An Optimization Scheme for SVAC Audio Encoder

Ruo Shu, Shibao Li, Xin Pan

An Optimization Scheme for SVAC Audio Encoder

Ruo Shu*, Shibao Li, Xin Pan*****

** China University of Petroleum*
Computer and Communication Engineering Institute
Qingdao 266580,
China
shuruo@upc.edu.cn

***China University of Petroleum*
Computer and Communication Engineering Institute
Qingdao 266580,
China
lishibao@upc.edu.cn

**** China University of Petroleum*
Computer and Communication Engineering Institute
Qingdao 266580,
China
langqin12390@126.com

ABSTRACT: *Both audio signals and MEL-frequency cepstral coefficients are encoded in SVAC audio encoder. These two independent processing leads to structural redundancies for the encoder design. This paper proposes an optimization scheme using MEL-frequency cepstral coefficients to realize high-frequency reconstruction and removes Bandwidth Extension module of the original encoder. Simulation results prove that the new encoder has considerable improvement on structure, reconstructed high-frequency precision and coding quality.*

KEYWORDS: *SVAC audio encoder, MEL-frequency ceptral coefficients, bandwidth expansion, linear predictive coding coefficients, quality evaluation.*

1. Introduction

Surveillance video and audio coding is designed for application in the field of national security and defense, and is important for maintaining public security and fighting and preventing criminal activities. Chinese government started to establish the national standard of surveillance video and audio coding (SVAC) in 2008, and the standard was approved and published in 2010. At present, the corresponding products are in promotion and have wide application future.

Different from conventional audio encoders mainly used in multi-media fields, SVAC audio encoder has some special technical characteristics for the particularity of applications, such as providing a high recognition rate for voiceprint recognition in back-end after signals are decoded and real-time responses when special events occur, etc. These new techniques lead to complex algorithms, and in addition, combining them with conventional coding technologies simply and directly result in many structural redundancies in SAVC audio encoder. Therefore structure optimization for the encoder needs further research.

This paper proposes an optimization scheme for SVAC audio encoder. The new method reconstructs high-frequency content of signal based on decoded MEL-frequency cepstral coefficients (MFCCs) extracted and coded in front-end, and removes the Bandwidth Extension (BWE) module in original encoder.

2. SVAC audio encoder

The overall framework of SVAC audio encoder is shown in Figure1. The coding processing can be divided into two parallel parts: audio coding and characteristic parameters coding. After sampling rate conversion, input PCM signal in one road is transferred to audio coding module, and can be encoded in different bitrates controlled by the levels of events detected in abnormal events detector. The other road signal is sent to characteristic parameters coding module in which the well-known MFCCs are extracted, quantified and coded in order to prevent the influence of speech coding distortion on voiceprint recognition in back-end.

In audio coding module, input signal is divided into low-frequency band and high-frequency band. Based on the signal type, low-frequency signal is encoded in two switch modes: Algebraic Code Excited Linear Prediction (ACELP) and Transform Audio Coding (TAC), and high-frequency signal is encoded through BWE technology which consumes little bitrates. In characteristic parameters extracting module, MFCCs are extracted, and then quantified and coded in subsequent modules. Here two coding modes are offered for MFCCs: direct coding mode and predictive coding mode (dotted line in Figure 1). The latter one firstly decodes bit stream of the audio encoder and acquires reconstructed signal, then

MFCCs of reconstructed signal and original signal are extracted respectively. At last, residual MFCCs are quantified and coded.

It can be seen that the connection between these two road modules lies only in MFCCs' predictive coding mode. From the perspective of encoder design, these two road signal processing (only at different sampling frequencies,) perhaps cause structural redundancies. In view of this, the structure of SVAC audio encoder has potential to be further optimized.

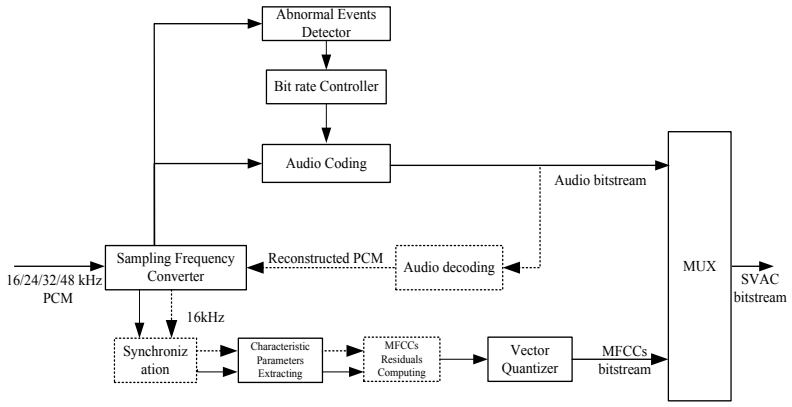

Figure 1. *SVAC audio encoder block diagram*

3. SVAC bandwidth expansion

As an enhancement technique for audio coding, even under the conditions of limited bit-rates, BWE can present high-frequency content with a very small amount of side information coded in front-end, and use part of low-frequency data combined with the side information to realize high-frequency reconstruction of signal, which further broadens the signal bandwidth of low bit-rate coding, and improves coding quality. BWE technology now has been employed in a variety of audio encoders.

BWE in SVAC audio encoder transmits energy gain factors and 8-order linear predictive coding (LPC) coefficients, and they occupy 16 bits only in one frame after quantization. In audio decoding, high-frequency excitation signal is obtained by gain adjustment of low-frequency excitation signal, and then synthesis filter is designed based on the decoded 8-order LPC coefficients. At last, high-frequency signal is reconstructed and then full-band signal is obtained.

It can be seen that BWE in SVAC audio encoder is based on LPC coefficients, however the encoder also chooses MFCCs as its characteristic parameters. Both of them are widely used speech signal characteristic parameters. Recent studies have

shown that the MFCCs have stronger correlation than LPC coefficients between high and low frequency of speech signals and the BWE technology based on MFCCs is superior to the current implementation scheme based on LPC coefficients.

4. Bandwidth expansion based on MFCCs

Firstly, we analyze the extraction steps of MFCCs. In order to be unified with SVAC audio encoder, we select speech signals whose frequency spectrums are in the range of 0 kHz ~ 8 kHz. Low-frequency is in 0 kHz ~ 4 kHz range, and high-frequency is in 4 kHz ~ 8 kHz range. Extraction steps are summarized briefly as follows:

a) *Pre-emphasis:* A single-pole high-pass filter is used to emphasize the high-frequency content.

b) *Windowing:* A Hamming window is used to mitigate the edge effect of discontinuities between frames.

c) *Power spectrum estimation:* Fast Fourier Transform(FFT) is applied to obtain the power spectrum;

d) *Mel-scale filters bank binning:* MEL scale triangular filters are applied. Record the number of low-pass filter with M, and the number of high-pass filter is denoted by N. The output energy of each MEL filter for low-frequency is $X_k\left(0 \le k \le M-1\right)$, and the output energy of each MEL filters for high-frequency is $Y_k\left(0 \le k \le N-1\right)$;

e) *Log operation:* The outputs of MEL filters log-energies are obtained.

f) *Discrete Cosine Transform(DCT):* DCT of the log-energies is applied to obtain the MFCCs as follows:

$$LF: \quad c_n = \sqrt{\frac{2}{M}} \sum_{k=0}^{M-1} (\log X_k) \cos\left(\frac{(2k-1)n\pi}{2M}\right), 0 \le n \le M-1$$
$$HF: \quad c_n = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} (\log Y_k) \cos\left(\frac{(2k-1)n\pi}{2N}\right), 0 \le n \le N-1 \qquad [1]$$

Where c_n is the *nth* MFCC.

Details of the derivation of MFCCs-based BWE are described as below.

Seen from the MFCCs extraction process above, two steps involve non-invertible loss of information; discarding phase information in step c) and the many-to-one mapping of the MEL filters in step d). And in the practical application of step f), such as speech recognition, also involves potential loss of information depending on whether the MFCCs vectors are truncated.

Assume that the MFCCs of high-frequency are under well preservation, so by Inverse DCT (IDCT) we can get the complete reconstruction of the energy Y_k of MEL filters. However the evaluation for Y_k in step *d)* is a many-to-one mapping, so

the MEL scale power spectrum cannot be reconstructed precisely through Y_k . Finer cepstral detail can, however, be obtained by interpolating from these log-energies of MFCCs by increasing the resolution of the IDCT as follows:

$$\log \hat{Y}_{k'} = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} c_n \cos\left(\frac{(2k'-1)n\pi}{2iN} \right), 0 \le k' \le iN-1 \quad [2]$$

Where i is an interpolation factor, which is decided by the resolution of MEL-scale. For example, when N＝6 and the MEL-scale resolution is 1, for the high-frequency ranges in 4 ~ 8 kHz, the calculation for i is:

$$i = \left| \frac{f_{mel}(8kHz) - f_{mel}(4kHz)}{N+1} \right| = 100 \quad [3]$$

Where $f_{mel}(\ )$ is the conversion formula between MEL frequency and the actual linear frequency:

$$f_{mel}(f_{Hz}) = 2595 \log_{10}(1 + f_{Hz}/700) \quad [4]$$

Through exponential transformation of $\hat{Y}_{k'}$ from Mel frequency to linear frequency, we can acquire high-frequency power spectrum, and then with Inverse Fast Fourier Transform (IFFT) we get the autocorrelation coefficients of high-frequency sequence. Furthermore we use Levinson-Durbin iterative algorithm to obtain the solution of Yule-Walker equation, finally, the LPC coefficients are achieved. The generation of high-frequency excitation signal will not be discussed in this paper, and we use the technology defined in SVAC.

5. Improvement on BWE

SVAC audio encoder codes audio signals and MFCCs simultaneously. So obviously MFCCs-based BWE mentioned above can be adopted to replace the original BWE based on LPC coefficients.

For the sampling frequency converter in Figure 1, the output signal transferred to MFCCs coding module is sampled in 16 kHz, and the output signal transferred to audio coding module is sampled in 12.8kHz～38.4kHz range. So bandwidth of encoded audio signal is in 6.4 kHz ～19.2 kHz range, and the bandwidth of BWE is half of encoded signals. In SVAC audio coder, there are many coding levels, and Table 1 lists three levels from 1.0 to 1.2. Each level has two internal sampling frequencies.

Technical parameter	Levels		
	1.0	*1.1*	*1.2*
internal sampling frequency(kHz)	12.8/16	24/25.6	32/38.4
samples per frame	512	512	512
maximum bitrates(bits/s)	24250	38800	58200

Table 1. *Three coding levels in SVAC audio encoder*

For example, when internal sampling frequency is 12.8kHz at level 1.0, high-frequency content of signal in 3.2 kHz ～6.4 kHz range is reconstructed by BWE and the maximal frequency is 6.4 kHz. At the same time, the signal sent to modules to extract and code MFCCs is sampled in 16 kHz, so its MFCCs extracted from high-frequency band can be used for high-frequency reconstruction. In this case, the maximal frequency reconstructed is 8 kHz higher than 6.4 kHz in the previous coder. So our optimization scheme is that the original BWE module can be removed completely when internal sampling frequency is lower than or equal to 16 kHz at low coding level because high-frequency reconstruction can be achieved from MFCCs.

Figure 2 is the framework of improved audio encoder for SVAC based on the above-mentioned idea. In this framework, sample frequency converter converts audio signal to a fixed 16 kHz sampling frequency. The converted signal is output to a low-pass filter (LP) and a high-pass filter (HP). The low-frequency signal is encoded by audio coding module and MFCCs are extracted and coded from full band signal. At last, SVAC bit streams are multiplexed.

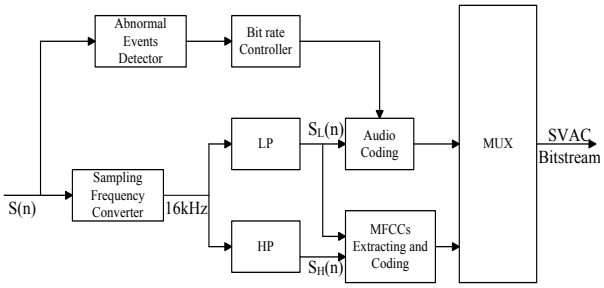

Figure 2. *New block diagram for SVAC audio encoder*

In decoder, the MFCCs of high-frequency signal decoded are converted to LPC coefficients to design the synthesis filter which is used to reconstruct the high-frequency content of original signal.

In this new SVAC audio encoder, BWE module is removed completely and the structure of encoder is simplified.

6. Simulation results

We compare our new encoder with SVAC audio encoder in simulation. 30 mono audio sequences are selected as experimental materials, including 18 male and female speech samples, 5 music samples and 7 abnormal events samples standardized in SVAC. Coding level is 1.0 with 3 different bit rates which are 10.8kbps, 17.2kbps and 24.4kbps respectively. Simulation concerns complexity and bit consumption, comparison of reconstructed LPC Coefficients, contrast of accuracy of reconstructed high-frequency band, and quality evaluation.

6.1. Analysis of complexity and bit consumption

Our encoder removes BWE module in SVAC and avoids calculation and quantization for gain factors and LPC coefficients, without introducing additional operations. So the computational complexity is reduced. The original encoder rebuilds LPC coefficients gained from inverse quantization and interpolation, and also decodes gain factors. However, the new decoder rebuilds LPC coefficients based on MFCCs. Their computational complexities are approximately equal. From the point of view of bit consumption, our encoder saves 16 bits consumption of BWE module in original SVAC, and the saved bits can be used for MFCCs coding or audio coding, which can increase the corresponding quantization accuracy and coding quality. The new encoder does not have to encode energy gain factors due to the energy information contained in MFCCs which namely are log-energies described above, so the new encoder saves bits again.

Here comes to the conclusion that our encoder achieves gains both on complexity and bit consumption.

6.2. Comparison of reconstructed LPC coefficients

In SVAC audio encoder, 8-order LPC coefficients of high band (4 kHz~8 kHz) are obtained and quantified. The dimensions of MFCCs in the new encoder is DIM (10, 6) which means 6 MFCCs are extracted from high band and quantified. The LPC coefficients are rebuilt based on these quantified MFCCs following the above method.

We compare these two sets of LPC coefficients with their real values and choose standard variance to describe error estimation. The definition is:

$$\sigma[i]=\sqrt{\frac{\sum_{n=1}^{M}\left(lpc'(i,n)-lpc(i,n)\right)^2}{M}} \qquad [5]$$

In the formula above, $lpc'(i,n)$ is distorted value of LPC coefficient and $lpc(i,n)$ is its real value. Also, i ($1 \le i \le 8$) is the index of LPC coefficient, and n represents the frame index. M is the number of frames. 2000 frames of different samples are chosen in our simulation. Table 2 presents average $\sigma[i]$($1 \le i \le 8$) in comparison.

	σ[1]	*σ[3]*	*σ[5]*	*σ[7]*
	σ[2]	*σ[4]*	*σ[6]*	*σ[8]*
original encoder	0.166	0.266	0.154	0.136
	0.318	0.200	0.133	0.071
new encoder	0.144	0.191	0.144	0.112
	0.205	0.173	0.119	0.058

Table 2. *Error estimation of LPC coefficients*

It can be seen from the table that compared to the original LPC coefficients, the error of reconstructed LPC coefficients in the new encoder is less than that in SVAC audio encoder. The main reason is that bit rates consumed by LPC coefficients in SVAC audio encoder are relatively less.

Here comes to the conclusion that reconstruction of LPC coefficients has higher accuracy in the new encoder.

6.3. Comparison of accuracy of reconstructed high-frequency spectrum

Fig.3 is an English female speech signal which is coded by both encoders. One frame is picked up to do comparison on high-frequency spectrum. In Figure 3, picture above is the frequency spectrum of original signal. The intermediate picture is spectrum of signals coded at level 1.0 when the internal sampling frequency is 12.8 kHz. In order to compare effectively, low-frequency part less than 4 kHz uses ACELP to encode, and frequency spectrum in 4 kHz～6.4 kHz range is reconstructed by BWE. Picture below is the reconstructed spectrum in our encoder in which low-frequency part under 4 KHz uses ACELP to encode and frequency spectrum in 4 kHz～8 kHz range is rebuilt based on MFCCs. We can see that two encoders have same performances at low-frequency band, but the new encoder

broadens spectrum from 6.4 kHz to 8 kHz, especially this broadening is realized without increasing coding bit rates.

Here comes to the conclusion that the new encoder has higher precision in high-frequency reconstruction.

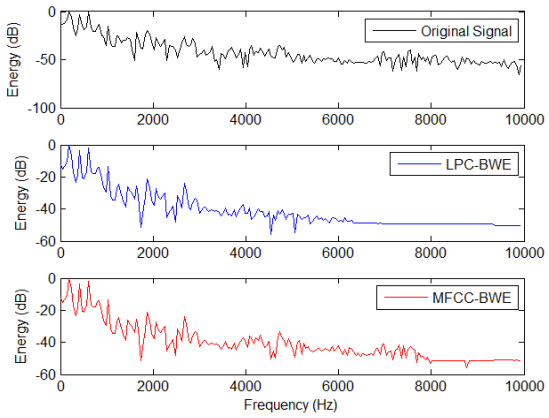

Figure 3. *Reconstructed high-frequency spectrum in two encoders*

6.4. Quality evaluation

We performed MUSHRA experiments for the two encoders. Considering when the spectrum of music signals are below 8 kHz, human's hearing is affected significantly more than speech signals, we only select speech samples and some abnormal events samples in our subjective evaluation.

There are 7 listeners who have rich listening experiences. The experiment follows the MUSHRA method. In order to ensure the score at reasonable range, the original audio samples are hidden in all samples, and 3.5 kHz and 7.0 kHz low-pass filtered signals are added as hidden anchors.

Figure 4 presents score curves of two encoders at different bit rates. The original encoder reconstructs high-frequency below 6.4 kHz, and the new encoder below 8 kHz. Their curves are near the curve of 7.0 kHz anchor. But we can see that coding quality of the new encoder significantly exceeds the original one, and the average gain is about 3.63.

We can get the following conclusion that coding quality of the SVAC audio encoder is improved by the new scheme at same bit rates.

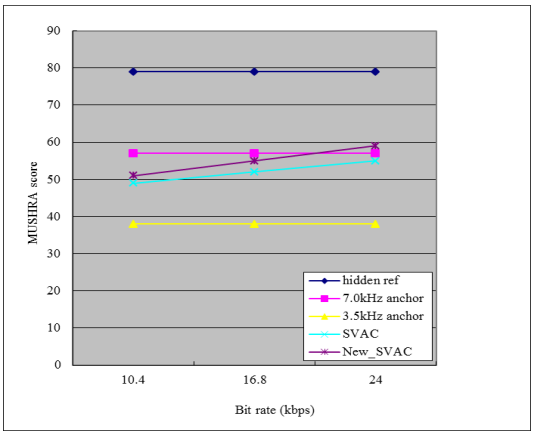

Figure 4. *MUSHRA evaluation*

7. Conclusion

We propose an optimization scheme for SVAC audio encoder. We reconstruct high-frequency content of audio signal only dependent on those MFCCs extracted in front-end, and eliminate BWE module to simplify the encoder's structure, along with improvement on coding quality without increasing bit rates. Simulation results show that the new encoder achieves apparent gains in various aspects.

On the other hand, the optimization scheme only works effectively at coding level 1.0. When at higher level, BWE based on MFCCs cannot reconstruct spectrum higher than 8 kHz because the signal for MFCCs extracting is only sampled at 16 kHz. In fact, BWE module in original SVAC audio encoder cannot reconstruct high-frequency content with good quality also. For example, when audio signal is sampled at 19.2 kHz, high band from 9.6 kHz to 19.2 kHz is estimated dependent on 16 bits side information only, and the reconstruction quality is not satisfying. How to design BWE module more effectively at different coding levels for SAVC audio encoder is our future research direction.

8. Acknowledgment

The work is supported by 'the Fundamental Research Funds for the Central universities' of China (No.12CX04078A).

The authors would like to thank the anonymous reviewers for their constructive comments and corrections and the MUSHRA listening test participants for their effort.

9. References

GB/T 25724-2010, “Technical Specification of Surveillance Video and Audio Coding”, *Beijing: Standardization Administration of the People's Republic of China (SAC)*, 2010.

Pulakka, H., Remes, U., Yrttiaho, S., Palomaki, K., Kurimo, M., Alku, P., "Bandwidth Extension of Telephone Speech to Low Frequencies Using Sinusoidal Synthesis and a Gaussian Mixture Model", *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 20, no. 8, pp.2219-2231, 2012.

Pulakka, H., Laaksonen, L., Myllyla, V., Yrttiaho, S., Alku, P., "Conversational evaluation of artificial bandwidth extension of telephone speech using a mobile handset", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP),* pp.4069-4072, 2012.

Chivukula, R.K., Reznik, Y.A., Devarajan, V., Jayendra-Lakshman, M., "Fast Algorithms for Low-Delay SBR Filter banks in MPEG-4 AAC-ELD", *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 20, no. 3, pp.1022-1031, 2012.

A. H. Nour-Eldin, P. Kabal., “Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech” , *In Proc. InterSpeech,* pp.2489–2492, 2007.

A. H. Nour-Eldin, P. Kabal., “Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech” . *In Proc. InterSpeech,* pp.53–56, 2008.

A. H. Nour-Eldin, P. Kabal., “Combining frontend-based memory with MFCC features for Bandwidth Extension of narrowband speech”, *In Proc. ICASSP,* pp.4001-4004, 2009.

T. Ramabadran, J. Meunier, M. Jasiuk, B. Kushner., “Enhancing distributed speech recognition with back-end speech reconstruction”, *In Proc. EuroSpeech,* pp.1859-1862, 2001.

Borgstrom, B.J., Alwan, A., “A Unified Framework for Designing Optimal STSA Estimators Assuming Maximum Likelihood Phase Equivalence of Speech and Noise ”, *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 19, no. 8, pp.2579-2590, 2011.

Thoma, H., “A system for subjective evaluation of audio, video and audiovisual quality using MUSHRA and SAMVIQ methods”, 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX),pp.31-32,2012.