ON STRETCHING GAUSSIAN NOISES WITH THE PHASE VOCODER

Wei-Hsiang Liao,Axel Roebel
UMR STMS IRCAM - CNRS - UPMC
Paris, France
wliao@ircam.fr

Alvin W.Y. Su
SCREAM Lab, Dept. of CSIE
Nat. Cheng-Kung Univ.
Tainan, Taiwan
alvinsu@mail.ncku.edu.tw

ABSTRACT

Recently, the processing of non-sinusoidal signals, or sound textures, has become an important topic in various areas. In general, the transformation is done by the phase vocoder techniques. Since the phase vocoder technique is based on a sinusoidal model, it's performance is not satisfying when applied to transform sound textures. The following article investigates into the problem using as example the most basic non-sinusoidal sounds, that are noise signals. We demonstrate the problems that arise when time stretching noise with the phase vocoder, provide a description of some relevant statistical properties of the time frequency representation of noise and introduce an algorithm that allows to preserve these statistical properties when time stretching noise with the phase vocoder. The resulting algorithm significantly improves the perceptual quality of the time stretched noise signals and therefore it is seen as a promising first step towards an algorithm for transformation of sound textures.

1. INTRODUCTION

In present time, there exist many algorithms that have been developed to achieve high quality transformation of sounds. These state-of-the-art algorithms mainly focused on sinusoidal sounds[1] [2] [3] [4]. Such as instruments and tonal sounds. While the properties of sinusoidal sounds are well-researched, a large class of sounds cannot be dealed with the way dealing with sinusoidal sounds. These sounds often driven by a stochastic process. They often called *sound textures*.

Sound textures are common in the environment. They could be either natural of artificial. For example, the sounds of wind blowing and rain dropping are natural textures, and the noise of crowd and the sound of train passing are artificial. Although there's no clear definition of sound texture[5], this class of sounds often exploits certain stable structure over a time period. The structure can be described by a series of statistic properties. There are some different approaches dealing with sound textures. For example, Saint-Arnaud[5] proposed an analysis/synthesis scheme which based on atomic features in sound textures, Hanna[6] used randomized sinusoids to synthesize stochastic noises. Schwarz[7] proposed a descriptor-driven, corpus-based approach to synthesize sound textures. But to enable the further analysis, synthesis and manipulation of sound textures, a parametrized modeling which describes a texture with a set of statistical properties is desired. Fortunately, some previous works on parametrization has been done for image textures, Portilla[8] proposed a texture model based on the statistics of wavelet coefficients. Bruna[9] proposed a new wavelet transform which provides a better view for textures while capturing high-order statistics. For sound textures, McDermott[10] proposed a parametrized model, adapted from [8] and which characterizes target sound texture with high order statistics and correlation between subband envelopes. An algorithm is proposed in his article to convert Gaussian noise into different target textures.

To manipulate sounds, such as time stretching or transposition, using phase vocoder[1] [2] [3] is a common approach. Phase vocoder could stretch a signal while preserving it's envelope and naturality of transient [11] [12] [13]. With phase vocoder, one could perform a large factor stretching which still yields a natural result. While the theory basis of phase vocoder works well for sinusoidal signal, it's not suitable for sound textures. Because of the constraint of phase continuity could no longer be used to estimate output phases. To stretch a sound texture with phase vocoder, one must maintain perceptive important properties of the texture during the stretching.

In this article, we try to resolve the properties being changed during the transformation. In order to aim the phenomenon appeared during the process, the time stretching of white Gaussian noise is used. Since Gaussian noise exploits independency along both time and frequency axes, it introduces no extra dependency to the analysis frames. With this simplified case, we found that it's crucial to compensate the change of temporal correlation during the time stretching transformation. A primitive algorithm is given to correct the distortion of temporal correlation.

The paper is organized as follows: section 1 describes previous works on the sound texture parametrization and the issues encountered when applying phase vocoder to sound texture. Section 2 lists the perceptually important statistical properties of sound textures. Section 3 formulates the time stretching procedure with phase vocoder and correct the result with iterative approaches. Section 4 conducts an experiment of the correction. Section 5 gives the conclusion and future prospective.

2. PROPERTIES OF SOUND TEXTURES

According to Julesz's conjecture[14] and portilla's work[8], the perceptually important properties of a texture can be described in statistical features. In Mcdermott's work[10], it suggests that the perceptually important properties of sound texture could be summarized in three categories : *marginal statistics, frequency correlation* and *temporal correlation*. These properties are further described in the following subsections:

2.1. Marginal Statistics

Marginal statistics are the statistical moments of the spectral coefficients in different time points or different subbands. Though the usage of higher orders are possible, the first four moments are used to form the marginal statistics in common cases [8].

2.2. Frequency Correlation

In general, frequency correlation means the cross-correlation between different frequency components. These components can be simply coefficients, subbands, or features of subbands, for example, the envelope of cochlear bands[15] [10]. Moreover, in [10], each subband is further filtered by a modulation filterbank, and the cross-correlation of the modulation bands are evaluated.

2.3. Temporal Correlation

Temporal correlation refers to the autocorrelation of a frequency component along time axis. It can be the autocorrelation of STFT coefficients[16] or the autocorrelation of a temporal feature. It is also the most important property when performing time stretching by phase vocoder. The detail is described in section 3.

3. TIME STRETCHING GAUSSIAN NOISES

To investigate the problem encountered during the stretching of Gaussian by the phase vocoder, the formulation of phase vocoder must be described. The transformation procedure of phase vocoder consists of analysis, modification and resynthesis [1] [2] [3]. The analysis step first applies STFT to the input signal, then, proceed modification in the spectral domain, and last, the resynthesis is done by inverse STFT and overlap-and-add. Assuming a given input signal s, along with a window function w and the size of Fourier transform N. Then the analysis frame S_l which centered at time point l can be written as:

$$S_l(k) = \sum_n s(n)w(n-l)e^{\frac{-j2\pi nk}{N}} \tag{1}$$

During the transformation, the coefficients of S_l may be modified as well as the centering position l might be moved [17] [3] [11]. If $f(l)$ is the function of the modified center of S_l, then the resynthesis procedure can be written as:

$$s'(n) = \sum_k S_l(k)e^{\frac{j2\pi nk}{N}} \tag{2}$$

$$\tilde{s}(n) = \frac{\sum_n w(n-f(l))s'(n)}{\sum w^2(n-f(l))} \tag{3}$$

If inconsistency happens between overlapping frames, (3) outputs the signal with the least square error respect to the corresponding analysis frames [18]. If we stretch a Gaussian noise (1)-(3) and assign the phase of each frame as the way processing sinusoidal signal, one would find that the noise component disappeared in the resynthesized signal. In fact, short distorted sinusoids will be perceived in the resynthesized signal. In this case, the phase and the time position of analysis frames are modified during the time stretching transformation. While the amplitude of the coefficients is unmodified, the marginal statistics of analysis frames remain the same. Therefore, the possible variation of correlation between STFT coefficients is investigated.

The correlation between two STFT coefficients in the analysis frames is the inner product between one coefficient and the complex conjugate of the other. By using (1), the correlation is:

$$\begin{aligned} C_{S_\alpha(k_1),S_\beta(k_2)} = &\sum_n s(n)w(n-\alpha)e^{\frac{-j2\pi nk_1}{N}} \\ &\overline{*\sum_n s(n)w(n-\beta)e^{\frac{-j2\pi nk_2}{N}}} \end{aligned} \tag{4}$$

The equation in (4) can be used to calculate the correlation in both time and frequency directions. Let the terms in (4) be column vectors, replace the window function $w(n-l)$ by $\mathbf{w}_l$, and $e^{\frac{-j2\pi nk}{N}}$ by $\mathbf{e}_N^k$, (4) can be rearranged into the summation of element-wise product(Hadamard product) of outer product matrices:

$$C_{S_\alpha(k_1),S_\beta(k_2)} = \sum(\mathbf{s} \cdot \mathbf{w}_\alpha \cdot \mathbf{e}_N^{k_1})(\mathbf{s} \cdot \mathbf{w}_\beta \cdot \overline{\mathbf{e}_N^{k_2}})^T \tag{5}$$

$$= \sum \mathbf{s}\mathbf{s}^T \cdot \mathbf{w}_\alpha \mathbf{w}_\beta^T \cdot \mathbf{e}_N^{k_1} \overline{\mathbf{e}_N^{k_2}}^T \tag{6}$$

Equation (6) consists of three parts, the outer product of $\mathbf{s}$, shifted outer product of $\mathbf{w}$, and the outer product between $\mathbf{e}_N^{k_1}$ and $\mathbf{e}_N^{k_2}$. In fact, the latter two terms form the shape of a Gabor filter[19]. The correlation respect to a direction is equal to apply the corresponding Gabor filter to the outer product matrix of s.

Since the stretching is done by moving the center of analysis frames to new positions, their relative distances in time are changed. Therefore the temporal correlation between the frames would no longer correct. It can be seen from (6), moving the position of frames affects only the second term, $\mathbf{w}_\alpha \mathbf{w}_\beta^T$. Assuming s is time-invariant and independent to w, L_S is the length of s. The estimated temporal correlation of STFT coefficients can be written as the form of complex autocorrelation function in time:

$$A_S(k,l) = L_S \cdot \sum E[\mathbf{s}\mathbf{s}^T] \cdot \mathbf{w}_0 \mathbf{w}_l^T \cdot \mathbf{e}_N^k \overline{\mathbf{e}_N^k}^T \tag{7}$$

The autocorrelation of analysis frames in (7) should be $A_S(k, dl)$ after time stretching. As one could see, even if s has no innate autocorrelation ($E[\mathbf{s}\mathbf{s}^T] = \mathbf{I}$), there still exists autocorrelation between the coefficients of STFT frames at the same bin index, which is induced by $\mathbf{w}_\alpha \mathbf{w}_\beta^T$. Originally, the autocorrelation introduced by the analysis window would be canceled after resynthesis. But since the position of analysis frames are moved, the cancellation fails. Fig.1 shows the difference of autocorrelation function of a Gaussian noise before and after a time stretching of factor 2 with an ordinary phase vocoder. Also, in Fig.1, the autocorrelation is not exactly stretched by twice in time. This is because the phase compensation in phase vocoder cannot fix the autocorrelation of a noise correctly.

To compensate the change of correlation, the coefficients of $S_l(k)$ must be altered to make their autocorrelation approaches $A_S(k, dl)$. The optimization which proposed in [8] is applied. Let $T(l') = A_S(k, dl)$ and the maximum lag is L, we could have:

$$\Phi = \sum_l^L \|T(l) - A_S(k,l)\|^2 \tag{8}$$

$$\frac{\partial\Phi}{\partial s_i(k)} \propto \sum_l^L (T(l) - A_S(k,l))(s_{i-l}(k) + s_{i+l}(k)) \tag{9}$$

Apply the gradient projection mentioned in [8], we could have:

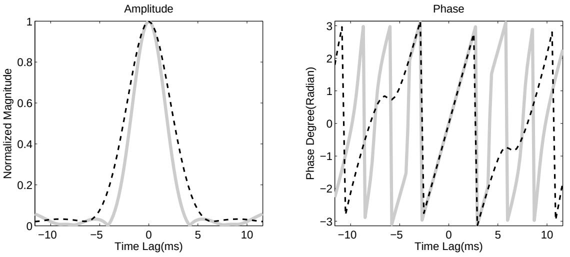

Figure 1: The change of temporal correlation before(thick solid line) and after(thin dotted line) phase vocoder stretching.($k = 64, -11.6 \le timelag \le 11.6$)

$$s'_i(k) = s_i(k) + \sum_{l}^{L} \lambda(l,k)(s)(s_{i-l}(k) + s_{i+l}(k)) \quad (10)$$

$$s'_i(k) = s_i(k) \otimes h(i,k) \quad (11)$$

$h(i,k)$ in (11) is a zero-phase filter function with length L. There are $N/2$ different filters for bin indices between 1 $N/2$, while the other half are the complex conjugates. At last, we use the algorithm in [20] to solve the coefficients of the filters.

4. EXPERIMENT

To verify the result of section 3, an experiment is conducted. In this experiment, we perform the time stretching on a random generated Gaussian noise. Since the marginal statistics and spectral correlation are unaffected by time stretching, only the temporal correlation has to be corrected after the transformation. The first step is the ordinary phase vocoder analysis as (1). Then the following step corrects temporal correlation for each frequency bins respecting to the stretch factor. The resynthesis is done as (3). The estimation of target autocorrelation follows (7), and $E[\mathbf{ss}^T]$ is replaced by $\sigma^2 I$ due to the nature of Gaussian noise. The detailed configuration is listed below:

- original signal : Gaussian noise (1 sec).
- sampling rate : 44.1khz
- stretching factor : 3.0
- hop size of input frame : 16
- size of Fourier transform : 256
- maximum lag of correlation : 64

The error of temporal correlation for each frequency bin before and after correction is shown in Fig.2. In most of the frequency bins, the error is lower than 48dB, through a better correction may still exist. The spectrogram of the Gaussian noise is shown in Fig.3. In the spectrogram of 'stretched w/o correction', the energy of the spectrogram is not uniformly distributed. Both frequency components and the gaps between them tend to persist longer in time, thus making the sound more sinusoidal and resulting perceivable artifacts such as short pitches. After the correction is applied, the intensity of short sinusoids are mitigated, thus reducing the arfefacts. The sound files could be found on [21]. The difference between the sounds is perceivable, but still not perfect. This may due to the temporal correlation was distorted after the resynthesis. A possible reason is that the phase coherence is not considered in the correction of temporal correlation. If so, certain constraint has to be applied to the correction mechanism.

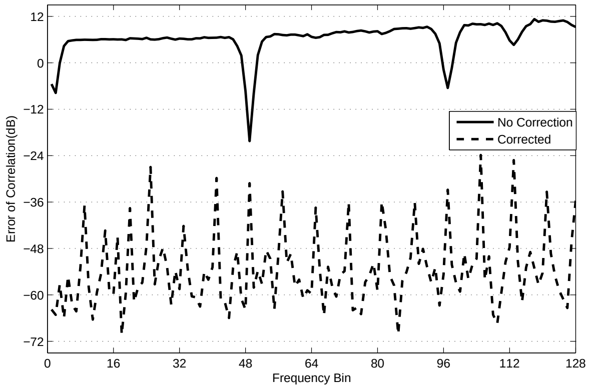

Figure 2: The error of temporal correlation before(solid line) and after(dotted line) the correction.

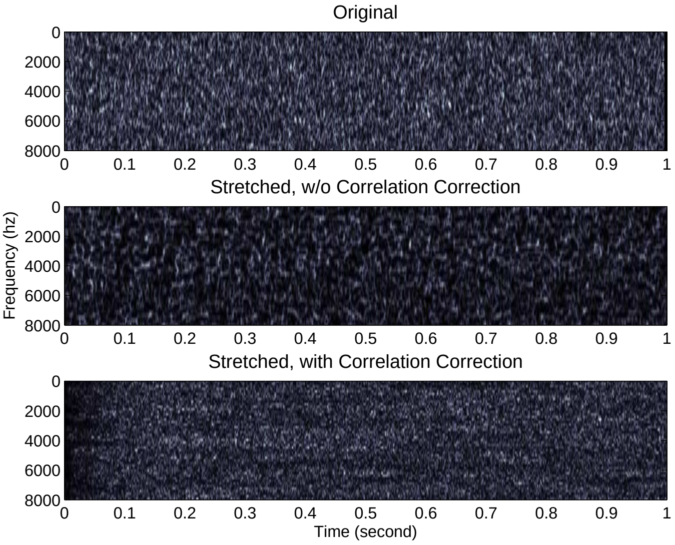

Figure 3: The spectrogram of original signal, stretched w/o correction, stretched with correction

5. CONCLUSION

With a proper parametrized model and proper modification, phase vocoder could be applied to non-sinusoidal signals. In this paper, we investigated the variation of temporal correlation during the time stretching process on a Gaussian noise by phase vocoder. From this simplified case, we found that the variation will always in place regardless the innate correlation of the signal. Also, the variation in the temporal correlation is perceptually significant. Therefore, to achieve a successful time stretching, it is necessary to correct temporal correlation respect to the stretching factor and analysis window. The temporal correlation function can be obtained by estimation under different stretch ratio. Currently, the correction procedure is done iteratively, but it's likely that there exists a non-iterative solution. However, this correction does not include the correction of correlation induced by the signal itself and the correction of correlation is not perfect. The discovery from the streching of Gaussian could be served as the basis of general sound texture transformation. This could lead to various applications like non-uniform stretching, texture morphing and parametric texture synthesizing. The future work includes considering the general sound texture and seek an efficient mechanism to correct statistical properties during the transformation by phase vocoder.

6. REFERENCES

[1] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[2] M.-H. Serra, "Musical signal processing, chapter introducing the phase vocoder," in *Studies on New Music Research. Swets & Zeitlinger*, pp. 31–91, 1997.

[3] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications," *Journal of the AES*, vol. 47, no. 11, pp. 928–936, 1999.

[4] J. Bloit, N. Rasamimanana, and F. Bevilacqua, "Towards morphological sound description using segmental models," in *Proceedings of DAFx*, 2009.

[5] N. Saint-Arnaud and K. Popat, "Computational auditory scene analysis," ch. Analysis and synthesis of sound textures, pp. 293–308, Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1998.

[6] P. Hanna and M. Desainte-catherine, "Time scale modification of noises using a spectral and statistical model," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on*, vol. 6, pp. 181–184, April 2003.

[7] D. Schwarz, *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Ircam - Centre Pompidou, Paris, France, January 2004.

[8] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision*, vol. 40, pp. 49–71, December 2000.

[9] J. Bruna and S. Mallat, "Classification with invariant scattering representations," *CoRR*, vol. abs/1112.1120, 2011.

[10] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, Sep 2011.

[11] A. Roebel, "A new approach to transient processing in the phase vocoder," in *6th International Conference on Digital Audio Effects (DAFx)*, (London, United Kingdom), pp. 344–349, Septembre 2003.

[12] A. Roebel and X. Rodet, "Real time signal transposition with envelope preservation in the phase vocoder," in *International Computer Music Conference*, (Barcelona, Spain), pp. 672–675, Septembre 2005.

[13] A. Roebel, "Shape-invariant speech transformation with the phase vocoder," in *InterSpeech*, (Makuhari, Japan), pp. 2146–2149, Septembre 2010.

[14] B. Julesz, "Visual pattern discrimination," *Information Theory, IRE Transactions on*, vol. 8, pp. 84–92, 1962.

[15] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, Aug. 1990.

[16] J. H. McDermott, A. J. Oxenham, and E. P. Simoncelli, "Sound texture synthesis via filter statistics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-09)*, (New Paltz, NY), pp. 297–300, IEEE Signal Processing Society, Oct 18-21 2009.

[17] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[18] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[19] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms*. Birkhaeuser, 1998.

[20] J. Dennis, L. Vicente, J. E. Dennis, and L. N. Vicente, "Trust-region interior-point algorithms for minimization problems with simple bounds," tech. rep., SIAM J. Control and Optimization, 1995.

[21] "http://anasynth.ircam.fr/home/english/media/whliao-experiment-stretching-gaussian-noise/."