ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 183–188 (2006)

HIGH QUALITY SPEECH CODEC EMPLOYING SINES+NOISE+TRANSIENTS MODEL

M. KULESZA, Ł. LITWIC, G. SZWOCH, A. CZYŻEWSKI

Gdańsk University of Technology
Multimedia Systems Department
Narutowicza 11/12, 80-952 Gdańsk, Poland
e-mail: maciej_k@sound.eti.pg.gda.pl

(*received June 15, 2006; accepted September 30, 2006*)

A method of high quality wideband speech signal representation employing sines+transients+noise model is presented. The need for a wideband speech coding approach as well as various methods for analysis and synthesis of sines, residual and transient states of speech signal is discussed. The perceptual criterion is applied in the proposed approach during encoding of sines amplitudes in order to reduce bandwidth requirements and to preserve high quality of speech. Therefore, the psychoacoustic model devised for perceptual speech coding is presented. The experimental results reveal that method for tonality estimation employed in the psychoacoustic model has a significant impact on perceptual coding accuracy. Various methods for tonality estimation are presented and compared.

Key words: speech coding, sines+noise+transients model, VoIP telephony.

1. Introduction

The importance of wideband speech coding, as well as a need for unique technology to encode speech, music and mixed content has been recognized in recent years. Although narrowband parametric speech codecs still dominate in today's telecommunication systems, it can be expected that wideband speech codecs will gradually replace them, as they provide more natural sound and reduce the listening effort. The important insight is that three novel wideband signal coding algorithms dedicated to both circuit and switched telecommunication systems were standardized by the ITU-T. Two of them, called AMR-WB and VMR-WB, can be viewed as pure speech coding algorithms based on ACELP technology, because they do not provide at least satisfactory quality of non-speech signals representation [2, 11]. Therefore, extended AMR-WB+ codec was introduced to overcome this limitation. Unfortunately, as the AMR-WB+ codec takes an advantage of hybrid ACELP/transform coding techniques, it introduces the coding delay up to 90 ms, and then in general it is not suitable for real-time two-way communication [13]. Accordingly, one can notice that there is still a need for wideband, high-quality, mid-delay speech codec with improved ability to encode non-speech signals.

Reversely to coding techniques based on the speech production model, that is insufficient for more complex signals, the codec proposed in this paper employs analysis technique allowing to extract sines, noise and transients parts of the signal. In the next step, each part of the entire signal is encoded using adequate technique, including the perceptual criteria. It has to be mentioned that sines+residual model is widely used as a powerful tool for signal modification (e.g. pitch, time-scale) [4]. The sines+residual signal representation was also employed for efficient narrowband speech coding at about 8 kbps rate. Additionally, it was found that it is a robust method for coding both speech signals and mixed audio content [1, 10]. Concerning this, it can be expected that extending the sines+residual model with transients selection module will further improve the signal representation accuracy. Similar to the AMR-WB codec, in the proposed approach the *Bandwidth Extension Technique* (BWE) is employed in order to restore the high frequency content of the encoded signal [6, 11]. It has to be mentioned that during various stages of encoding process the perceptual criterion is applied, allowing to reduce the bit-rate requirements for codec bit stream [9].

The organization of the paper is as follows. Section 2 presents a brief description of the proposed codec architecture. In Sec. 3, the details of the psychoacoustic model are discussed, and in Sec. 4 the results of experiments revealing the impact of the tonality estimation procedure on model accuracy are discussed. Finally, in Sec. 5 we summarize our contributions.

2. Wideband speech coding employing sines+noise+transients model

The structure of the encoder employing sines+noise+transients model is presented in Fig. 1.

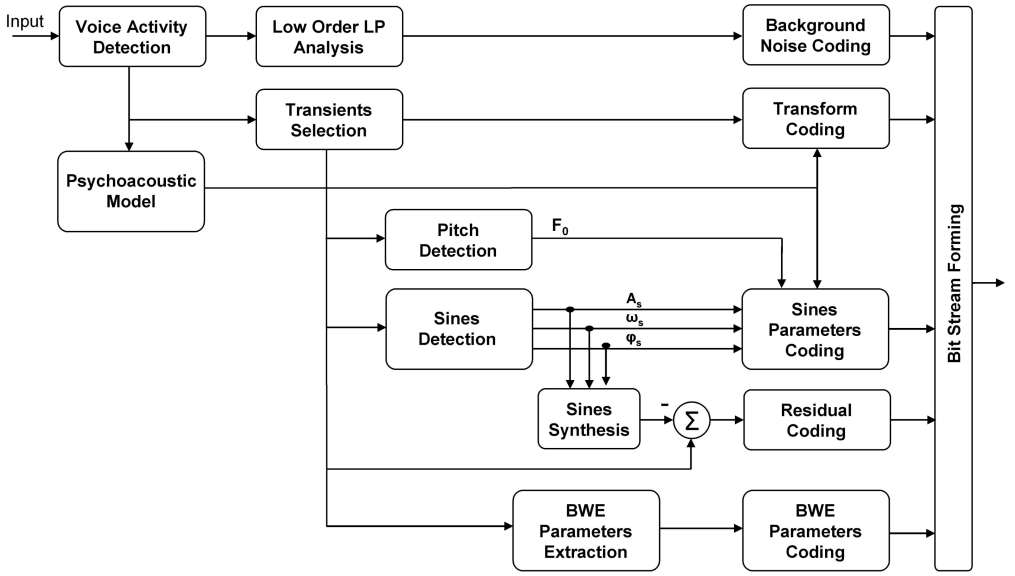

Fig. 1. Block diagram of proposed codec.

It has to be mentioned that during typical conversation each speaker remains mute for about 40% of time [2]. Therefore, in the proposed codec inactive frames are first detected and low-order LP analysis is performed in order to encode the background noise with negligible bit-rate (about 1 kbps). The background noise is synthesized in the decoder employing *Comfort Noise Generation* (CNG) technique [2]. If the frame is classified as active, transients signal classification is applied. As the proper transients representation has a significant impact on signal quality, transform coding method is employed for these segments of signal [7]. In this case, the psychoacoustic criterion is applied in order to reduce the bit-rate requirements. All remaining frames that are not classified as a transient are then processed according to the sines+noise model rules [14].

The amplitudes, frequencies and phases of sines are obtained using STFT approach. For each frame, local maxima of the amplitude spectrum are detected and *Sinusoidal Likeness Measure* parameter is estimated in order to allow distinguishing sines from noise-like components [12]. Basing on the set of sines parameters, local synthesis is performed and the residual (stochastic) signal is calculated. An important insight is that the LPC coefficients for residual signal are estimated with increased time resolution, as it is essential for preserving the high quality of signal [14]. Two methods for sines parameters encoding are employed depending on the spectral structure of the input signal. If voiced parts of speech or any other harmonic signal is processed, the fundamental frequency is estimated and transmitted to the decoder [1, 14]. If any frequencies of detected sines are not related to the pitch, the perceptual criterion is applied during partial tracking in order to eliminate noisy and masked partials, and then the components related to the sines are encoded separately [9]. Accordingly, the codec is able to encode both speech signal with reasonable bit-rate and high quality and audio/mixed content with a quality significantly higher then possible to obtain with CELP based codecs.

Although bandwidth extension techniques do not provide the signal quality as high as true wideband codecs, they allow improving the quality of narrowband signals significantly without sacrificing the bit-rate. In the presented codec architecture, BWE technique is employed in order to extend the synthesized signal bandwidth above 7 kHz. Thus the appropriate set of parameters allowing to reconstruct the upper band of the signal basing on the lower band signal properties are extracted and transmitted to the decoder [5, 6].

3. Psychoacoustic model

In most modern audio and speech coding systems psychoacoustics fundamentals are applied to achieve efficient quantization of signal parameters (for foundations of perceptual coding of speech and audio see [3]). The psychoacoustic models included in standards such as MPEG 1, AAC are based on excitation pattern model in which the amount of masking results from excitation. In the sines+residual model the psychoacoustic model can also benefit the tonal vs. non-tonal component determination. It is assumed that some noise or side-lobe components which are erroneously detected by

sine detection procedure, e.g. *Sinusoidal Likeness Measure*, can be discarded when the global masking curve or signal to mask ratios have low values [9].

In both psychoacoustic MPEG models tonality measure is used for masking threshold calculation. In the Model I determination of tonal and non-tonal components (maskers) is evaluated and different procedures are applied to estimate the masking. The tonal vs. non-tonal determination is based on empirically-based criterion gives a good masking curve estimation, however it pertains cases where the precision is not needed [12]. In the psychoacoustic Model 2 the hard tonal vs. non-tonal determination is substituted by a tonality measure estimation [3]. The tonality measure is calculated using simple linear predictor:

$$\hat{r}_k^t = 2r_k^{t-1} - r_k^{t-2}, \tag{1}$$

$$\hat{\phi}_k^t = 2\phi_k^{t-1} - \phi_k^{t-2}, \tag{2}$$

where r indicates magnitude and ϕ indicate phase values.

The Euclidean distance between the predicted and the actual values of magnitude and phase spectra for every frequency line is determined and is called *unpredictability measure* [3]. If the component is tonal the value of unpredictability measure is small resulting from good prediction. However, it should be noticed, that the tonality estimation is limited to individual frequency lines. For the real-life speech signals it is likely that tonal components would have slowly varying frequency and would spread over a few adjacent frequency lines. In such situation the prediction error (Eqs. (1), (2)) would be significant and would result in erroneous tonal vs. non-tonal characterization.

Some modifications to unpredictability measure can be applied to overcome this problem. One way is to predict only phase values since the magnitude values of components vary in a nonlinear way and the simple prediction (Eq. (1)) may be insufficient. The other way is to perform the phase prediction, however it is done on the basis of one preceding frame instead of two of them as in Eq. (2). The prediction based on one frame information should minimise the error in case of occurrence of component fluctuation between adjacent frequency lines. Such a phase prediction can be performed by means of a following procedure using additional analysis parameters:

$$\hat{\phi}_k^t = \phi_k^{t-1} + 2\pi k \frac{R}{N_{\mathrm{FFT}}}, \tag{3}$$

where R is hop distance and N_{FFT} is frame size.

The modified unpredictability measure is proposed in this paper using phase predictor defined in Eq. (3) instead of the predictor in Eq. (2).

4. Experiments

The experiments concerned the application of unpredictability measure and modified unpredictability measure in order to check how these measures correspond to actual signal's features. A voiced frame of speech signal was processed during the experi-

ments. The frequency values of significant sinusoidal components are presented in Table 1. Figure 2 presents the values of unpredictability measure (upper plot) and modified unpredictability measure (lower plot). It can be noticed that the values of unpredictability measure indicate more noise components (values above 0.5) than tonal components in the analysed frame, which does not correspond to the actual situation. On the other hand the values of modified unpredictability measure coincide with the determined tonal components (see Table 1).

Table 1. Values of unpredictability measure and modified unpredictability measure for evaluated tonal components.

Freq. of sinusoidal components [Hz]	123.8	247.6	376.8	1372	1523	1636	3407
Unpredictability measure	0.18	0.44	0.36	0.75	0.65	0.99	0.76
Modified unpredictability measure	0.04	0.02	0.05	0.65	0.78	0.06	0.48

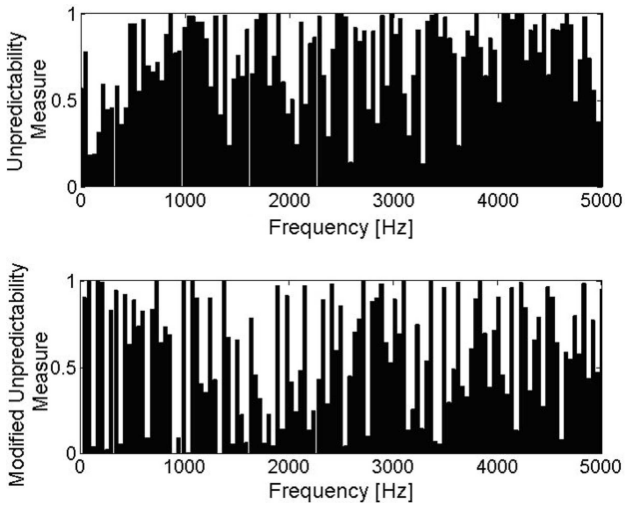

Fig. 2. Values of unpredictability measure (upper) and modified unpredictability measure (lower).

5. Conclusions

The architecture of wideband speech codec employing sines+noise+transients model has been proposed and presented. As the psychoacoustic model constitutes one of its key module, the experiments has been focused on its properties optimising. Although unpredictability measure has been successfully applied to audio coding standards such as MPEG and AAC, it was found that it often does not correspond to actual signal characteristics. In audio signals, containing many components the precise tonal measurements may not be of much importance. However, in speech signals built of a small number of components, the exact tonality measurement may improve coding efficiency. It has been found that modified unpredictability measure estimates the tonality better than standard

unpredictability measure but further experiments must be taken in order to validate the applicability of this measure to psychoacoustic models.

Acknowledgments

Research funded by the Polish Ministry of Science and Higher Education within the Grant No. 3 T11D 004 28.

References

[1] ANNADANA R., FERREIRA A., SINHA D., *A new low bit rate speech coding scheme for mixed content*, [in:] Preprint 120-th AES Convention, Paris, France, May 2006.

[2] AHMADI S., JELINEK M., *On the architecture, operation, and applications of VMR-WB: The new cdma2000 wideband speech coding standard*, IEEE Communication Magazine, **44**, 5, 74–81 (2006).

[3] BRANDENBURG K., *Perceptual coding of high quality digital audio*, [in:] Applications of Digital Signal Processing to Audio and Acoustics, KAHRS M., BRANDENBURG K. [Eds.], Kluwer Academic Publishers, 2002.

[4] CHAZAN D., HOORY R., SAGI A., SHECHTMAN S., SORIN A., SHUANG Z., BAKIS R., *High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification*, IEEE International Conference on Acoustic, Speech, and Signal Processing – ICASSP, Toulouse, May 2006.

[5] EHRET A., DIETZ M., KJORLING K., *State-of-the-art audio coding for broadcasting and mobile applications*, [in:] Preprint 114-th AES Convention, Amsterdam, The Netherlands, March 2003.

[6] FUEMMELER J., HARDIE R., GARDNER W., *Techniques for the regeneration of wideband speech form narrow band speech*, EURASIP Journal on Applied Signal Processing, 4, 266–274 (2001).

[7] KULESZA M., SZWOCH G., CZYZEWSKI A., *High quality speech coding using combined parametric and perceptual modules*, Proceedings of 13-th World Enformatika Conference, vol. 13, pp. 244–249, Budapest, May 2006.

[8] LAGRANGE M., MARCHAND S., RAULT J. B., *Sinusoidal parameters extraction and component selection in non-stationary model*, Proc. of 5-th International Conference on Digital Audio Effects (DAFx), Hamburg, September 2002.

[9] LEVINE S., SMITH III J., *Improvements to the switched parametric & transform audio coder*, Proc. 1999 IEEE Workshop on Application of Signal Processing to Audio and Acoustics, New York, October 1999.

[10] NAJAFZADEH-AZGHANDI H., KABAL P., *Perceptual coding of narrowband audio signals at 8 kbit/s*, Proc. IEEE Workshop Speech Coding, Pocono Manor, 1997.

[11] OJALA P., LAKANIEMI A., LEPHANAHO H., JOKIMIES M., *The adaptive multirate wideband speech codec: system characteristics, quality advances, and deployment strategies*, IEEE Communication Magazine, **44**, 5, 59–65 (2006).

[12] RODET X., *Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models*, Proceedings of the IEEE Time-Frequency and Time-Scale Workshop (TFTS'97), Coventry, UK, August 1997.

[13] SALAMI R., LEFEBVRE R., LAKANIEMI A., KONTOLA K., BRUHN A., TALEB A., *Extended AMR-WB for high-quality audio on mobile devices*, IEEE Communication Magazine, **44**, 5, 90–97 (2006).

[14] ZOLZER U., *DAFX – digital audio effects*, John Wiley&Sons, 2002.