n°3393 (H-3), 1992 October.

[14] X. Rodet, "Musical Sound Signal Analysis/Synthesis :Sinusoidal + Residual and Elementary Waveform Models", TFTS'97, IEEE Time-Frequency and Time-scale Workshop 97, Coventry, GB, Aug 1997. http://mediatheque.ircam.fr/articles/textes/Rodet97/

[15] E. Zwicker, H. Fastl, "Psychoacoustics Facts and Models", Springer-Verlag Berlin Heidelberg 1990.

[16] R. Kronland-Martinet, "The Wavelet Transform for Analysis, Synthesis, and Processing of Speech and Music Sound", Computer Music Journal, vol 12 :4 1988, pp. 11-20.

[17] C. Roads, " L'audionumérique " traduit par Jean de Reydellet, Traduction française de " The Computer Music Tutorial " publié aux Etats-Unis par MIT Press, Dunod 1998.