Audio Engineering Society

Convention Paper 7946

Presented at the 127th Convention
2009 October 9–12 New York, NY, USA

Audio Bandwidth Extension using Cluster Weighted Modeling of Spectral Envelopes

Nikolay Lyubimov[1], and Alexey Lukin[2]

Dept. of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia

[1] lubimov.nicolas@gmail.com

[2] lukin@graphics.cs.msu.ru

ABSTRACT

This paper presents a method for blind bandwidth extension of band-limited audio signals. A rough generation of the high-frequency content is performed by nonlinear distortion (waveshaping) applied to the mid-range band of the input signal. The second stage is shaping of the high-frequency spectrum envelope. It is done by a Cluster Weighted Model for MFCC coefficients, trained on full-band width audio material. An objective quality measure is introduced and the results of listening tests are presented.

1. INTRODUCTION

When the audio signal is transmitted and processed, it is often subject to noise and distortion. One widespread type of distortion is reduction of frequency range (bandwidth) of the signal. It can happen during transmission of signals through phone lines, perceptual audio coding or sampling rate conversion of digital signals. Loss of high-frequency signal components is perceived as subjective quality degradation.

Bandwidth extension is the process of re-synthesizing missing frequency components of the signal in order to improve the subjective quality. Bandwidth extension methods are often found in modern perceptual audio (de)coders. Such methods can be blind, when no information about missing signal components is available, and non-blind, when certain information about missing components is available during the synthesis stage.

One popular example of a non-blind bandwidth extension technology is SBR – Spectral Band Replication [1]. This algorithm can be found in mp3pro and AAC+ audio coding formats. The first part of the algorithm is working in the encoder; it saves compressed information about high-frequency signal energy envelope in the bit stream of only few kb/s. The second part of the algorithm is working in the decoder; it synthesizes high-frequency signal by frequency

shifting the mid-frequency signal and then shapes the spectrum of the synthesized signal using the encoded side information from the bit stream.

The problem of blind bandwidth extension is more complex because naturalness of the resulting signal strongly depends on the energy envelope of synthesized high frequencies. Another parameter that is important for natural sound is harmonicity of the synthesized signal. One aspect of harmonicity is preservation of harmonic relationships in the synthesized signal. Another aspect is the relation of tonal and noisy components in the synthesized signal.

Blind methods of bandwidth extension try to predict parameters of the reconstructed high-frequency signal using parameters of the existing mid-frequency signal. The simplest possible way is described in [2]: it assumes that there is a linear slope of energy at high frequencies for every moment of time. The slope is estimated from the mid-frequency material, and the synthesized high-frequency material is shaped to fit this slope. In [3] a low-complexity method is proposed. It uses nonlinear transformation of the input signal that creates high frequency spectral content. A simple algorithm that attempts to maintain the continuity of spectral envelope is introduced. The interesting approach was described in [4], where the bandwidth extension is done by linear predictive extrapolation both in time and frequency domains. More complicated method is proposed in [5] with the focus on speech applications. The vector quantization technique and codebook mapping is used there in order to predict the shape of high-frequency energy envelopes. Another way is to use statistical models to predict envelope features, like Gaussian Mixture Model [6].

This paper presents a new way of more accurate prediction of the high-frequency energy envelope using a Cluster Weighted Model for MFCC coefficients. The general scheme of proposed algorithm is introduced in section 1. In sections 2 and 3 the main steps of algorithm are given in details. The evaluation results are done in section 4, and then the conclusion and future works are presented.

2. ALGORITHM DESCRIPTION

Like in most often used blind bandwidth extension techniques, the method proposed in this paper consists of two consecutive steps:

1. "Rough" generation of high-frequency spectral content

2. Shaping of the energy spectrum envelope of the generated content

which are performed in small frames independently. The general scheme of our algorithm is depicted in figure 1.

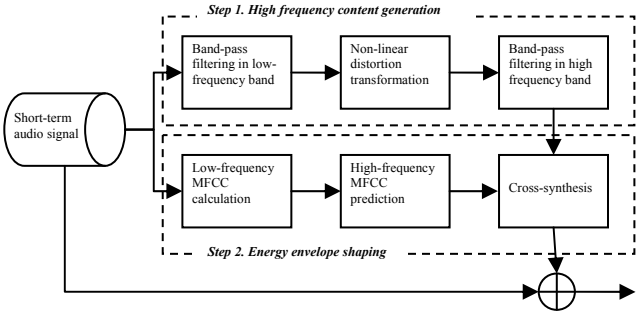

Figure 1 The general scheme of proposed algorithm.

First we split low-frequency audio signal in frames of 25 ms length and 75% overlap. Each frame is weighted by Hamming window in order to allow proper signal reconstruction using overlap-and-add technique. We perform "rough" high frequency content generation in time domain and envelope shaping in frequency domain. High frequency content generation is based on non-linear distortion, or waveshaping method, that has property of spectrum expansion [7]. Then high-frequency energy envelope adjustment stage is started. In this work we utilize MFCC coefficients to parameterize spectral envelopes. In order to predict the parameters of high frequency spectral envelopes we have trained the Cluster Weighted Model using full-bandwidth sound examples. In each time frame of input signal we calculate the MFCC vector of low-frequency audio spectrum. The corresponding high-frequency MFCC vector is generated like a weighted sum of linear functions. Each function describes a cluster that could be interpreted as particular envelope shape. Finally cross-synthesis technique is used and re-synthesized high-frequency content is summed with input low-frequency signal.

2.1. High-frequency content generation

In this work we have tried to find the method of "rough" high frequency content generation that complies with the following requirements:

- It has to be fast and easy to implement
- It should preserve the harmonic relationship at high frequencies
- It should be independent of a signal power level

Full Wave Rectifier method is suitable for these purposes. It can be written in form

$$y(t) = abs(x(t)) \quad (1)$$

where $x(t)$ and $y(t)$ are input and output signal respectively. Before non linear transformation and after this the signal is filtered using band-pass filters [8]. Pre-filtering and post-filtering blocks are required in order to avoid some undesirable effects like intermodulation and aliasing.

2.2. Energy envelope shaping

After "rough" generation stage the envelope shaping stage starts. As possible way of envelope parameterization we utilize well-*known Mel Frequency Cepstral Coefficient* (MFCC) approach. In [9] MFCC coefficients are considered to be appropriate features for musical signals as well as for speech processing purposes.

The process of high-frequency energy envelope prediction is performed by statistical modeling of relationship between low-frequency and high-frequency spectral envelopes.

Unlike GMM-like methods we use more flexible technique called *Cluster Weighted Modeling (CWM),* originally presented in [10]. The flexibility of this technique consists in the fact that we can synthesize high-frequency spectral envelope parameters as weighted sum of any arbitrary functions $f_i(x)$, called *local functions*, where x denotes input low-frequency envelope parameters vector and weights depend on this input:

$$\tilde{y}(x) = \sum_{i=1}^{M} w_i(x) f_i(x) \quad (2)$$

If we will fit the property $\sum_{i=1}^{M} w_i(x) = 1$ for any input x, the equation (3) could be considered as fuzzy logic codebook mapping. Actually each local function with its weight represents a cluster, and can be interpreted as particular high-frequency envelope shape that depends on input low-frequency energy envelope shape. This is illustrated in figure bellow:

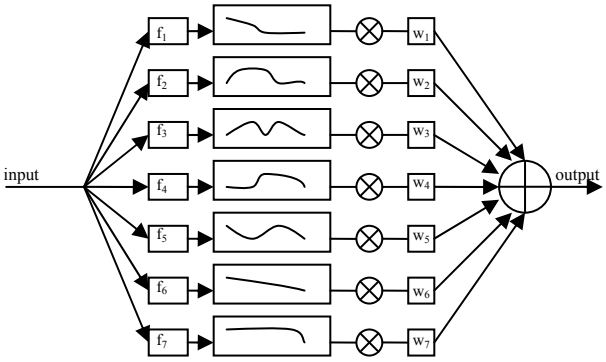

Figure 2 Cluster weighted modeling for high frequency envelope prediction using 7 atomic envelopes shapes

This allows avoiding the problem of dictionary widening in normal codebook mapping. Using this method we can have only a few number of possible low-frequency *atomic shapes* in order to create the infinite number of output shapes. Otherwise, we fit the parameters of weighting functions $w_i(x)$ and local functions $f_i(x)$ using maximum likelihood criteria on chosen dataset of envelope features. Thereby CWM includes vector codebook mapping as well as statistical model properties.

In order to simplify the calculations on training stage we have used local functions of linear form:

$$f_i(x) = \xi_i + A_i(x - \mu_i) \quad (3)$$

and weighting functions in form

$$w_i(x) = \frac{\alpha_i g_i(x)}{\sum_{m=1}^{M} \alpha_m g_m(x)} \quad (4)$$

where $g_i(x)$ is a normal probability density function with expectation value μ_i that represents i-th atomic low-frequency envelope shape, and covariance matrix Σ_i, and α_i is a scalar weight of each mixture component.

In this case, there are explicit formulas to compute CWM parameters which are derived from Expectation-Maximization formalism [11].

Using the predicted MFCC coefficient for high-frequency spectral envelope, we can modify the input envelope in cepstral domain due to linearity of cosine transform. First we calculate the scaling cepstral coefficient using equation:

$$s_n = c_n^{target} - c_n^{source} \quad (5)$$

Here c_n^{target} denotes target envelope MFCC representation that is predicted by CWM model, and c_n^{source} is representation of actual (source) envelope that is presented in spectrum after non linear distortion.

In order to get the full envelope from MFCC representation, we use the linear interpolation between inverse DCT coefficients of obtained parameters:

$$S(\omega) = \frac{\omega_{k+1} - \omega}{\omega_{k+1} - \omega_k} b_k + \frac{\omega - \omega_k}{\omega_{k+1} - \omega_k} b_{k+1}, \quad \omega_k \le \omega \le \omega_{k+1} \quad (6)$$

where ω_k is k-th band on mel-spaced frequency grid,

$$b_k = \exp\left\{ \frac{s_0}{N} + \frac{2}{N} \sum_{n=1}^{N-1} s_n \cos\left(\frac{\pi n (k + 0.5)}{N}\right) \right\} \quad (7)$$

We use this envelope as the *scaling function* for modifying the presented envelope above cutoff frequency. The final full-bandwidth spectrum is obtained via formula:

$$Y(\omega) = \begin{cases} X(\omega), & \omega \le \omega_{cutoff} \\ S(\omega) X_{dist}(\omega), & \omega > \omega_{cutoff} \end{cases} \quad (8)$$

where $X(\omega)$ denotes complex spectrum points of input band-limited signal, and $X_{dist}(\omega)$ is a spectrum of nonlinear distorted signal.

3. EVALUATION AND RESULTS

The goals of evaluation stage were:

- Find the best audio material for statistical training of high-frequency envelope prediction model
- Explore the coherence between objective (automatic error calculation) and subjective (evaluation on listening test) quality assessments
- Compare proposed method with the methods
  - that uses envelope parameterization without psychoacoustic model instead of MFCC
  - that uses blind bandwidth extension method without envelope shaping stage (similar to that is presented in [8])

For our statistical model training, we have collected 3 types of audio material: monophonic (single-source) audio signals, polyphonic (multiple-source) signals from RWC Music Genre database and some speech samples. Using 44 KHz 16-bit audio format, we have chosen various audio fragments of 5 minutes total duration for each type. Our goal was to examine the ability of each model to predict the high-frequency envelopes for different audio type. The cutoff frequency has been set on 5 KHz, and the maximal Euclidian distance between predicted and actual envelope in range 5 – 18 KHz has been calculated:

$$error = \max_t \left\| y_p(t) - y_a(t) \right\| \quad (9)$$

where $y_p(t)$ and $y_a(t)$ are predicted and actual MFCC representations of high-frequency spectral envelopes at time t, and $\|\square\|$ denotes Euclidian distance.

Here the evaluation results are presented. The numbers in upper rows indicate the number of clusters used in specified model.

	3	7	10	16	32	64
Single instrument 1 (French horn)	42	37.39	34.22	35.81	52.09	46.39
Single instrument 2 (Guitar note)	27.17	33.46	47.55	32.87	39.33	43.27
Woman speaking	78.46	68.21	72.30	72.54	78.13	83.94
Man speaking	41.97	55.94	44.70	40.28	53.46	53.85
Polyphonic music 1 (classic)	62.20	78.94	72.08	68.73	83.42	94.07
Polyphonic music 2 (hard rock)	35.19	38.65	40.34	37.10	35.30	40.34
Polyphonic music 3 (pop)	67.19	66.42	62.37	61.72	49.87	60.09

Table 1 Polyphonic music model errors

	3	7	10	16	32	64
Single instrument 1 (French horn)	27.88	33.1	27.46	25.31	27.38	35.50
Single instrument 2 (Guitar note)	35.67	58.13	50.47	45.66	66.96	64.14
Woman speaking	76.37	96.68	112.71	129.87	122.56	107.82
Man speaking	67.68	62.57	75.89	93.95	76.31	72.72
Polyphonic music 1 (classic)	77.47	98.25	105.36	109.64	92.57	101.74
Polyphonic music 2 (hard rock)	44.05	64.23	74.45	91.39	72.42	64.04
Polyphonic music 3 (pop)	78.94	67.32	81.66	103.50	73.01	69.67

Table 2 Single instrument model errors

	3	7	10	16	32	64
Single instrument 1 (French horn)	38.20	48.00	53.06	66.49	78.41	60.88
Single instrument 2 (Guitar note)	73.81	62.41	61.71	74.15	66.51	61.11
Woman speaking	77.21	84.70	87.82	93.20	98.19	81.82
Man speaking	52.31	51.03	49.22	50.00	46.70	48.15
Polyphonic music 1 (classic)	95.50	95.65	99.58	92.91	88.77	102.43
Polyphonic music 2 (hard rock)	68.39	65.29	58.17	66.30	64.14	42.91
Polyphonic music 3 (pop)	59.36	59.69	64.94	64.34	71.52	44.28

Table 3 Speech model errors

If we examine the best results (minimum errors) and the worst results (maximum errors) in each row for each table, we can derive the table bellow that describes the means of best and worst error for each type of model:

	polyphonic	**single instrument**	**speech**
best	45.30	55.53	57.02
worst	55.03	90.12	77.91

Table 4 The mean values of minimum ("best") and maximum ("worst") errors depicted in tables 1-3

This simplifies the view of the tables above. Apparently, the model constructed on polyphonic music has the minimal mean error both for the best and the worst results unlike the other models. This fact is well explicated, because in polyphonic music case vocal lines and single instrument passages are also included.

In order to show the coherence between chosen objective metric and perception of extended sounds, we perform the informal listening test. The listeners compared the original full bandwidth sound ("the reference") with the filtered sound ("the anchor") and some kinds of restored sound with different type of models. The models have been chosen corresponding to the minimum (MIN) and maximum (MAX) value in each table. Three audio samples (pop music, French horn and speech) have been presented to the listeners. The values in tables are in range [0, 100] where 0 signifies perceptual dissimilarity between timbre of reference signal and obtained examples, and 100 means that the timbre of these audio samples is similar. The evaluation results are depicted in table below.

	Polyphonic (MIN)	Polyphonic (MAX)	Single instrument (MIN)	Single instrument (MAX)	Speech (MIN)	Speech (MAX)
French horn	59.99	41.40	95.97	91.22	36.79	19.90
Pop-music	14.45	10.32	3.91	5.15	6.71	7.37
Speech	69.55	71.03	45.82	40.44	75.87	70.62

Table 5 Evaluation results on perceptual listener test using different kind of models

Comparing the last table with the tables 1-3, the correlation between objective error metric and subjective evaluations is observed. Actually if we look at results obtained from monophonic model of French horn applied to another French horn audio sample (first row of table 2) we will see the best envelope predictions among all results. Listener assessments confirm this result. The low assessments from pop-music audio samples occur due to undesirable intermodulation distortion effect.

Finally we have tested our method with other bandwidth extension algorithms. Through the similar listening test we have obtained the following assessments for

- algorithm proposed in this paper (CWM+MFCC)
- algorithm that uses cluster weighted modeling of sub-band energy without psychoacoustic model (CWM+SBE)

- algorithm that doesn't use envelope shaping stage, that is similar to method presented in [8] (NLD)

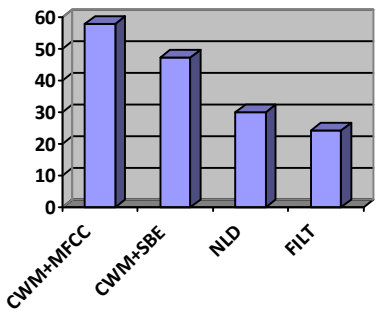

Figure 3 Comparative assessments with different bandwidth extension algorithms obtained via listening test

As depicted in this figure, the proposed method shows better performance on listening test with respect to other methods. The last bar (FILT) shows the comparative assessment of non-processed low frequency audio samples with 5 KHz cutoff frequency. It shows that after bandwidth extension algorithm the subjective quality of audio samples improves significantly.

4. CONCLUSION AND FUTURE WORK

In this paper we have presented the new algorithm of audio bandwidth extension. Unlike widely used Spectral Band Replication (SBR) technology, our algorithm creates high frequency content of short term spectrum without any additional information. The non-linear distortion method produces first the "rough" approximation of high frequency spectral content. After that the HF energy envelope is predicted using Cluster Weighted Model for MFCC coefficients, and envelope is shaped corresponding to predicted parameters.

We have examined our method using three different evaluation stages:

1. by choosing the best audio material using in machine learning model construction for HF envelope prediction

2. by exploring the coherence between objective and subjective evaluations for different types of constructed models

3. by comparing developed algorithm with another technique, like one presented in [8]

We have shown the best generalization properties for the model trained on polyphonic pop music audio material from RWC Music Genre database compared with single-instrument and speech audio material. Furthermore we have shown coherence between objective quality metric and subjective evaluations. Objective quality metric has been introduced as the maximal Euclidian distance between predicted and actual envelopes in MFCC feature space. Subjective evaluations have been done through informal listening test. Finally on listening test of our algorithm we have shown better performance compared with the algorithms with sub-band energy envelope parameterization instead of MFCC, and algorithm without envelope shaping at all.

The main drawback of the proposed method is that the quality of results degrades at low cutoff frequencies on polyphonic signals: the generated high-frequency extension gets noise-like character. This probably happens because of the intermodulation effect produced by a non-linear transformation of polyphonic audio signals. In further work we will try to avoid this problem by preliminary separation of audio sources.

5. REFERENCES

[1] M. Dietz, L. Liljeryd, K. Kjorling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," presented at the 112th AES convention, Munich, Germany, 2002, May 10-13.

[2] Chi-Min Liu, Wen-Chieh Lee, and Han-Wen Hsu "High Frequency Reconstruction for Band-Limited Audio Signals", Proc. of the 6th Int. Conference on DAFX, London, UK, September 8-11, 2003

[3] Manish Arora, Joonhyun Lee, and Sangil Park "High Quality Blind Bandwidth Extension of Audio for Portable Player Applications", presented at the 120th AES convention, Paris, France, 2006 May 20-23

[4] Chatree Budsabathon, Akinori Nishihara "Bandwidth Extension with Hybrid Signal Extrapolation for Audio Coding", IEICE Trans. Fundamentals, Vol. E90-A, No. 8, August 2007

[5] N. Enbom and W.B.Klein "Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients", in IEEE Workshop on Speech Coding, Porvoo, Finland, pp. 171-173, 1999

[6] M. Nilsson, H. Gustafsson, S.V. Andersen, and W.B. Kleijn. "Gaussian mixture model based mutual information estimation between frequency bands in speech." Acoustics, Speech, and Signal Processing, 2002. Proceedings (ICASSP '02). IEEE International Conference on, 1 :525-528, 2002.

[7] Marc Le Brun "Digital Waveshaping Synthesis", Journal of the Audio Engineering Society, 27(4), 1979, pp. 250-266.

[8] Eric Larsen, Ronald Aarts, Michael Danessis "Efficient High-Frequency Bandwidth Extension of Music and Speech", presented at the 112th AES convention, Munich, 2002

[9] Beth Logan "Mel-Frequency Cepstral Coefficients for Music Modeling", In Int. Symp. on Music Information Retrieval, 2000

[10] Neil Gershenfeld, "Cluster-Weighted Modeling: Probabilistic Time Series Prediction, Characterization and Synthesis", Nature of Mathematical Modeling, MIT Press, 1998, Ch. 15, pp. 365-386

[11] Danil V. Prokhorov, L.A. Feldkamp, T.M. Feldkamp "A New Approach to Cluster-Weighted Modeling", Int. Joint Conference on Neural Networks (IJCNN) 2001 Proceedings IEEE, Vol.3, pp. 1669-1674