EMULATING ROUGH AND GROWL VOICE IN SPECTRAL DOMAIN

Alex Loscos, Jordi Bonada

Music Technology Group of the Institut Universitari Audiovisual
Universitat Pompeu Fabra, Barcelona, Spain
{alex.loscos|jordi.bonada}@iua.upf.es

ABSTRACT

This paper presents a new approach on transforming a modal voice into a rough or growl voice. The goal of such transformations is to be able to enhance voice expressiveness in singing voice productions. Both techniques work with spectral models and are based on adding sub-harmonics in frequency domain to the original input voice spectrum.

1. INTRODUCTION

Vocal disorders have been largely studied as pathology in the field of phoniatry. However, in the context of popular singing, vocal disorders not always come from pathologies but sometimes healthy voices use it as an expressive recourse. The goal of the algorithms presented here is to achieve natural rough and growl effects in order to enhance singing voice in music productions. Decide where and how much effect to apply is a critic issue that will not be discussed here.

Unlike many of the studies concerning vocal disorders, the algorithms presented in this paper arise from spectral models and work with frequency domain techniques instead of working with physical models and use time domain techniques. More concretely, both rough and growl algorithms have been implemented on top of phase-locked vocoder techniques [1], [2].

1.1. Voice production mechanism

The first step in the voice production cycle takes place when air enters the lungs via the normal breathing mechanism. When this air in the lungs is pushed out by muscle force it excites the vocal mechanism through the bronchi and trachea. When the vocal folds are tensed, the airflow causes them to vibrate (Figure 1), producing voiced speech sound. In this case the airflow is chopped by the vocal folds into quasi-periodic pulses.

When the vocal folds are relaxed, in order to produce a noise, the airflow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sound, or it can build up pressure behind a point of total closure within the vocal tract (e.g. the lips), and when the closure is opened, the pressure is suddenly released, causing a brief transient sound.

The most common approach to model the voice production system is based on a source-filter decomposition assumption [4]. The speech processing discipline has been using widely this source-filter model in which two types of source signals (noise for unvoiced and a periodic pulse-train for voiced) are filtered by a dynamic filter that emulates the vocal tract (supra-laryngeal filter).

Figure 1: *Video capture of a female larynx taken from* [3] *with permission of author. Vocal folds and aryepiglottic folds can be observed in the center and lower part of the laryngoscopic view respectively*

Note that from the voice-source model point of view, the vocal disorders that are being considered here come basically from the aperiodicities of the voiced excitation, that is, from the periodic train of pulse-like waveforms that corresponds to the voiced glottal excitation.

2. ROUGHNESS

Roughness in voice can come from different pathologies such as biphonia, or diplophonia, and can combine with many other voice tags such as hoarse or creaky [5]. In this paper we will not stick to the rigorous rough voice definition but we will refer to rough voice as the one due to cycle to cycle variations of the fundamental frequency (jitter), and the period amplitude (shimmer).

Being aware of such nomenclature, we can say the most common techniques used to synthesize rough voices work with the source - filter model and reproduce the jitter and shimmer aperiodicities in time domain [6]. These aperiodicities can be applied to the voiced pulse-train excitation by taking real patterns that have been extracted from rough voices recordings or by using statistical models [7].

2.1. The roughness algorithm

The main idea underneath our algorithm for turning a normal phonation voice into a rough voice is to take the original input signal, transpose it down a certain number of octaves, take then the transposed signal and shift it and overlap it with a certain amount of randomness (Figure 2) to re-synthesize the original voice with its new rough character.

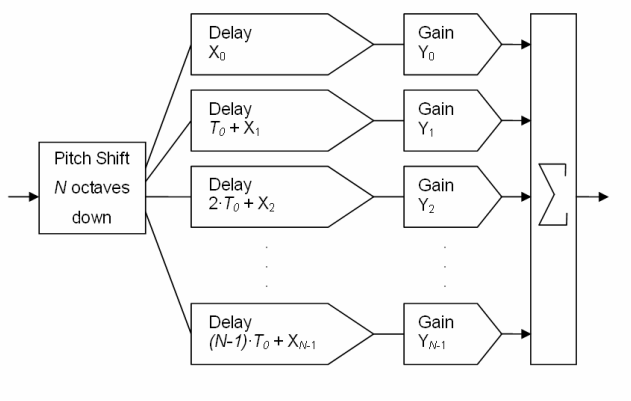

Figure 2: *Block diagram of the rough emulator.*

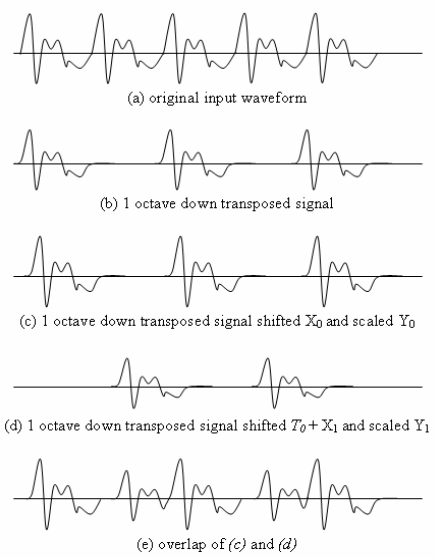

Figure 3: *Figurative representation of the waveforms at different points of the algorithm for N=2.*

The shift applied to each of the N shifted versions of the transposed signal is:

$$Shift_i = i \cdot T_0 + \mathrm{X}_i \quad for\ i=0,1,..,\ N\text{-}1 \tag{1}$$

where T_0 is the original period and X_i is a zero mean random variable. These differently shifted N versions of the transposed signal are then scaled by a unity mean random variable Y_i and finally overlapped.

In order to take in what is the outcome of such system, Figure 3 shows a figurative time domain representation of all steps for $N=2$.

Figure 3 does not illustrate real results since our frame-based implementation does not take into account the relationship between the frame rate and the input period, nor the analyzed frame history. Also, the real implementation changes X and Y stochastic variables values at every frame time. This scenario does not allow generating patterns such as the ones represented in the figure. This is also the reason why even though theoretically, with such algorithm, the higher N is the more control we may have over isolated periods of the signal, the implemented system does not fulfil this rule.

2.2. The simplified roughness implementation

The reason for implementing a very simplified version of the algorithm presented in previous Section is that the effect had to fit in a real time voice transformation environment. All performed simplifications are described next in this Section.

The one octave down transposition is accomplished by adding pure sinusoids to the spectrum in the sub-harmonic frequencies. More precisely, adding the main lobe bins of the analysis window and taking the phase from the closest harmonic peak and shifting it with the corresponding offset as in [8]:

$$\Delta\varphi = 2\pi \cdot f_h \cdot \left(\frac{f_{sh}}{f_h} - 1 \right) \cdot \Delta t \tag{2}$$

where f_{sh} is the sub-harmonic frequency to fill, f_h is the frequency of the closest harmonic peak, and Δt is the frame time.

Since the greater N is, the more computationally expensive the effect is, we have taken the minimum N value, N=2.

The jitter and shimmer stochastic variables of the first channel are set to its mean value X_0=0 and Y_0=1. Thus, the output of this first channel will be a one octave down transposition of the original input. This is a not very risky simplification for N=2 since it can be seen as moving X_0 and Y_0 randomness to X_1 and Y_1.

The stochastic variables X_1 and Y_1 are defined to have a normal distribution with variances 3% of the input signal period, and 3 dBs respectively.

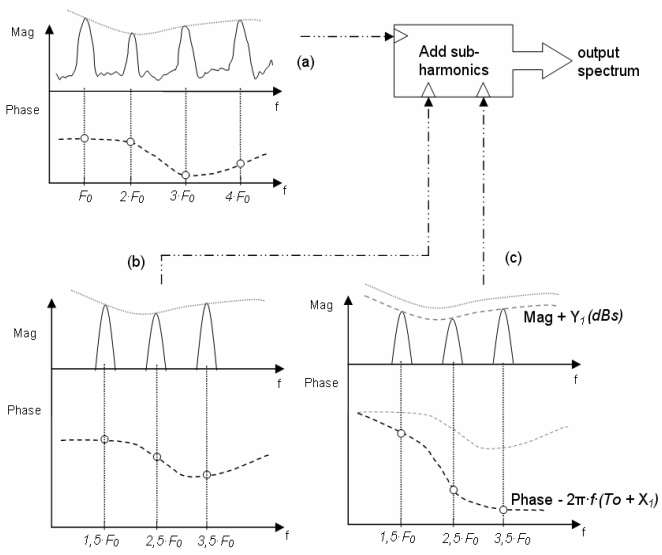

Figure 4: *Representation of the roughness straight forward implemented system where (a) is the original input signal spectrum (b) are the sub-harmonics that generate the transposed version of the signal, and (c) are the sub-harmonics that generate Figure 3d signal*

The random scaling due to Y_1, as well as the random time shift due to $T_0 + X_1$ are only applied to the sub-harmonics. The only reason for doing such oversimplification is to reduce the computational cost of the algorithm since with this only half of the peaks to which the random variables should be computed and applied are actually processed. The time shift is applied in frequency domain by adding the corresponding constant slope phase offset to the phase of the sub-harmonics spectrum as represented in the spectrum of Figure 4c.

Only sub-harmonics inside the F_0-8000 Hz band are added to the spectrum. Upper sub-harmonics are not significantly relevant in terms of acoustic perception to reproduce the rough effect, and the first sub-harmonic (placed at $0.5F_0$) is assumed to be, based on the observations, always masked by the amplitude of the fundamental peak.

3. GROWL

Singers in jazz, blues, pop and other music styles often use the growl phonation as an expressive accent. Perceptually, growl voices are close to other dysphonic voices such as hoarse or creaky, however, unlike these others, growl is always a vocal effect and not a permanent vocal disorder.

According to [9] growl comes from simultaneous vibrations of the vocal folds and supra glottal structures of the larynx. The vocals folds vibrate half periodically to the aryepiglottic fold vibration generating sub-harmonics.

The growl algorithm presented here adds these sub-harmonics in frequency domain to the original input voice spectrum to try to emulate the growl phonation. These sub-harmonics follow certain magnitude and phase patterns that have been extracted from the spectral analysis and observation of real growl voice recordings.

3.1. The growl observation

The behaviour of the growl sub-harmonics in terms of magnitude and phase vary quite a lot from one voice to another, from one pitch to another, from one phrase to another, etcetera. However certain patterns appear quite frequently. These patterns, which are explained next, are the ones that the growl effect applies.

If a growl utterance is observed in time domain, it is most of the times easy to recognize which is the real period of the signal and which is the macro period due to growling as it is in Figure 5. In the observations made growl phonation appeared to have from two to five sub-harmonics. In Figure 5 example, the spectrum presents three sub-harmonics placed at $F_0 \cdot (m+k/4)$ (for m=0..number of harmonics, and k=1..3). Thus, three inner periods can be distinguished in between a growl macro period.

Regarding the magnitudes, in the band that goes from the fundamental frequency up to approximately 1500 Hz, the sub-harmonic peaks are commonly located below the spectral envelope (defined by the harmonic peaks). In this band, the closer the sub-harmonic is to the nearest harmonic, the higher its magnitude is. In the upper band, from approximately 1500 Hz to half the sampling rate, sub-harmonics go along with the harmonic spectral shape.

Regarding the phase, for a growl utterance with N sub-harmonics, the typical behaviour of the phases of the sub-harmonics is to get approximately aligned with the phase of the left harmonic peak every N+1 periods as illustrated in Figure 6.

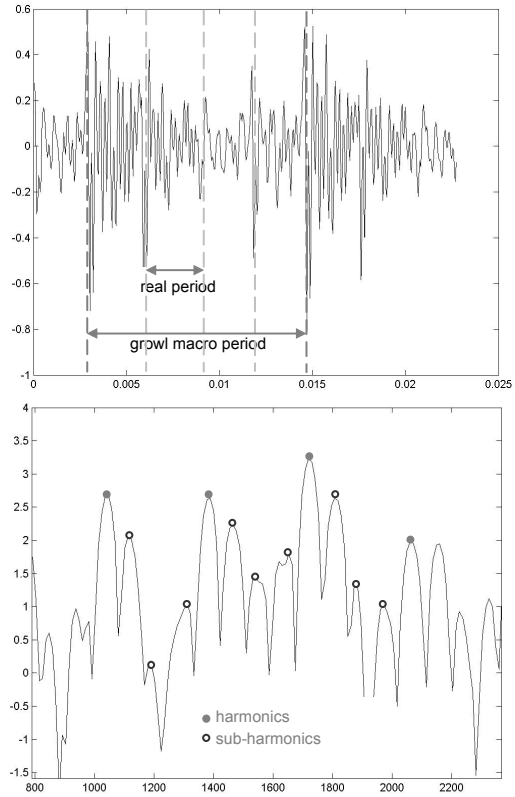

Figure 5: *Partial waveform (upper) and spectrum representation of a growl utterance*

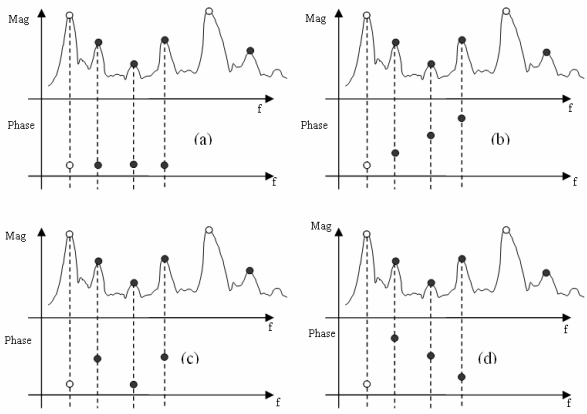

Figure 6: *Representation of the spectrum phase behavior of a growl voice in the beginning of four consecutive periods (a,b,c,d) for N=3 sub-harmonics*

For the harmonic peaks, harmonic i is always 1 cycle below harmonic i+1. In between them, for the sub-harmonics peaks, the peak phase can be generally expressed as:

$$\varphi_{p,k}^{sh} = \varphi_i^h + \frac{2\pi}{N+1}\cdot(k+1)\cdot p\,, \quad \text{for } k=0,1,2 \text{ and } p=0,1,2,3 \qquad (3)$$

being p the inner period index (p=0 for Figure 6a and p=3 for Figure 6d), k the sub-harmonic peak index in between consecutive harmonic peaks, and N the number of sub-harmonics.

3.2. The growl effect implementation

Based on the most frequently observed growl spectral symptoms, the implemented system fills the original spectrum with sub-harmonics. However, since growl is not a permanent disorder, the effect can not be applied all along the performance. For this reason the implementation includes an automatic growl control (as shown in Figure 7) by which we determine how much of the effect has to be applied at each time depending of the input singing voice. This control is mainly based on the first derivatives of the fundamental frequency and energy and has control on how many sub-harmonics have to be added, and their phase and magnitude patterns (including the gain of the sub-harmonics).
With such implementation, the system is able to reproduce growl sub-period amplitude patterns as the one shown in Figure 8. In the waveform view of the transformed voice we can observe how each of the four periods of the growl macro period is set to have different amplitude. This amplitude modification is achieved by applying phase alignment patterns extracted from real growl analysis to the sub-harmonics.

4. CONCLUSIONS AND FURTHER WORK

The rough and growl algorithms presented in this paper have proven to be suitable in changing the voice character. However, the naturalness of the effect is highly dependent on input voice.

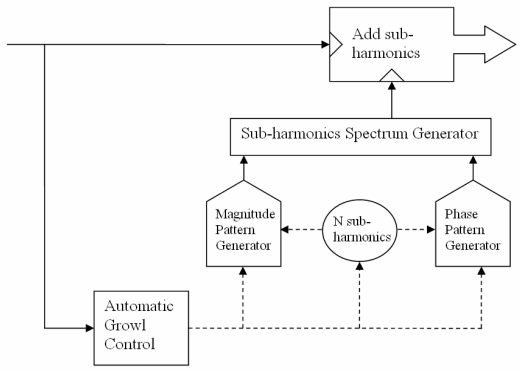

Figure 7: *Block diagram of the growl implementation.*

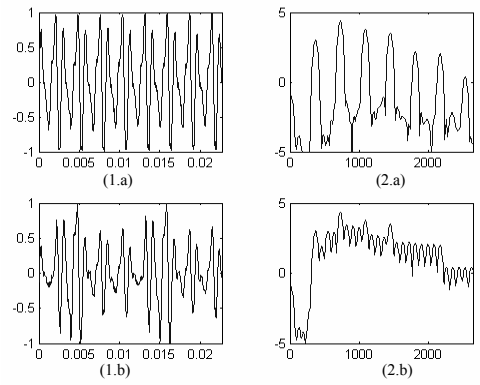

Figure 8: *Partial views of waveform (1) in seconds and magnitude spectra (2) in Hz from both original (a) and transformed (b)*

For different types of voice, different tessitura, different expressions, etcetera different values of the transformation parameters are required. In that sense, a dynamic automatic control over these parameters has to be found. In the growl effect, this control would have to combine and work together with the current automatic growl control.

In the growl effect patterns extracted from real growl recordings are roughly reproduced in synthesis. This means the period to period amplitude envelope inside a growl macro-period is not only included in the phase alignment of the sub-harmonics but also in the sub-harmonics amplitudes. However, it is a tedious job to find the sub-harmonic amplitudes and phase alignment required for a certain made-up amplitude envelope. It is also remarkable no control over the jitter is available with the current growl implementation.

Concerning the rough effect, two interesting directions come up from the current implementation. First, perform a study of the system without any of the simplifications implemented. Second, take into account the frame rate / input period relationship and the analyzed frame history so that the system could follow the period. In that situation, increasing *N* would really improve the resolution of the algorithm. This could be considered as going towards a fusion of both techniques in which we would have control over the period to period jitter and the shimmer.

5. REFERENCES

[1] M. S. Puckette, "Phase-locked vocoder" in *Proc. of IEEE Conf. on Applications of Signal Processing to Audio and Acoustics*, Mohonk, 1995.

[2] J. Laroche, M. Dolson, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects" in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[3] J. P. Thomas, "Voice Disorders," in `www.voicedoctor.net`.

[4] G. Fant, "The source filter concept in voice production," Royal Institute of Technology, Stockholm, Sweden, pp. 21–37, 1981.

[5] I. R. Titze, "Workshop on Acoustic Voice Analysis, Summary Statement," Nat. Center for Voice and Speech, Denver, Colorado, 1994.

[6] D. G. Childers "Speech Processing and Synthesis for Assessing Vocal Disorders" in *Engineering in Medicine and Biology Magazine, IEEE*, Mar. 1990, vol. 9, no. 1, pp 69–71.

[7] J. Schoentgen, "Stochastic models of jitter" *J. of Acoust. Soc. of America*, April 2001, vol. 109, no. 4, pp. 1631–1650.

[8] J. Bonada, A. Loscos, "Sample-based singing voice synthesizer by spectral concatenation," *J. of AES*, May 2000, vol. 48, no. 5, pp. 490–498.

[9] K.-I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama, "Growl voice in ethnic and pop styles" in *Proc. Int. Symp. on Musical Acoustics (ISMA 2004)*, Nara, Japan, April 2004.