DESIGNING A UNIFIED SPEECH/AUDIO CODEC BY ADOPTING A SINGLE CHANNEL HARMONIC SOURCE SEPARATION MODULE

Sang-Wook Shin, Chang-Heon Lee, Hyen-O Oh and Hong-Goo Kang*

School of Electrical and Electronic Engineering, Yonsei University, Korea
**Digital TV Lab. LG Electronics Inc., Korea*
E-mail: {swshin, leech, hgkang}@dsp.yonsei.ac.kr

ABSTRACT

This paper proposes a unified speech/audio codec by adopting a single channel harmonic separation module as a pre-processor. A modulation frequency analysis method is used for harmonic separation, and the separated components are first encoded by an appropriate codec, e.g. speech codec. The error between input and the encoded signal is re-encoded by another codec. Though any type of codec can be used for the purpose, we adopt two state-of-the-art international standards (AMR-WB and HE-AAC) to provide an interoperability option. The amount of allocated bits to each stage is controlled by a power ratio of separated harmonic components to input signal. Subjective listening tests verify the consistency of the proposed method in speech, music and mixed signal inputs.

Index Terms—Modulation frequency analysis, AMR-WB, HE-AAC, hybrid coding,

1. INTRODUCTION

The objectives of speech and audio coding are same in a sense that both of them reduce the amount of signal information to efficiently utilize communication bandwidths or to minimize memory usage while keeping perceptual quality as high as possible [1]. However, key paradigms on detailed compression methods are quite different because of inherent constraints on specified application areas. The purpose of speech coding technology is communication between two parties, but that of audio coding is one-way streaming service for entertainment [2][3]. In general, delay and complexity requirements as well as limit of transmission bandwidth for speech coding are stricter than audio coding. The differences on application areas also result in the variation of core paradigm coding technologies. How to model human's voice production system is the core of speech coding, but how to perceive the sound quality is the one of audio coding [2][3].

Most speech coding standards assume that voice can be synthesized by passing periodic pulse-like or random excitation signals through a vocal tract modeling system. Therefore, encoding parameters consist of excitation signals, filter coefficients to model vocal tract, and signal gains. AMR-WB codec, operating at a 16 kHz sampling, divides frequency bands by two parts, 50-6400 Hz and 6400-7000 Hz, to decrease complexity and it distributes the bits into the subjectively more important frequency range. The lower frequency band is coded using an ACELP algorithm while the higher frequency band is artificially synthesized in the decoder using the parameters from the lower band encoder and a random excitation [4].

Since audio coding standards focus on deceiving human auditory systems, how to quantize transformed coefficients is the key component. The high efficiency AAC (HE-AAC) which is a technology of combining spectral band replication (SBR) techniques with advanced audio coding (AAC) has been standardized by moving picture experts group (MPEG-4) [5].

Standardization process is handled by different institutions: moving picture experts group (MPEG) for audio coding and international telecommunication union – telecommunication sector (ITU-T) or wireless community for speech coding [3][6]. As the trends of convergence in communication and broadcasting systems becomes popular, however, needs to efficiently encode both speech and audio signals with one unified codec are increasing. However it is really difficult to design a unified coding method which can efficiently encode both types of signals at a low bit rate because traditional speech coding methods are designed for speech signal but not for audio signal and vice versa. A simple but powerful method is applying a different coding method by estimating the segmental characteristics of input signal [7]. Though it is a simple approach, it causes severe degradation if classification accuracy is low or in transition region where speech- and music-like characteristics exist simultaneously.

This paper proposes an efficient coding method for both speech and audio signals. We first separate harmonic components from the input signal and encode it by a speech coding approach. The remaining part or error between input and the encoded signal is re-encoded by audio coding approach. Therefore, key components of the proposed codec are a separation module and how to efficiently distribute the number of bits to encoding modules.

2. HARMONIC SOURCE SEPARATION BASED ON MODULATION FREQUENCY ANALYSIS

A filterbank, followed by subband envelope detection and frequency analysis of the subband envelopes is the framework of modulation frequency analysis [8]. The filterbank is implemented using the short-time Fourier transform (STFT). Subband envelope is detected using the spectral magnitude of the Fourier transform [8]. For a discrete signal $x(n)$, the STFT can be expressed as

$$X_k(k) = \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn}, \quad for\ k = 0,...,K-1, \tag{1}$$

and the envelope detection and modulation frequency analysis as

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(lL-m)\left|X_k(m)\right|W_I^{im}, \quad for\ i = 0,...,I-1, \tag{2}$$

where $W_k = e^{-j(2\pi/K)}$. $h(n)$ and $g(n)$ are the acoustic and modulation frequency analysis windows with shift length M and L, respectively.

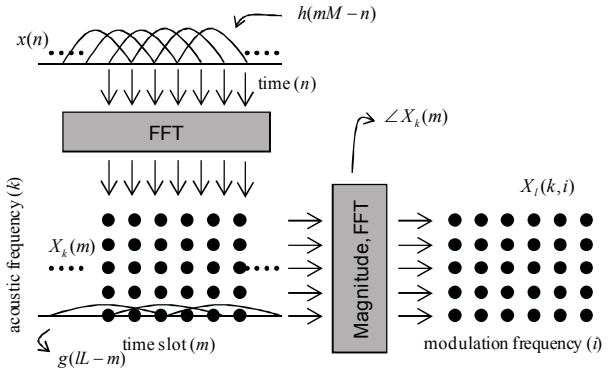

Figure 1 Framework of Modulation Frequency Analysis

The modulation analysis framework is illustrated in Figure 1. The magnitude of the subband envelope spectra is typically displayed in a modulation spectrogram representation. The modulation spectrogram of speech, music, and mixed signal are plotted in (a), (b), and (c) of Figure 2, respectively. High energy caused by the pitch region of speech signal is a prominent feature of modulation spectrum [8]. Thus the energy sum along all the acoustic frequencies will also represent high levels at modulation frequencies related to the pitch region as shown (d), (e) and (f) of Figure 2. Harmonic components that will be efficiently encoded by speech codecs can be separated using the distributions. The pitch searching range in Figure 2 is limited to that of AMR-WB [4]. The first peak point is calculated based on the convex hull algorithm [9].

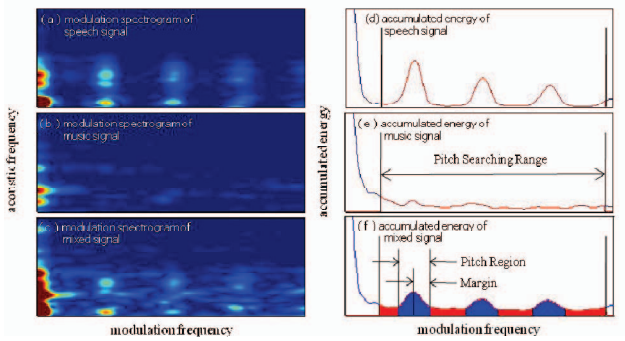

Figure 2 Modulation Spectrogram and its accumulated energy

The pitch region P of the harmonic component can be estimated by taking some margins at the location of first peak point and its harmonics in the pitch searching range. At frames in which no peak location is found, moving average value of previous frames can be used to estimate pitch region. Define $Q = \{i : i(f_s / IM) \in P\}$ as the set of modulation frequency indices i in the pitch region P when f_s is sampling frequency of input signal [8]. Modulation frequency energy over the harmonic signal's pitch region is represented as

$$E_l^h(k) = \sum_{i \in Q} \left|X_l(k,i)\right|^2. \tag{3}$$

We may consider that the range for non-harmonic signal is the outside of the pitch region in the pitch searching range.

$$E_l^r(k) = \sum_{i \notin Q} \left|X_l(k,i)\right|^2. \tag{4}$$

For each frame l, i.e. at the time instances $n = l(LM)$, the frequency suppressing function is determined from the ratio between target(harmonic) and remaining region [8].

$$F_l(k) = \frac{E_l^h(k)}{E_l^h(k) + E_l^r(k)}. \tag{5}$$

The obtained value is multiplied to the magnitude of each acoustic frequency in Eq. (1) to suppress non-harmonic components of input signal.

3. HYBRID CODING STRUCTURE OF AMR-WB AND HE-AAC

HE-AAC and AMR-WB are one of the representative codecs for audio and speech signal, respectively. The performance degrades when speech signals are coded by

HE-AAC, especially in low bitrates. Likewise, AMR-WB has a weak point in music signals. A structure to design a unified speech/audio codec covering all these weakness is proposed in this chapter.

Overall structure of the proposed algorithm is shown in Figure 3. In the structure, AMR-WB and HE-AAC must be operated simultaneously. The frame length of two codecs should be identical in that reason. At a 16kHz sampling data, HE-AAC has frame length of 1024 samples while AMR-WB has 320 samples. We modified the length of frame to 256-samples by removing downsampling operation in AMR-WB, and united four frames in sequences to make identical frame length with HE-AAC.

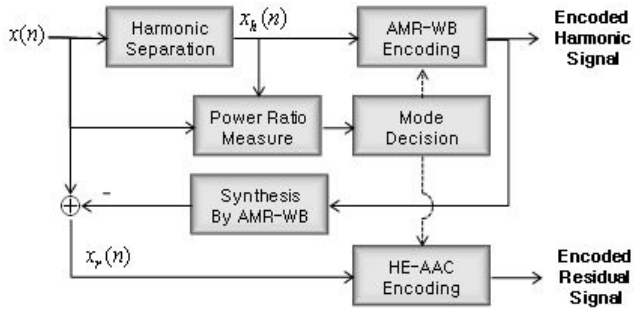

Figure 3 Structure of Proposed Hybrid Encoder using Single Channel Harmonic Source separation

Harmonic signal is separated from input signal using the separation method described in previous chapter. Then the ratio of power between the separated signal and the original signal is calculated by

$$Power\ Ratio = \frac{\sum_{frame}\left[x_h(n)\right]^2}{\sum_{frame}\left[x(n)\right]^2} \qquad (6)$$

From the value of this ratio, the amount of bits to be allocated of both codec is decided. The AMR-WB codec operates at 16 kHz sampling rate and consists of nine modes of bitrates whose maximum value is 23.85 kbit/s, while the HE-AAC uses bitrates under 20 kbit/s at the 16 kHz sampling rate. Therefore, signals of 16 kHz sampling rate for bitrates of 19.85 kbits/s are target environment of this paper.

AMR-WB encoder operates at nine different modes such as 6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 and 23.85 kbit/s [4]. With the constraints of total 19.85 kbit/s, we take only two lowest modes. Because a mode operated by AMR-WB is once decided, only the remaining to the total bitrates 19.85 kbit/s can be allocated to HE-AAC encoder. Modes of bitrates over 8.85 kbit/s result in too small bit allocation to HE-AAC. As expected, bitrates of 6.6 and 8.85 kbit/s also have poor performance in AMR-WB, but all the errors will be covered by HE-AAC by the structure of Figure 3. In addition to the usage of two encoding modes simultaneously, we also introduce special modes in which only one codec operates. Including these two cases, total four cases of allocating bits are possible. Parameters to determine which mode to select are the power ratios given in Eq. (6). Thresholds for each mode and amount of bits to be allocated to AMR-WB and HE-AAC are summarized in Table 1.

Table 1 Four different modes of the proposed codec

Mode	Criterion	Bitrate for AMR-WB (kbps)	Bitrate for HE-AAC (kbps)
A	$0 \le Pow_{ratio} \le Thr_C$	0	19.85
B	$Thr_C < Pow_{ratio} \le Thr_B$	6.60	13.25
C	$Thr_B < Pow_{ratio} \le Thr_A$	8.85	11.00
D	$Thr_A < Pow_{ratio} \le 1$	19.85	0

There can be a situation of operating with mode A after mode D, where all signals should be encoded by HE-AAC following the results of all signals encoded by AMR-WB. Perceivable discontinuity of two frames in this case can be occurred depending on the characteristics of the input signal. We avoid the problem by not allowing the rapid change from the mode D to A directly, and vice versa. It is plotted in Figure 4.

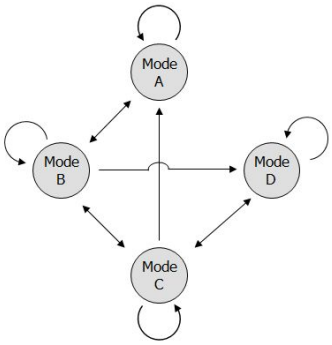

Figure 4 State Diagram of possible paths in each mode

4. PERFORMANCE EVALUATION

For the harmonic separation of signals sampled at 16kHz, variables in Eq. (1) and (2) were set to M=16, K=512, L=32, and I=512. The 48-point and the 64-point Hanning window were used for $h(n)$ and $g(n)$, respectively. The peak search range was set from 70Hz to 485Hz considering the pitch search interval of AMR-WB coder. And the margin to find out pitch region is set to 20Hz. Thresholds for mode decision in Table 1 are set to Thr_A = 0.5, Thr_B = 0.4 and Thr_C = 0.3.

We performed the multiple stimuli with hidden reference and anchor (MUSHRA) test [10] for subjective quality evaluation. The method provides an absolute measure of the audio quality of a codec which can be directly compared

with the reference signal. In the test, ten listeners who were trained and familiar with test environment were participated. Each listener used headphones. Results denote mean values and 95% confidence levels of test scores. Test materials were selected from MPEG conference [11] which were downsampled to 16kHz and mono recorded. Data sets were partitioned into four clusters as shown in Table 2. In second cluster, music signals are followed by speech signals or speech signals are followed by music signals.

Table 2 Data Sets from MPEG conference

Speech	Concatenated	Music	Mixed
Female (Chinese)	Jazz + male (Korean)	Classic	Pop + female (Korean)
Female (English)	Male (Korean) + jazz	Bass guitar	Commercial + male (English)
Male (German)	Male (German) + rock	Rock	Guitar + male (English)
Female(English)	Wedding + male (English)	Pop	Classic + male (English)
Male (Korean)		Pop	Accordion + female (English)
Male (English)		Symphony	Jazz + female (French)
		Jazz	

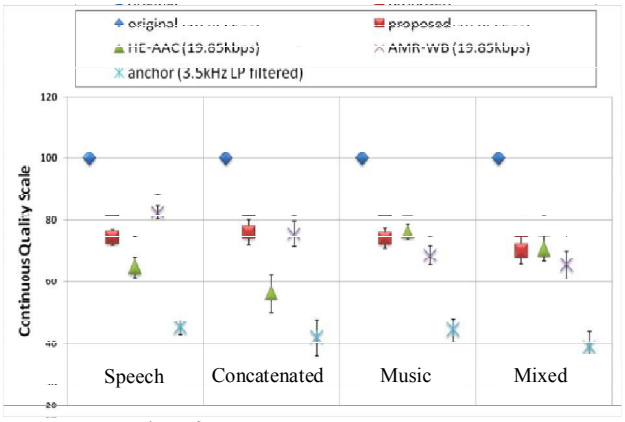

Figure 5 Results of MUSHRA test

As shown in Figure 5, the HE-AAC has poor performance compared to that of AMR-WB for speech signal, which confirms the results reported in the standard meeting [12]. Similar perceiving distortions are found in the result of concatenated input signals. In these two clusters, the proposed algorithm totally outperforms the HE-AAC. From the results, we can conclude that the proposed algorithm can cover the weak point of HE-AAC for speech signals. And the proposed algorithm shows higher scores than those of the AMR-WB and shows similar performance compared with the results of HE-AAC for music signals. It means that the proposed algorithm does not loose the superior performance of the HE-AAC in music signals. No degradation of performance is appeared in the proposed algorithm in mixed signals. From the results of Figure 5, we verified that the proposed algorithm showed consistent performance regardless of input signal's characteristics.

5. CONCLUSIONS AND DISCUSSION

We introduced an algorithm of separating harmonic signals from mixed input data by utilizing a framework of modulation frequency analysis. It is used as a preprocess of the unified speech and audio coding method designed by HE-AAC and AMR-WB. Under the condition of total bits being fixed, proper amount of bits are allocated to both codecs using the power ratio of input signal and separated harmonic signal. MUSHRA tests verified the superior performance of proposed coding method to standard codecs regardless of input data types.

6. REFERENCES

[1] Peter Noll, "Wideband Speech and Audio Coding", *IEEE Communications Magazine,* Vol. 31, Issue 11, Nov. 1993.

[2] Ming Yang, "Low bit rate speech coding" *IEEE Potentials,* Vol. 23, Issue 4, Oct-Nov 2004

[3] Peter Noll, "MPEG digital audio coding" *IEEE Signal Processing Magazine,* Vol. 14, Issue 5, Sept. 1997

[4] 3GPP TS 26.190 v6.1.1,"AMR-WB Speech Codec; Transcoding functions", July, 2005.

[5] M. Wolters, K. Kjorling, D. Homm and H. Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," *115th AES Convention, Preprint 5871*, Oct. 2003.

[6] Gibson, J.D, "Speech coding methods, standards, and applications", *IEEE Circuits and Systems Magazine,* Vol. 5, Issue 4, Fourth Quarter 2005

[7] 3GPP TS 26.290 v6.3.0, "AMR-WB+ Codec ; Transcoding functions", June, 2005.

[8] S.M. Schimmel, L.E. Atlas and Kaibao Nie, "Feasibility of Single Channel Speaker Separation based on Modulation Frequency Analysis," *Pro. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* 2007.

[9] Mermelstein, P. "Automatic segmentation of speech into syllabic units" *Journal of the Acoustical Society of America 58(4): 880-883* October 1975

[10] European Broadcasting Union, "MUSHRA – EBU method for subjective listening tests of intermediate audio quality." *Draft Technical Recommendation BMC607(rev.1) B/AIM 022(rev.9), Technical Department*, 2000, January.

[11] ISO/IEC JTC1/SC29/WG11, "Workplan for Exploration of Speech and Audio Coding", *MPEG2007/N9096* April 2007, San Jose, USA

[12] ISO/IEC JTC1/SC29/WG11, "Report on Evaluation of Speech and Audio Framework Database", *MPEG2007/w9252,* July 2007, Lausanne, Switzerland