TRANSFORMING SINGING VOICE EXPRESSION – THE SWEETNESS EFFECT

Lars Fabig, Jordi Janer

Music Technology Group – IUA
Universitat Pompeu Fabra, Barcelona, Spain
{lfabig,jjaner}@iua.upf.es

ABSTRACT

We propose a real-time system which is targeted to music production in the context of vocal recordings. The aim is to transform the singer's voice characteristics in order to achieve a sweet sounding voice. It combines three different transformations namely Sub-Harmonic Component Reduction (reduction of sub-harmonics, which are found in voices with vocal disorders), Vocal Tract Excitation Modification (to achieve a change in loudness) and the Intonation Modification (to achieve smoother transitions in pitch). The transformations are done in the frequency domain based on an enhanced phase-locked vocoder. The Expression Adaptive Control estimates the amount of present vocal disorder in the singer's voice. This estimate automatically controls the amount of Sub-Harmonic Component reduction to assure a natural sounding transformation.

1. INTRODUCTION

1.1. Voice characters

A singer's voice may have characteristics which on the one hand are described scientifically as vocal disorders, like a growl, creak, rough or hoarse voice. On the other hand, from the musical point of view, one can consider these characteristics as highly relevant expressiveness features, which creates its singer's voice unique timbre. There are singers whose voices always have a timbre with vocal disorders, like Louis Amstrong, Janis Joplin and Tom Waits, but others control the amount of vocal disorders intentionally like Joe Cocker, Sting and Brian Adams.
Obviously intentional vocal disorder is not the only possible expression in singing voice. Among many others there are vibrato, intonation, breathiness and loudness.

In this article we concentrate on Sweetness in the singing voice. The idea behind is to transform whatever type of singing voice into a sweet sounding voice, for example a sweet jazz voice, like the jazz singer Silje Nergaard or a smooth Bossa Nova voice like the singer Astrud Gilberto. Instead of adding expression we want to flatten it. This can be of interest in music production and post-production to be able to easily adjust the expressivity of a recorded singer's voice posteriorly. A positive side effect is that we can avoid additional recording sessions, if we just want to retouch just a few phrases. Or a singer just wants to obtain a certain timbre that he is not able to sing or because of a steady hoarse voice.

To achieve the Sweetness effect we propose a combination of three different transformations:

- *Sub-Harmonic Reduction* intends to minimize the amount of vocal disorders.
- *Intonation Modification* smoothes pitch changes to achieve a sort of portamento effect.
- *Vocal Tract Excitation Modification* modifies the loudness of the voice.

1.2. Spectral View of Vocal Disorders

Voice disorders are caused by a variety of factors. Abuse or misuse of the voice with yelling, singing, or speaking is one common reason. Physical alterations in the vocal folds due to abusive lifestyle (smoking, alcohol) or aging are possible. Lack of movement and poor or improper function of the vocal folds are other causes.
But also we are able to intentionally control our voice the way that vocal disorders become audible. Singers exploit it to enhance their timbre with personal expression. It may be used as an effect in a particular passage or phrase, e.g.: in a loud extended sung tone.

In this article we focus on the particular case of voice disorder that provokes "hoarseness". This pathology is often referred in medical literature as Muscle Tension Dysphonia. This phenomenon occurs in the Larynx, the organ that includes the vocal folds (the source of voice production) due to an excessive muscular effort and usually due to pressed phonation (high sub-glottal pressure) [5].

A spectral analysis of a hoarse or growl voices reveals additional sub-harmonics which are superposed with the main harmonic partials. They appear due to modulations in amplitude and/or frequency of the glottal cycle periodicity in the voice source [1]. In one particular case depicted in figure 1 there is not only one harmonic series present but several commensurable harmonic series, which are superposed. You see a harmonic series (a) which has been superposed with a sub-harmonic series in (b). The fundamental frequency is of the sub-harmonic series is approximately one third of f_0. In other cases when the irregularities of the vocal folds vibrations behave more stochastically, there may appear many noisy components in the spectrum, which can not be interpreted as a commensurable harmonic series anymore.

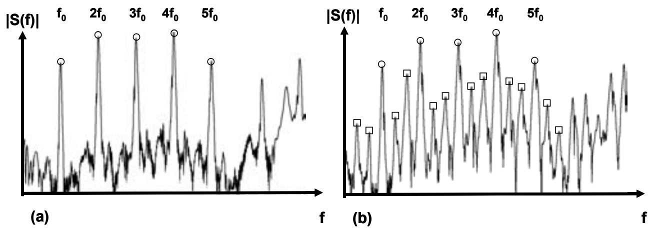

Figure 1: Spectra of singing voice without significant sub-harmonics (a) and with sub-harmonics (b). Sub-harmonics are indicated with square boxes.

2. SWEETNESS ALGORITHM OVERVIEW

The proposed Sweetness Effect algorithm is represented by the block diagram seen in figure 2.

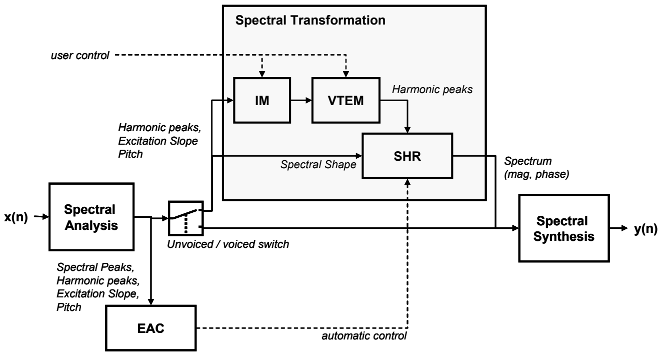

Figure 2: Sweetness Effect processing scheme.

The applied processing technique is an enhanced version of the rigid phase-locked vocoder based on Short-Time-Fourier-Transformation (STFT) using an Analysis - Transformation - Synthesis scheme. The main difference is a Fundamental Frequency Estimator which allows us to classify spectral peaks as harmonic or spurious peaks (see section 3.2 for details). First a frame-based-*Spectral Analysis* is done, which extracts the following spectral data:

- *Spectral peaks* (frequency, magnitude and phase)
- *Fundamental frequency*
- *Harmonic peaks* (frequency, magnitude and phase)
- The approximated *Spectral Shape*
- *Vocal Tract Excitation Descriptors*.

The estimated spectral data is used as control data and/or as data to be transformed in the Spectral Transformation block. Furthermore they are used as input data for the Expression Adaptive Control, which derives a control signal to regulate the necessary amount of Sub-harmonic Component Reduction depending on the amount of vocal disorder in the singer's voice.

The Transformation block combines three types of transformation which are namely the *Intonation Modification* (IM) (stretches or smoothes pitch changes), *Vocal Tract Excitation Modification* (VTEM) and the and *Sub-Harmonic Reduction* (SHR). While the amount of Intonation Modification and the amount of Vocal Tract Excitation Modification is controlled manually by the user, the amount of Sub-Harmonic Reduction is controlled automatically by the *Expression Adaptive Control* (EAC). The whole transformation section is bypassed when the analysis spectral frame is considered to be unvoiced. The reason is that roughness and growl is only perceived in the stationary part of sung vowels. The spectral processing techniques used by the three transformations are transposition, equalization and spectral interpolation.

The transformed spectral data is then re-synthesized in the *Spectral Synthesis* block, which consist in an Inverse-Fast-Fourier-Transformation, inverse windowing and Overlap & Add procedure.

The algorithm has been implemented as a real-time application based on the Spectral Peak Processing technique [2], which was especially developed for transforming singing voice with high sound quality, using the C++ Library for Audio and Music (CLAM) [7].

In section 3 we are going to give an overview about used techniques for spectral analysis and in section 4 we explain the spectral transformations we apply to achieve the Sweetness effect. Section 5 deals with the Expression Adaptive Control and we conclude with results and improvements in section 6.

3. SPECTRAL ANALYSIS

3.1. Peak and Pitch Detection

The Spectral Analysis is done using Short-Time-Fourier-Transformation. The time-signal is multiplied with the function of a Kaiser-Bessel window of 46 ms length and then a Fast-Fourier-Transformation is calculated. A simple peak detection algorithm is used to detect the local maxima in the magnitude spectrum (Spectral Peaks). The Fundamental Frequency Detector estimates the singer's pitch from the detected *Spectral Peaks*. The Fundamental Frequency Estimator furthermore includes a decision algorithm to label a frame as voiced or unvoiced.

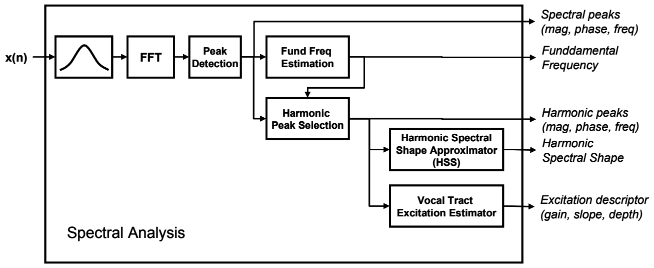

Figure 3: Spectral Analysis block diagram.

3.2. Harmonic Peak Selection

Considering the estimated fundamental frequency the spectrum is divided in regions of perfect harmonics according to multiples of the fundamental frequency f_0. For each of the regions the

Spectral Peak of maximum magnitude is searched which is supposed to be the representative peak for this region, called Harmonic Peak [3]. It is represented as well as the Spectral Peaks as a data triplet of frequency, magnitude and phase.

3.3. Harmonic Spectral Shape Approximation

We use a 3rd order spline-interpolation to approximate the Harmonic Spectral Shape, which interpolates between data pairs of logarithmic magnitude and frequency of the selected harmonic peaks.

3.4. Vocal Tract Excitation Estimation

The Vocal Tract Excitation Estimation takes advantage of the Excitation plus Resonance Model [2] (EpR), which is based on an extension of the well known source/filter approach [4]. The EpR filter can be decomposed into two cascade filters. The first of them models the differentiated glottal pulse frequency response and the second the vocal tract (resonance filter). The EpR source is modeled as a frequency domain curve and one source resonance, see figure 4. The curve is defined by a gain and an exponential decay as follows:

$$Source_{dB} = Gain_{dB} + SlopeDepth_{dB}\left(e^{Slope \cdot f} - 1\right) \qquad (1)$$

It is obtained from an approximation to the harmonic spectral shape (HSS) determined by the harmonics identified in the Harmonic Peak Selection:

$$HSS(f) = envelope_{i=0..n-1}\left[f_i,\ 20\log(a_i)\right] \qquad (2)$$

where i is the index of the harmonic, n is the number of harmonics, f_i and a_i are the frequency and amplitude of the i_{th} harmonic. On top of the curve, we add a second resonance in order to model the low frequency content of the spectrum below the first formant. The vocal tract resonance model has no impact on hoarseness or growl so that we do not consider it for our needs.

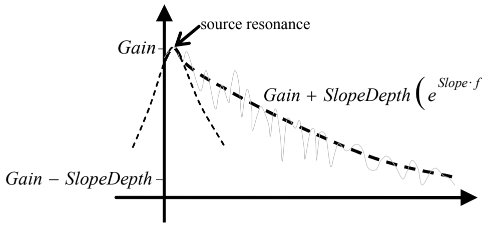

Figure 4: Approximation of the EpR source from [2].

4. SPECTRAL TRANSFORMATIONS

4.1. Sub-Harmonic Reduction

The Sub-Harmonic Reduction reduces the amount of perceived hoarseness or growl minimizing the energy of sub-harmonics. Assuming that a pure harmonic series without sub-harmonics represents a clean sounding voice without vocal disorders, the approach is to synthesize the voice from the pure harmonic series, i.e. synthesizing just the harmonic peaks.
To model the bandwidth of the harmonics, the Harmonic Peaks are convolved with the Fourier-transform of the analysis window function. This is done in the *Sinusoidal Renderer* which fills the surrounding areas of each harmonic.
In contrast, the *Spectrum Renderer* fills the spectrum with the data taken from the original spectrum (maintains the original shape of the magnitude spectrum).
Afterwards both complex spectra are interpolated according to the interpolation factor *k* which is controlled by the user to adjust the amount sub-harmonic component reduction, see figure 5.

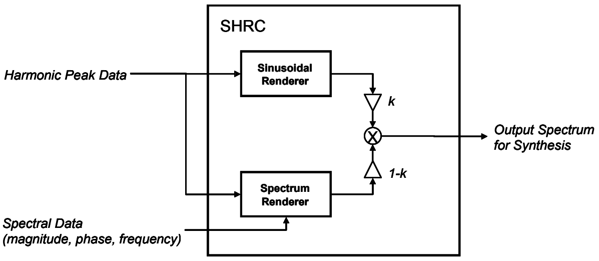
SHRC
Harmonic Peak Data
Sinusoidal Renderer
k
Output Spectrum for Synthesis
1-k
Spectrum Renderer
Spectral Data (magnitude, phase, frequency)

Figure 5: Sub-Harmonic Component Reduction.

In first experiments we found out that the pure harmonic series sounds unnatural, because of missing noisy components. We used two approaches to tackle this problem:

- The naturalness of the voice improves when the Sub-Harmonic Reduction is only applied in the frequency range from 0 Hz to 13 kHz, (with a given sample rate of 44.1 kHz). The frequency range from 13 kHz to 22.05 kHz remains untouched.

- Apply just the minimum amount of Sub-Harmonic Reduction so that perceptively we are not able listen the hoarseness or growl anymore.

Obviously in singer's performance where roughness and growl are part of his expression the existing amount of sub-harmonics may be time-varying, so that we continuously had to adjust the interpolation factor *k* to achieve the best compromise between naturalness on the one hand and non-perceptive sub-harmonics on the other hand. To get rid of the continuous manual adjustments we propose the *Expression Adaptive Control*, which measures the amount of sub-harmonic energy in the original spectrum and adaptively controls the amount of interpolation between original spectrum and the PHS (see section 5).

4.2. Vocal Tract Excitation Modification

The Vocal Tract Excitation Modification algorithm is based on the Excitation plus Resonance Model [1] (EpR). In singing voice the energy of higher harmonics depends on its singer's vocal tract excitation. The EpR model describes this behaviour with a decaying exponential function.
The idea behind Vocal Tract Excitation Modification is to equalize the voice the way as it would have less excitation. It calculates the difference of the measured EpR curve and the

desired one for all harmonics. The difference of both spectral envelopes is the resulting frequency domain filter curve to be applied to the signal.

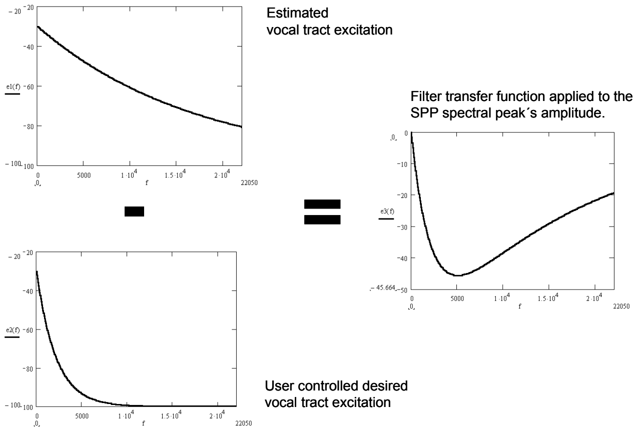

Figure 6: Measured and desired vocal tract excitation and difference of both.

The user controls the excitation slope parameter of the desired EpR curve in a meaningful range that has be measured a-priori from a number of voice recordings.

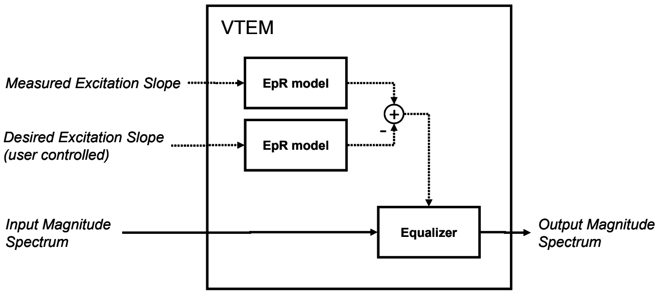

Figure 7: Vocal Tract Excitation Modification

4.3. Intonation Modification

The Intonation Modification transformation is achieved by low-pass filtering the detected pitch, see figure 8 (original and smoothed, i.e. low pass filtered). According to the amount of deviation from the original pitch the original spectrum is pitch transposed towards the smoothed pitch.

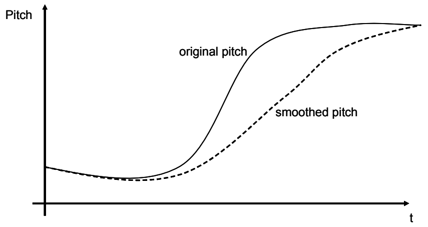

Figure 8: Original and smoothed pitch curve.

5. EXPRESSION ADAPTIVE CONTROL

As we have mentioned before, the vocal disorders that we intend to minimise are perceptually present mostly in steady vowels sounds. Therefore, it seems convenient to apply a control signal, so that the original signal is processed by the Sub-Harmonic Reduction algorithm only when the disorders are perceptually relevant. We call this signal *Expression Adaptive Control* (EAC), since it controls the Sub-Harmonic Reduction depending on the expressiveness of the singer's voice. Basically, the EAC drives the Sub-Harmonic Reduction algorithm in such a way that the degree of interpolation between original spectrum and the pure sinusoidal spectrum, varies dynamically. The original voice is kept unaltered in case of transients and healthy phonation, which ensures a more natural sound.

Our first task is to define a method for identifying a voice with vocal disorders. As described in section 1.2 in a growl or hoarse voice besides the harmonic partials additional sub-harmonics are found. In the field of *Perceptual Audio Coding* appears often the idea of identifying the noisiness of an audio signal. Common methods are the *Spectral Flatness Measure* (SFM) and *Tonality* [6]. Here we use a similar concept for identifying the sub-harmonic components.

As input data, the Expression Adaptive Control takes values from the Spectral Analysis block: Spectral Peaks, Harmonic Peaks, Pitch and Excitation Slope. The *Sub-Harmonic Factor* is computed using a formula that derives from Harmonic Peaks and Spectral Peaks values, which are stored in arrays. Both arrays contain peak information of magnitude, frequency and phase. The first step is to divide the spectrum in regions around each Harmonic Peak [2]. We assume that peaks of frequency above 3.5kHz are not relevant for our estimation, thus we consider only the lower frequency range. In the figure 9, we observe the Harmonic Peaks, the Harmonic Region (centred on each harmonic partial), and all sub-harmonics with lower energy.

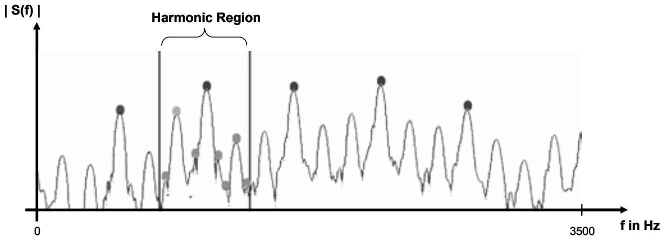

Figure 9: Example of vocal disorder. The spectrum (0-3500 *Hz*) shows clearly the presence of sub-harmonics, in addition to the harmonic partials (marked with dark circles).

5.1. Sub-Harmonic Factor

In a first study, we attempted to compute the Sub-Harmonic Factor (SHF) using the Spectral Flatness Measure of each

Harmonic Region. It is defined by Johnston [6] as the ratio between arithmetic and geometric means of the spectral power density function, and computed directly from the FFT. Experimental results tended to be misleading, when the analysed frame contained harmonics of high bandwidth, for instance in case of vibrato.

Therefore, we developed a new approach considering for each region only the magnitude of the spectral peaks as valid data. In the equation 3, the Spectral Peaks are represented by the vector *SPeak[]*, and the region's Harmonic Peak is represented by $HPeak_r$. We call it *Region Sub-Harmonicity* (RSH_r).

$$RSH_r = \frac{\left|HPeak_r\right|}{\sum_i \left|SPeak\left[i\right]\right|} \tag{3}$$

Then we calculate the Sub-Harmonic Factor (SHF) as the average of all Region Sub-Harmonicity (RSH_r) values. We only consider up to a frequency of 3.5 kHz (Equation 5). Note that for frames with a high number of sub-harmonics, RSH_r tends to 0. Thus, in the final formula (Equation 5), *SHF* tends to *1.0* for a high sub-harmonicity, and is *0* in case of a signal with solely pure harmonics.

$$SHF = \frac{1}{R}\sum_r \left(1 - RSH_r\right), \qquad f_r \le 3500Hz \tag{5}$$

5.2. Excitation Slope and Smoothing

Additionally to SHF, the EAC signal is controlled by the Excitation Slope. The Excitation Slope is an output parameter of the EpR model, and describes the exponential decay of the harmonic partials in the vocal excitation, see section 3.4. After analysing a number of voice recording with vocal disorders, we observed that the ”hoarseness” effect was perceptually more present in vowel sounds such as: /a/ and /e/. We studied the results thoroughly using the EpR model [2] and observed that the Excitation Slope parameter was highly correlated with the mentioned vowel sounds.

The resulting EAC signal is the Sub-Harmonic Factor, weighted with the Excitation Slope and smoothed in time using a moving-average filter. Since our system works in frequency domain at a frame rate, the temporal smoothing here is also referred to a frame rate.

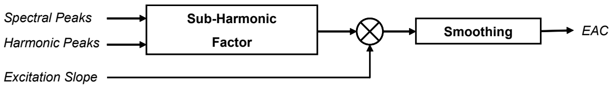

Figure 10: The EAC signal consists of the Sub-Harmonic Factor, which is weighted by Excitation Slope and finally smoothed.

Figure 11 shows a timeline with the evolution of different parameters: *original audio*, *pitch*, *Sub-Harmonic Factor*, *Excitation Slope* and EAC. In this example, the original audio only presents disorders in the first part. The EAC signal indicates the amount of sub-harmonic reduction that the Sub-Harmonic Reduction algorithm will apply. As we can see the sub-harmonic reduction will be applied mostly in the third note (/iee/). It is also noticeable that the first two notes (/uh/), also with a high SHF, are weighted with the Slope signal.

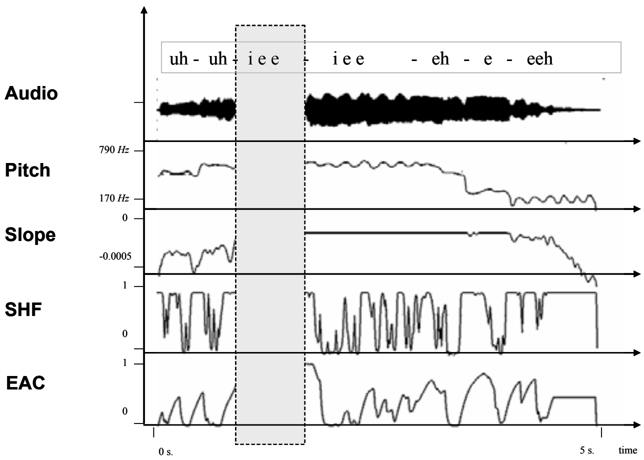

Figure 11: The evolution of the different signals involved in the EA is shown in this example. The *Pitch* signal shows the different notes. The grey area marks the segment where hoarseness is present in the voice.

6. RESULTS AND CONCLUSIONS

The proposed system is able to transform a growl or hoarse a singing voice into a sweet sounding voice using Sub-Harmonic Component Reduction combined with Intonation Modification and Vocal Tract Excitation Modification. The system works offline or in real-time so that it can also be used for a live performance. The system provides professional sound quality at a sampling rate of 44.1 kHz.

The Sub-Harmonic Component Reduction works for most types of voices with vocal disorders. Singing voices performing at low pitch with a high amount of growl turned out to be still problematically. In this case harmonics and sub-harmonics are strongly frequency modulated so it becomes difficult to distinguish them in the spectral domain. If applied without Expression Adaptive Control with maximum reduction of sub-harmonics the transformed voice sounds less natural because of its pure sinusoidality. An approach to improve this behaviour is to preserve subtle noisy components of the original voice which do not belong to the category of sub-harmonics. Using the Expression Adaptive Control the unwanted effect of sinusoidality mostly disappears, except for voices with a strong growl expression.

Experimenting with a number of vocal recordings taken from several sampling libraries the Expression Adaptive Control performed well for most examples. Nevertheless it lacks robustness, because the control signal may deviate significantly depending on the number of spectral peaks detected for each harmonic region. Therefore unnecessarily Sub-Harmonic Component Reduction is applied although no sub-harmonics are present in the spectrum. This is perceived as a degradation of the voice's naturalness.

Although the system requires improvements regarding its robustness we consider it a very useful tool for any recording or post-production studio as well as for singer's live performance.

Different results can be auditioned at:
http://www.iua.upf.es/~lfabig/sweetnesseffect.

7. REFERENCES

[1] J. Schoentgen, "Stochastic models of jitter", J.Acoustic Soc. Am. 109 (4), April 2001.

[2] J. Bonada, A.Loscos, "Sample-based singing voice Synthesizer by spectral concatenation", Proceedings of Stockholm Music Acoustics Conference, August 6-9, 2003.

[3] J. Laroche, "Frequency-Domain Techniques for High-Quality Voice Modifications", Proc. of the 6th Int.Conference on Digital Audio Effects, September 8-11, 2003.

[4] Childers, D.G., "Measuring and Modeling Vocal Source-Tract Interaction", IEEE Transactions on Biomedical Engineering, 1994.

[5] J. Sundberg, "The Science of the Singing Voice", Northern Illinois Press, 1987.3

[6] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria" in *IEEE on Selectesd Areas in Communications*, 1988.

[7] CLAM 2004. "C++ Library for Audio and Music". Music Technology Group, Audiovisual Institute, Universitat Pompeu Fabra, http://www.iua.upf.es/mtg/clam.