GENERATION OF GROWL-TYPE VOICE QUALITIES BY SPECTRAL MORPHING

Jordi Bonada *Merlijn Blaauw*

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
Email: {jordi.bonada, merlijn.blaauw}@upf.edu

ABSTRACT

In this paper we introduce a morph-based approach for generating voice source aperiodicities frequently associated with strong vocal expressions, especially in singing. In our approach the excitation characteristics of one signal are combined the fundamental frequency and spectral envelope characteristics of another signal. An exemplar sustained sample of the target voice quality is looped and resampled in the time domain in order to generate a continuous signal matching the input voice's fundamental frequency. This operation preserves most of the sample's voice quality characteristics at the cost of scaling time and frequency dimensions. While we found the temporal scaling to be acceptable in many contexts, the frequency scaling has to be inverted in order to generate appropriate spectral content for the source excitation's entire bandwidth. This can be accomplished by the phase-locked vocoder method of modulating harmonic bands in frequency. Finally, the input signal's harmonic amplitudes and phases are applied to the transformed morph sample, allowing for a simple one-dimensional control of morph amount by linear interpolation with the input signal. The proposed system is evaluated and the results are discussed.

Index Terms— speech synthesis, singing synthesis, human voice, harmonic analysis, voice quality

1. INTRODUCTION

Over the past years many researchers have identified the importance of voice qualities outside of the "modal" range in transmitting emotion and mood [1]. Varying voice quality is also considered important for obtaining a natural-sounding results in a wide range of speech processing applications such as TTS, voice conversion and voice transformation. The focus of this paper is on a class of voice qualities that are often described as "harsh", "rough", "hoarse" or "growl", which we will refer to as "growl-type" in this paper. In singing, these growl-type voice qualities are frequently used as an expressive resource or may be part of certain types of singing voices (e.g. a Louis Armstrong-type singing voice) [2].

The spectrum of "modal" singing voice can be approximated by a series of relatively stable harmonics and low energy noise. In contrast, what we consider the defining property of growl-type voice qualities is the presence of rapid changes of timbre, timing and strength of source excitation events. These changes result in the appearance of sub-harmonics in the spectrum, and modulation patterns in the time-domain waveform (jitter and shimmer). If those modulations are periodic, the fundamental period and the period of modulation (macro period) can be clearly visible. Within the scope of this paper we assume that for growl-type voice qualities the perceived effect is mainly a result of the signal's source excitation rather than its spectral envelope (aside from pulse-to-pulse variations).

Our goal is to be able to generate growl-type voice qualities in a given, principally modal, singing voice signal, whether it is recorded or synthetic. In other words, we want our output signal to contain a source excitation appropriate for a growl-type voice quality, while at the same time containing pitch and spectral envelope obtained from the input voice. We can approach this problem from two directions. The more common approach is to model the characteristics of the source excitation associated with growl-type voice qualities and then apply these to the modal input signal. A second approach, the one that is pursued in this paper, is to take a recording that already contains the desired voice quality and then modify its fundamental frequency and spectral envelope to match that of the input voice.

There has been a considerable amount of work on the first type of approach, the parametric modeling of growl-type source excitation. Schoentgen [3] for instance gives an extensive overview of different methods of generating jitter and shimmer patterns, which is typically derived by statistical analysis of real recordings, or using models inspired by the physiology of the voice source. The modulation patterns are then typically applied by identifying and transforming individual voice source pulses in the input signal using (e.g. using TD-PSOLA). Examples of applications using this method include voice transformation [4] and voice conversion [5]. An alternative approach proposed by Loscos [6] models the amplitude and phase behavior of sub-harmonics in the time-frequency domain. The main problem with these approaches is that accurate estimation of parameters of these models can be difficult and costly, and the underlying model itself may have certain limitations, such as not considering pulse-to-pulse variations in timbre. There can be a high variability in properties of growl-type voice qualities between different singers or even between different utterances of the same singer, which makes coming up with a single model that covers all cases difficult. Finally, a problem that is often men-

tioned with these methods is that the results are very dependent on the type of input voice used.

The second type of approach has been relatively unexplored. However, it could be considered a case of voice morphing, a concept that, in many variations, has been applied to a wide range of speech processing applications. Typically these use parallel recordings, apply a time-alignment between source and target, separate source excitation and spectral envelope for both, and do some form of interpolation or replacement of either or both of the components. Cano [7] applies the time-aligned spectral envelope of one singer with the excitation (pitch mainly) of another to allow real-time singer impersonation for Karaoke. Another application is to smoothly interpolate between two speakers by independently interpolating source excitation and spectral envelope [8]. A more restrictive case of morphing has been applied to interpolate between spectral envelopes associated with certain voice qualities or emotions of a single speaker or singer [9, 10]. The main problem with applying any of these existing studies to our problem is that in our application, and many others, parallel recordings are not available.

Another basic difficulty with using a morph-based approach is that processing voice signals with growl-type voice qualities can be difficult. Most vocoders used to transform voice signals are designed for principally harmonic signals and are not particularly suited for handling sub-harmonics or detailed jitter/shimmer patterns. Using these methods directly will typically result in low quality outputs or a loss of voice quality. For instance, the phase-locked vocoder [11] where harmonic frequency bands are shifted in order to pitch-shift signals works well when there is only one partial present per harmonic band. However, if a harmonic band contains sub-harmonics, all partials should be frequency-shifted by different amounts and their phases propagated differently. Even the more advanced approaches that explicitly include a model of the glottal source (e.g. [12, 13]) are typically limited to the lax/modal/tense range of voice qualities. Kawahara [14] recently proposed some vocoder methods designed to specifically deal with the fine-grained fundamental frequency variations often observed in growl-type voice qualities. While these approaches seem to work well for resynthesis without modification, whether they can provide satisfactory generation or transformation of growl-type voice qualities has not yet been fully explored.

2. PROPOSED SYSTEM

Our aim is to modify the voice quality characteristics of a given voiced utterance, without having to explicitly model the inner structure of the spectrum beyond its harmonic structure. For that purpose, we propose to morph the input signal with a sample that already posses the desired voice quality. The idea is to combine the excitation characteristics of this sample with the timbre characteristics of the input voice utterance. The sample is not constrained to be phonetically equivalent to the input voice. On the contrary, the idea is to use a reduced database of voice quality samples for any target utterance. In addition, we intend to integrate the morph effect into an existing singing voice synthesizer, and to be able to control the degree of morph. Figure 1 shows the steps of the proposed morphing system.

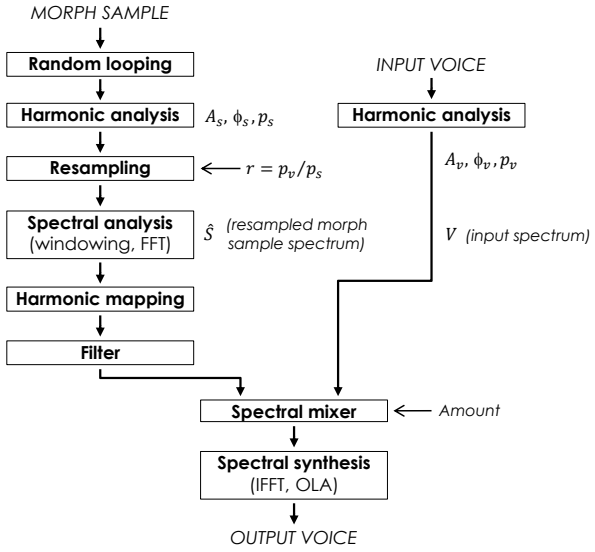

Fig. 1. Block diagram of the proposed system.

First a sample of the target voice quality must be prepared. A sustained vowel with rather stable fundamental and consistent voice quality is preferred. The fundamental frequency can be precomputed and, if needed, manually corrected. To produce a signal of any duration, two loop points are manually set. Looping is done alternating directions and applying small random offsets to the loop points to reduce perceived repetition. In order to match the fundamental frequency of the input voice, the looped sample is then resampled in the time-domain using a windowed sinc. Resampling is especially well-suited for our purposes because it does not require identifying the signal's spectral structure (harmonics, sub-harmonics, noise) while it preserves the excitation characteristics with high quality. The drawback is the inherent scaling of time and frequency dimensions.

For a given frame, we define the resampling factor as, $r = p_v/p_s$, where p_v and p_s are the fundamental frequencies of the input and sample voices respectively. A value greater than 1 means that the morph sample is played back faster than it was recorded, and therefore the duration gets shorter and frequencies higher. In other words, $s_t = 1/r$ and $s_f = r$, where s_t is the temporal scaling and s_f the frequency scaling. Temporal scaling can be acceptable for certain voice qualities, or in general if the resampling factor is small enough. This can be minimized using several samples at different fundamental frequencies for each target voice quality. One appropriate scenario is an arpeggio of vowels, where notes are looped and interpolated. In addition, the arpeggio can also be good for dealing with pitch-dependent effects within a given voice quality.

Frequency scaling affects both the frequency of the harmon-

ics and the voice timbre, defined here as the harmonic spectral envelope. Since harmonics are shifted in frequency, the noise present in their frequency bands is also shifted, producing unnatural synthesis results. A solution is to invert the frequency scaling by shifting in frequency the harmonic bands. Depending on the resampling factor, this means dropping or repeating some harmonics frequency bands. The mapping function is defined as,

$$m_i = \left\lfloor \frac{i}{r} + 0.5 \right\rfloor \qquad i = 0, 1, \ldots, N-1 \tag{1}$$

where i is the harmonic index and N the number of harmonics. Note that the mapped index is constrained to be a positive integer. The phase-locked vocoder [11] is an appropriate technique here. One argument is that the resampled signal already matches the target fundamental frequency, and that therefore the frequency and phase relation between harmonics and their surrounding sub-harmonics is correct. By contrast, if phase-locked vocoder was used for transposition, in order to maintain the correct frequency relation between harmonics and sub-harmonics, they would have to be shifted by different amounts. This implies having to estimate their parameters. This would be especially difficult when the macro period is not stable since then the spectrum of a single frame would contain different sub-harmonic structures.

The synthesis spectrum Y for the bins k corresponding to the i^{th} harmonic band becomes,

$$Y[k] = \widehat{S}[k + d_i] g_i e^{j\theta_i} \tag{2}$$

$$d_i = \left\lfloor p_v \left(m_i - i\right) \frac{L}{f_s} + 0.5 \right\rfloor \tag{3}$$

where $\widehat{S}$ is the resampled morph sample spectrum, k is the spectral bin index, d_i the frequency shift (in bins) for the i^{th} harmonic band, L the frame length, f_s the sampling rate, g_i is a gain, and θ_i a phase correction. The gain factor compensates for the timbre differences between the input and the resampled sample voices. It is computed as, $g_i = A_v\left[i\right] / A_s\left[m_i\right]$, where A_v and A_s are the harmonic amplitudes for input and morph sample respectively. Harmonic amplitudes are computed using parabolic interpolation of the bins surrounding the main peak of the magnitude spectrum within each harmonic band. The phase correction θ_i compensates for the differences between harmonic phases of the morph sample and the input voice utterance. It is computed as, $\theta_i = \phi_v\left[i\right] / \phi_s\left[m_i\right]$, where ϕ_v and ϕ_s the harmonic phases of input and morph sample respectively. Note that when the sample is played backwards in time, the sign of its harmonic phases has to be inverted.

A potential problem arises when the morph sample is transposed to a higher fundamental frequency, i.e. $r > 1$. In this case, some of the high frequency content of the spectrum may fall above the Nyquist frequency, therefore introducing some aliasing. If we apply a low-pass filter to correct this, than we eliminate the high frequency harmonics, and we cannot use them at synthesis (harmonic mapping). This can be quite noticeable, especially for high transposition factors. A better approach, although increasing computational cost, is to oversample the signal to a sampling rate high enough as to raise the Nyquist frequency above the last harmonic frequency band after resampling. This implies a longer analysis window length, so more computations involved in computing the FFT. However, the remaining steps of the system are not affected at all.

In our experiments, we initially computed harmonic parameters from the resampled morph sample. We found that often the growl morph effect was reduced for low target fundamental frequencies. This is related to the change of the ratio between window length and fundamental period. In our experiments the synthesis window length is constant (2048 points at 44.1 kHz). This is motivated by the intended integration of the morph effect into an existing singing synthesizer. Therefore the change of fundamental frequency during resampling modifies the relative window length respect to the fundamental period. When the analysis window is longer than the macro period, sub-harmonics appear in the spectrum. However, when the morph sample is transposed down by resampling, the relative window length gets shorter and sub-harmonics disappear. In other words, we could say that instead of quasi-stable sinusoids plus sub-harmonics, the spectrum exhibits non-stationary (or modulated) sinusoids, and harmonic parameters vary significantly in consecutive frames. In such case, the harmonic gains g_i actually remove these inter-frame amplitude variations, and at synthesis we obtain quasi-stable sinusoids.

Nevertheless, the situation can be significantly and easily improved if we estimate harmonic parameters from the original morph sample, before resampling, ensuring we apply a window long enough so that sub-harmonics appear in the spectrum. This is illustrated in Figure 2. Note that harmonic parameters no longer match the actual spectrum of the resampled signal, but are the ones used to compute the gain and phase corrections, and to transform the spectrum. Using this method the synthesized voice mostly preserves the modulations of the morph sample, and the degree of growl-type voice quality remains similar regardless of resampling factor.

Once harmonic frequencies, phases and amplitudes of the transformed morph sample and the input signal match, they can be freely mixed. We can vary between modal and growl voice qualities, for instance, by simply controlling a linear interpolation factor. While other mixing strategies, such as frequency-dependent mixing, are possible, this approach does have its limitation. For instance, we cannot directly control aspects such as jitter/shimmer amount or macro period duration.

3. EVALUATION

In order to get an idea of the performance of our system we conducted two MOS-type listening tests. In total 17 subjects participated, most having a background in music and signal processing.

The first test consisted of singing voice recordings morphed

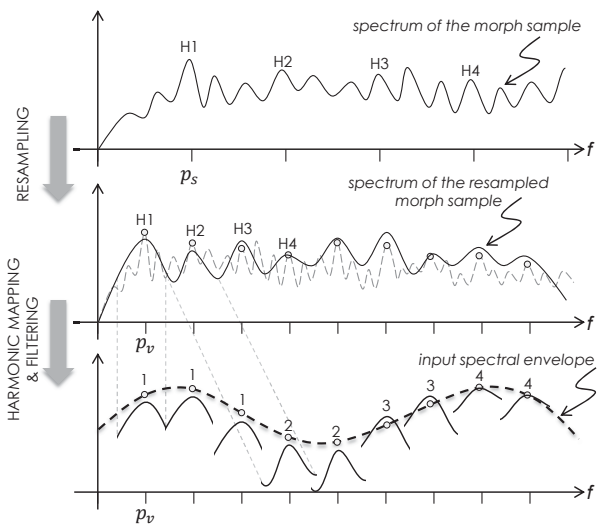

Fig. 2. An illustration of the process of resampling and harmonic mapping. Note that in the solid line of the middle plot sub-harmonics are no longer visible (window size L), while in the dashed line they are (window size $\widehat{L} = L/r$).

with growl sustain samples. We combined 4 recordings containing growl with 5 excerpts where growl was added synthetically by morph (in random order). We used input recordings from 4 different singers; 2 male and 2 female, with pitches spanning about one octave. In all cases the morph sample and input voice were provided by different singers. We asked subjects to rate the "overall quality and naturalness" from 1 to 5, and whether they thought the excerpt was synthetic or a recording.

The second test consisted of a single song generated using a singing voice synthesizer based on concatenation of diphone samples with modal voice quality. To increase naturalness of the synthesis, pitch and phonetic timing information was extracted from a recording. We processed the output of the synthesizer to artificially add growl voice quality in certain places. The morph amount (interpolation factor) control was manually set so that the usage of growls was similar to the recording. Subjects would rate a version with growl, a version without growl and the target recording (with growl), not only for "quality", but also for "expressiveness" (simultaneously).

The results of the listening tests are shown in figure 3. The recorded growls were rated as slightly below "good", while the growls generated by morph were rated as above "average". It is interesting to see the relatively high rate of confusion between real and synthetic growls, shown in Table 1.

Using voice quality morph to add expression to synthesized song resulted in an increase of perceived expressiveness, while perceived quality and naturalness decreases by a similar amount (but still well above "average").

A number of sound examples that we have used in the listening tests can be found online at:
`http://www.dtic.upf.edu/~mblaauw/icassp2013/`

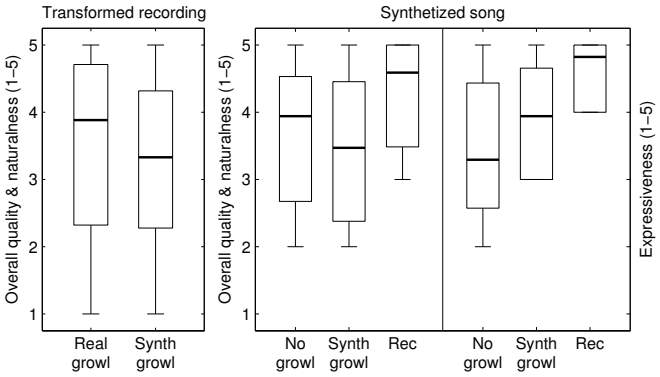

Fig. 3. Results of listening tests showing mean opinion scores (thick line), standard deviations above and under mean (box) and minimums and maximums (whiskers).

		Answer		
		Real	Synthetic	Not sure
Actual	Real	66.17%	19.12%	14.71%
	Synthetic	34.12%	54.12%	11.76%

Table 1. Confusion matrix between recorded growl excerpts and excerpts with synthetic growl generated by morph.

4. CONCLUSIONS

In this paper we have introduced a new method of generating growl-type voice quality in a given input signal based on spectral morphing. Unlike most existing methods, our method does not rely on a parametric model and does not require any parameter estimation of the growl-type signal beyond its harmonic parameters. While similar morph approaches often require parallel recordings, our method instead uses a looped exemplar sample of the target voice quality. This sample is not restricted to exactly match the input phonetically or in pitch. Pitch-shifting is achieved by method based on resampling which is suitable for signals containing sub-harmonics, which is not typically the case with existing vocoders.

While the scope of this paper was limited to growl-type voice qualities, another big advantage this morph-based method is that it is relatively widely applicable. The method shows promising results for other voice qualities such as "tense" and "breathy". However, we have found that for some voice quality transformations a corresponding change in spectral envelope is very significant. While our method only changes source excitation, for the class of growl-type voice qualities we have found this to be sufficient in order to produce convincing results in many cases. The listening tests seem to validate these findings.

5. ACKNOWLEDGMENTS

This paper is the outcome of joint research with Yamaha Corporation. We would like to thank Mr. Hisaminato for his support.

6. REFERENCES

[1] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1-2, pp. 189–212, 2003.

[2] K. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama, "Growl voice in ethnic and pop styles," in *Proc. International Symposium on Musical Acoustics (ISMA)*, 2004.

[3] J. Schoentgen, "Stochastic models of jitter," *J. Acoust. Soc. Am.*, vol. 109, pp. 1631–1650, 2001.

[4] D. Ruinskiy and Y. Lavner, "Stochastic models of pitch jitter and amplitude shimmer for voice modification," in *Proc. IEEE 25th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2008, pp. 489–493.

[5] A. Verma and A. Kumar, "Introducing roughness in individuality transformation through jitter modeling and modification," in *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[6] A. Loscos and J. Bonada, "Emulating rough and growl voice in spectral domain," in *Proc. 7th Int. Conference on Digital Audio Effects (DAFX)*, 2004.

[7] P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications," in *Proc. 2000 International Computer Music Conference (ICMC)*, 2000.

[8] H. Banno, K. Takeda, K. Shikano, and F. Itakura, "Speech morphing by independent interpolation of a spectral envelope and source excitation," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 82, no. 3, pp. 22–30, 1999.

[9] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[10] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure, "Gradually changing expression of singing voice based on morphing," in *Proc. Interspeech*, 2005, pp. 541–544.

[11] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in *Proc. 6th Int. Conference on Digital Audio Effects (DAFX)*, 2003.

[12] A. Roebel, S. Huber, X. Rodet, and G. Degottex, "Analysis and modification of excitation source characteristics for singing voice synthesis," in *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 5381–5384.

[13] H.-L. Lu and J. O. Smith, "Glottal source modeling for singing voice synthesis," in *Proc. 2000 International Computer Music Conference (ICMC)*, 2000, pp. 90–97.

[14] H. Kawahara and M. Morise, "Analysis and synthesis of strong vocal expressions: Extension and application of audio texture features to singing voice," in *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 5389–5392.