BLIND BANDWIDTH EXTENSION USING K-MEANS AND SUPPORT VECTOR REGRESSION

Chih-Wei Wu[1]* *and Mark Vinton*[2]

[1] Center for Music Technology, Georgia Institute of Technology, Atlanta, GA, 30318
[2] Dolby Laboratories, San Francisco, CA, 94103

ABSTRACT

In this paper, a blind bandwidth extension algorithm for music signals has been proposed. This method applies the K-means algorithm to firstly cluster audio data in the feature space, and constructs multiple envelope predictors for each cluster accordingly using Support Vector Regression (SVR). A set of well-established audio features for Music Information Retrieval (MIR) has been used to characterize the audio content. The resulting system is applied to a variety of music signals without any side information provided. The subjective listening test results show that this method can improve the perceptual quality successfully, but the minor artifacts still leave room for future improvements.

Index Terms— Bandwidth extension, K-means, Support Vector Regression

1. INTRODUCTION

With the increasing popularity of mobile devices (i.e., smartphones, tablets) and online music streaming services (i.e., Apple Music, Pandora, Spotify...etc), the capability of providing high quality audio content with minimum data requirement becomes more important. To ensure a fluent user experience, the audio content could be heavily compressed and lose its high frequency (HF) information during the transmission. This compression process may cause degradation to the perceptual quality of the content. An audio Bandwidth Extension (BWE) method can be used to address this problem and restore the HF information to improve the perceptual quality [1]. In general, audio bandwidth extension can be categorized into two types of approaches: 1) *Non-blind* 2) *Blind*.

In the first type of approaches (*Non-blind*), the signal is reconstructed at the decoder with side information provided. This type of approach can generate high quality results since more information is available. However, it also increases the data requirement and might not be applicable in some use cases. The most well-known method in this category is Spectral Band Replication (SBR) [2, 3]. SBR is a technique that has been used in the existing audio codecs such as MPEG-4 High-Efficiency Advanced Audio Coding (HE-AAC). It can improve the efficiency of the audio coder at low-bit rate by encapsulating the HF content and recreating it based on the transmitted low frequency (LF) signal with side information. Being a simple and efficient algorithm, SBR still introduces some artifacts to the signals [4]. One of the most obvious issues is the mismatch in the harmonic structures caused by the process of the band replication to create the missing HF content. To improve the patching algorithm, a sinusoidal modeling based method was proposed to generate the missing tonal components in SBR [5]. Another approach is to use a phase vocoder to create the HF content by pitch shifting the LF part [6]. The other approaches, such as offset adjustment between the replicated spectrum [7] or a better inverse filtering process [8], have also been proposed to improve the patching algorithm in SBR.

In the second approach (*Blind*), the signal is reconstructed at the decoder without availability of side information. This type of approach mainly focuses on general improvement instead of faithful reconstruction. One approach is to use a wave-rectifier to generate the HF content, and use different filters to shape the resulting spectrum [9]. This approach has a lower model complexity and does not require a training process. However, the filter design becomes crucial and could be difficult to optimize. The other approaches, such as linear predictive extrapolation [10] and chaotic prediction theory [11], also predict the missing values without any training process. Recently, machine learning based approaches gain more popularity. For example, envelope estimation using Gaussian Mixture Model (GMM) [12], Hidden Markov Model (HMM) [13] and Neural Network [14] has been used. These approaches generally work well when the training data is sufficient, but the model complexity could be higher than traditional methods.

For methods focusing on blind BWE of speech signals, Linear Prediction Coefficients (LPC) is commonly used to extract the spectral envelope and excitation from the speech. A codebook can then be used to map the envelope or excitation from narrowband to wideband [15]. Other approaches, such as linear mapping [16], GMM [17] and HMM [18], have been proposed to predict the wide-band spectral envelopes. Combining the extended envelope and excitation, the band-

[*] The first author performed the work while at Dolby Laboratories

width extended speech can be re-synthesized at the decoder. However, comparing with speech signals, music has a more complicated excitation signal and spectral shape. Therefore, an LPC based method might not be directly applicable.

In this paper, we focus on blind BWE methods for music signals. More specifically, we propose a method to extend the bandwidth of a given music signal from 7 kHz to 22.05 kHz . In the field of MIR, it is shown that audio features are useful for characterizing the audio content [19]. Inspired by the audio content analysis approaches, we propose to apply an unsupervised clustering algorithm followed by a machine learning based approach to build HF envelope predictors for signals with similar characteristics. The rest of the paper is structured as follows: In Sec. 2, the algorithmic details of the proposed method are described. In Sec. 3, the datasets, metrics, and results from a listening test are discussed. Finally, the conclusions and future directions of are presented in Sec. 4.

2. METHOD

2.1. Algorithm Description

The flowchart of the proposed method is shown in Fig. 1. It consists of two phases: training and testing. In the training phase, the audio signals are firstly converted into time-frequency representations using Complex Quadrature Mirror Filter (CQMF) transformation as specified in [2]. The CQMF filter-bank decomposes the signal into 64 complex valued sub-bands using blocks of 64 samples. Next, the spectral envelopes of each block are extracted and separated into HF and LF parts with a cutoff frequency of 7 kHz . A set of commonly used audio features are extracted from the LF signals, and these features are further clustered using K-means algorithm. For each cluster, a set of M HF envelope predictors are trained using Support Vector Regression (SVR) with the audio features and the actual HF spectral envelopes as targets; M equals to the number of coefficients representing the HF spectral envelopes. Finally, the resulting K by M envelope predictors and K centroids are stored and sent to the decoder.

In the testing phase, the audio signals are converted into time-frequency representations with the same CQMF transformation. The LF part of the signals (cutoff frequency = 7 kHz) are then separated, followed by a similar feature extraction process as in the training phase. For each block, the best set of envelope predictors is selected by calculating the distances between the current feature vector and the K centroids. These predictors are used to generate the predicted HF spectral envelopes. The HF complex CQMF coefficients are created by replicating the values from LF part and adjusting the spectral shape to match the predicted HF spectral envelopes. Finally, the resulting CQMF representation, which combines the original LF part and the generated HF part, is converted back to the time-domain using an inverse CQMF transformation.

2.1.1. K-means algorithm

The basic assumption of the proposed method is that audio signals with similar characteristics (such as genre) could be more likely to have similar spectral shapes. To explore the underlying simliarity of the audio content, one of the most popular unsupervised clustering algorithm, *K-means* [20], is used. The algorithm can be summarized as follows:

1 Initialize K centroids by randomly selecting K samples from the data pool.

2 Classify every sample with a class label of 1 to K based on their distances to the K centroids.

3 Compute the new K centroids by taking the average of each class.

4 Update the centroids

5 Repeat step 2 to 4 until convergence.

In a preliminary experiment of the proposed method, $K = 20$ to 40 was tested, and $K = 20$ was selected for achieving the best result in terms of the objective measurement (see Sec. 3.2). The maximum iteration is set to 500. However, the algorithm usually converges after 200 to 300 iterations. Finally, the distance measure used in our K-means implementation is Euclidean distance.

2.1.2. Support Vector Regression (SVR)

Support Vector Machine (SVM) [21] is one of the state of the art machine learning algorithms that has been proven successful for various classification tasks, and Support Vector Regression (SVR) is the variant of SVM for regression tasks. In general, SVM is a linear classifier that defines an optimal hyperplane to separate the data in the feature space, and the optimization problem is solved by finding the support vectors that can maximize the margins nearby the decision boundary. Comparing with the other classification and regression algorithms, SVM has the flexibility of defining the tolerance of error within the margins, leading toward a more generic solution. For implementation, a MATLAB version of the SVM library *LIBSVM* [22] is used.

In this paper, the basic idea is to predict the HF spectral shape based on the audio features extracted from the LF signal. Since the predicting values are continuous, a regression version of the SVM (nu-SVR) is used as the predictor. To introduce non-linearity into the model, a Radial Basis Function (RBF) kernel is used. The rest of the parameters follow the default settings in *LIBSVM*.

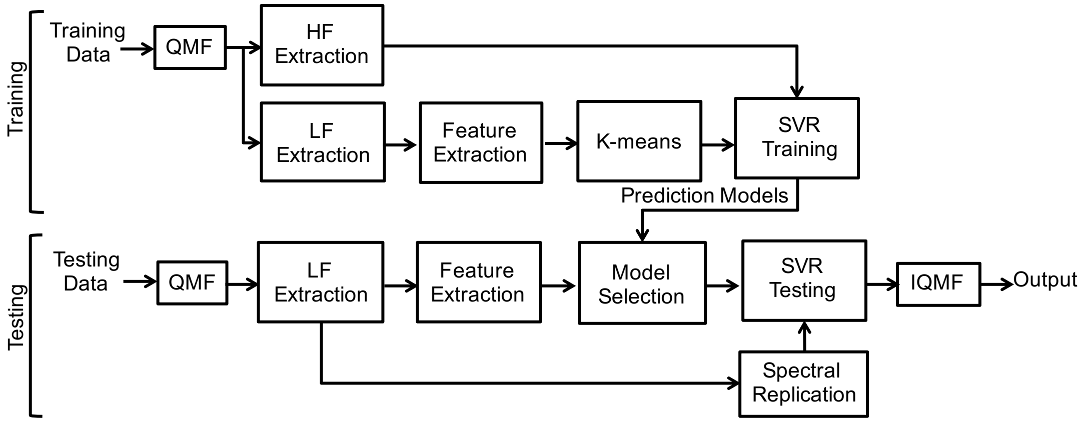

Fig. 1. Overview of the proposed blind bandwidth extension method

Table 1. List of the extracted audio features

Domain	Name	Dimensionality
Spectral	Centroid	1
Spectral	Flatness	1
Spectral	Skewness	1
Spectral	Spread	1
Spectral	Flux	1
Spectral	MFCC	13
Spectral	Tonal Power Ratio	1
Temporal	RMS	1
Temporal	Zero Crossing Rate	1
Temporal	ACF	10

2.2. Feature Extraction

The features used in this paper are listed in Table 1. These features are commonly used in audio content analysis. More implementation details of the selected features can be found in [23] and [1]. In this paper, the spectral envelopes are calculated by taking the absolute value of the complex QMF coefficients. The spectral features, as listed in Table 1, are computed from the spectral envelopes of the LF part of the input signal, and the temporal features are computed from the waveform of the same LF signal with non-overlapping blocks. The block size for calculating the temporal features is chosen to synchronize with the block size of the CQMF decomposition. Finally, the features are normalized using a standard z-score normalization process.

3. EXPERIMENTS

3.1. Datasets

Two datasets are used for training and testing purposes in this paper. The training set is a large collection of stereo signals with a variety of contents such as music, instrumental sounds, and singing voices. The entire folder contains 791 wav files. The length of the recordings varies from 30 seconds to 42 minutes, however, most of the tracks are within the range of 1 to 6 minutes. The testing set is a small collection of stereo signals, which includes 35 songs of different genres such as Classical, Pop, Jazz, Country and Rock. This collection is suitable for testing the system for its diversity. The length of each song is approximately 1 to 6 minutes.

As a pre-processing step, all of the audio tracks are downmixed to mono and resampled to a sampling rate of 44.1 kHz. To fasten the training process, only a short excerpt of 10 seconds from each track is used.

3.2. Metrics

The objective measurement used in this paper is the average spectral distortion as described in [16]. The equation is shown in Equation 1, in which S is the target spectral envelope (in dB), $\hat{S}$ is the predicted spectral envelope (in dB), N is the total number of blocks and W is the total number of frequency bins. The spectral envelopes are calculated as discribed in Sect. 2.2. In general, a lower spectral distortion D implies a higher similarity between the predicted and the actual spectral envelopes. This metric provides a reasonable quantitative measurement of the quality of the resulting signal. However, it is sensitive to small fluctuation and might not necessarily reflect the perceptual quality.

$$D = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{f=1}^{W}\frac{(S^(f,n) - \hat{S}^(f,n))^2}{W}\right)} \qquad (1)$$

3.3. Listening Test

A MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test was conducted to subjectively evaluate the proposed method. 10 songs from the testing set have been chosen to create 10 sets of stimuli. Each set contains 4 different versions of the 20 second excerpt of a song. The first version is the processed audio file using the proposed method.

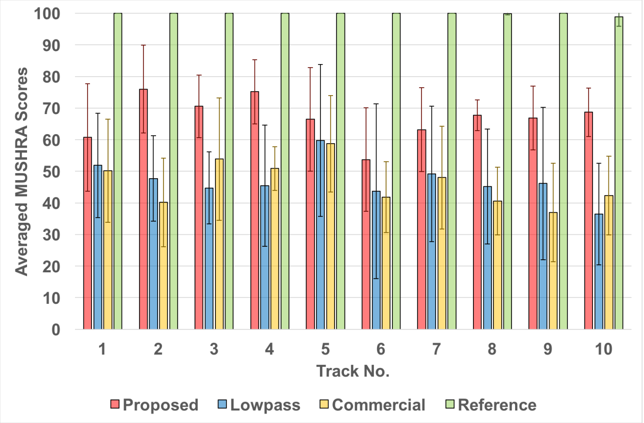

Fig. 2. Results of the MUSHRA test

Table 2. Averaged spectral distortion of the selected tracks

Track No.	1	2	3	4	5
D (dB)	5.53	5.89	10.01	7.26	6.02
Track No.	6	7	8	9	10
D (dB)	9.94	12.98	6.33	6.24	10.36

The second version is the anchor file, which is the low-passed audio file with a cutoff frequency equal to 7 kHz. The third version is processed audio files using a commercially available blind bandwidth extension system. The fourth version is the hidden reference, which is identical as the original input file with a bandwidth equal to 22.05 kHz. There were 7 subjects that participated in the listening test under the same configuration of a controlled listening environment. The subjects were instructed to grade the perceptual quality of the audio files with a scale between 0 and 100. A higher score indicates a higher perceptual quality. The results of the listening test and objective measurement are shown in Fig. 2 and Table 2.

3.4. Results and Discussions

From the results of the listening test, it can be observed that the proposed method has the highest mean scores on all tracks compared with the other versions. This result shows that the proposed method can successfully improve the perceptual quality of the low-passed signal. The objective measurement of the selected tracks is not highly correlated with the listening test scores. However, for certain items, it still reflects the trend of the perceptual quality. For example, the proposed method on track No. 2 and No. 6 has the highest and lowest mean score respectively, and their corresponding averaged spectral distortion are 5.89 dB and 9.94 dB.

In general, the tracks featuring strong human voices, such as track No. 1 and 5, have larger standard deviations on the scores of the proposed method, whereas the tracks focusing on strong background music, such as track No. 8 and 10, have smaller standard deviations. The reason could be that the artifacts in the first group of tracks are more noticeable, while in the second group they are more subtle. These artifacts might be caused by the mismatch in the harmonic structure after the spectral replication. Additionally, since the training set contains more music contents than singing voices, the envelope predictors might not be well-trained for the singing voices and could generate poor estimations.

Track No. 2 and 4 have the largest margins between the mean scores of the proposed method and the low-passed one. Both of these tracks feature strong instrumental sounds with almost no human voices. This could imply that the artifacts introduced by the proposed method are less pronounced on instrumental sounds. However, a more specific testing set is needed to verify this observation.

In certain tracks, a strong clicking sounds can be observed. The cause of the artifacts might be the non-overlapping blocks used in the system, which may create discontinuity and introduce fluctuations to the predicted envelopes.

4. CONCLUSION

In this paper, an audio content analysis inspired blind BWE method has been proposed. Based on the extracted audio features, the proposed method applies the unsupervised clustering technique to group the training data in the feature space, and trains different models separately to better predict the unknown spectral envelopes. The evaluation results show that the proposed method can improve the perceptual quality of the low-passed music signals successfully, and it is especially effective for instrumental sounds.

The future directions are: first, there are some existing artifacts reported by the subjects after the listening test, such as clicking, high pitch spikes and short distortions. Since these artifacts are most likely to be caused by transients, a signal adaptive method based on a transient detection algorithm could be developed to address these issues. Additionally, a signal adaptive noise blending process could be implemented to potentially improve the perceptual quality by masking the artifacts. Second, a larger training set with more emphasis on singing voices could be beneficial to train a better model for improving the quality of singing voices. Last but not least, a better patching algorithm can significantly reduce the artifacts by generating a smoother artificial HF content. A signal adaptive method that switches between simple replication and harmonic extension might provide a more flexible scheme to process different types of music signals.

5. REFERENCES

[1] Erik Larsen and Ronald M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Pro-*

cessing and Loudspeaker Design, John Wiley & Sons, 2004.

[2] Per Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA)*, Leuven, Belgium, 2002.

[3] Martin Dietz, Lars Liljeryd, Kristofer Kjörling, and Oliver Kunz, "Spectral Band Replication, a novel approach in audio coding," in *Proc. of the Audio Engineering Society Convention (AES)*, 2002.

[4] Chi-Min Liu, Han-Wen Hsu, and Wen-Chieh Lee, "Compression artifacts in perceptual audio coding," *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[5] Tomasz Zernicki and Marek Domanski, "Improved coding of tonal components in MPEG-4 AAC with SBR," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, 2008.

[6] Frederik Nagel and Sascha Disch, "A harmonic bandwidth extension method for audio codecs," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[7] Frederik Nagel, Sascha Disch, and Stephan Wilde, "A continuous modulated single sideband bandwidth extension," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[8] Han-Wen Hsu and Chi-Min Liu, "Decimation-whitening filter in spectral band replication," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, 2011.

[9] Manish Arora, Joonhyun Lee, and Sangil Park, "High Quality Blind Bandwidth Extension of Audio for Portable Player Applications," in *Proc. of the Audio Engineering Society Convention (AES)*, Paris, France, 2006.

[10] Chatree Budsabathon and Akinori Nishihara, "Bandwidth extension with hybrid signal extrapolation for audio coding," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 90, no. 8, pp. 1564–1569, 2007.

[11] Yong-tao Sha, Chang-chun Bao, Mao-Shen Jia, and Xin Liu, "High frequency reconstruction of audio signal based on chaotic prediction theory," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 381–384.

[12] Xin Liu, Chang-chun Bao, Mao-shen Jia, and Yong-tao Sha, "A harmonic bandwidth extension based on Gaussian mixture model," in *Proc. of the IEEE International Conference on Signal Processing (ICSP)*, 2010, pp. 474–477.

[13] Xin Liu and Chang-Chun Bao, "Blind bandwidth extension of audio signals based on non-linear prediction and hidden Markov model," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[14] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[15] Jonggeun Jeon, Yaxing Li, Sangwon Kang, Kihyun Choo, Eunmi Oh, and Hosang Sung, "Robust artificial bandwidth extension technique using enhanced parameter estimation," in *Proc. of the Audio Engineering Society Convention (AES)*, Los Angeles, USA, 2014.

[16] Yoshihisa Nakatoh, Mineo Tsushima, and Takeshi Norimatsu, "Generation of broadband speech from narrowband speech based on linear mapping," *Electronics and Communications in Japan, Part II: Electronics (English translation of Denshi Tsushin Gakkai Ronbunshi)*, vol. 85, no. 8, pp. 44–53, 2002.

[17] Kun-Youl Park and Hyung Soon Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.

[18] Peter Jax and Peter Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[19] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[20] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition*, Academic Press, 4 edition, 2009.

[21] Vladimir Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[23] Alexander Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, John Wiley & Sons, 2012.