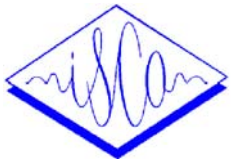

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NOVEL APPROACH OF LOW BIT-RATE SPEECH CODING BASED ON SINUSOIDAL REPRESENTATION AND AUDITORY MODEL

Wanggen Wan, Oscar C. Au**, Cyan L. Keung, Chi H. Yim*

Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
Email: *eewwg@ee.ust.hk, **eeau@ust.hk

ABSTRACT

In this paper, a new auditory spectrum based speech feature is proposed using sinusoidal representation and auditory model. The feature is optimized using the properties of auditory perception and masking. After quantizing and encoding the optimized feature parameters, a new speech-coding algorithm with average bit-rate of 3.25kbps is developed. The experimental results show that the synthetic speech retains most of the intelligibility and clearness of articulation of the original speech. Compared with the conventional algorithms, no voiced/unvoiced decision and pitch estimation are needed, complexity of the algorithm is much reduced, robustness and adaptation are both raised. The algorithm makes it possible to be realized with single DSP chip.

1. INTRODUCTION

Linear prediction (LP) model is more or less used in the conventional speech coding such as multi-pulse linear predictive coding (MPLPC), code-excited linear predictive coding (CELPC) and multi-band excitation (MBE) speech coding [1]. Such kind of speech coding method has got wider application and some of them have become the standards [2]. The speech coders based on LPC are quite sensitive to the variation of speakers and background noise because the LPC model is a kind of speech production model. The synthetic speech will degrade drastically in the both conditions, especially in the background noise condition. Instead of LPC method, harmonic representation is used to analyze the voiced speech [3] or voiced speech can be modeled by the summation of sinusoidal signal [4]. Another sinusoidal representation model for speech was proposed by McAulay et al [5][6]. In their paper, both voiced and unvoiced speech are modeled using the summation of sinusoidal signals, but some limitations are set for the unvoiced speech, i.e. the frequency interval should be small enough, otherwise there will be much artifacts in the synthetic speech.

In the sinusoidal representation model, all the local peaks in FFT amplitude spectrum of a frame of speech are picked out and each peak is represented by a sinusoidal function with different amplitude, frequency and phase. A frame of speech can be represented by the summation of all these sinusoidal functions. There are many advantages for this representation. For instance, no voiced/unvoiced decision is needed, even pitch estimation is not needed, the representation is not only suitable for speech signal, but also applicable for no speech signal such as music, and it is not so much sensitive to the variation of speakers as LPC model. It is proved that this representation can produce much better synthetic speech in high bit-rate speech coding system. In order to reduce the bit-rate, Ghizta[7] combined a simplified auditory model with sinusoidal representation and got a quite good results in his speech analysis/synthesis system. His algorithm might be applied to the middle bit-rate speech coding system.

Based on the work of both McAulay and Ghizta, we present a new auditory model consisting of second-order difference cochlear model (SDCM) and primary auditory nerve processing model (PANPM), which is used to extract the auditory spectrum-based feature parameters. The feature is then optimized using auditory perception and masking function. After quantizing and encoding these optimized feature parameters, a new speech-coding algorithm with average bit-rate of 3.25kbps is developed.

In this paper, we will first introduce auditory model, the auditory spectrum-based speech feature and its extraction method. Then we will present quantization and coding scheme for these parameters. The experimental results are finally presented.

2. AUDITORY MODEL

The auditory model is composed of SDCM and PANPM. The SDCM has the following form [8]:

$$\begin{aligned} & y_k(n) + b_{1k} y_k(n-1) + b_{2k} y_k(n-2) \\ & = A_k a_{0k} [u_s(n) - u_s(n-2)] \end{aligned} \qquad (1)$$

The corresponding transfer function can be obtained in the following form:

$$H_k(z) = A_k a_{0k} (1 - z^{-2}) / (1 + b_{1k} z^{-1} + b_{2k} z^{-2}) \qquad (2)$$

In Equation (1), $u_s(n)$ is the stape's velocity, $y_k(n)$ is the basilar membrane(BM) displacement in position x_k, parameters b_{1k}, b_{2k}, A_k and a_{0k} are all the coefficients with respect to position x_k along the BM[8].

The SDCM is a mathematical model for the BM vibration. With this model, characteristic frequency (CF) can be found for each position on the BM, and it will

gradually decrease from the base to the apex along the BM, but the 3dB bandwidth for each position on the BM will increase. These properties can be viewed in reference [8]. The SDCM is a kind of overlapped cochlear filter bank if considered in the view of electrical property. Its amplitude-frequency characteristic of transfer function is presented in Figure 1.

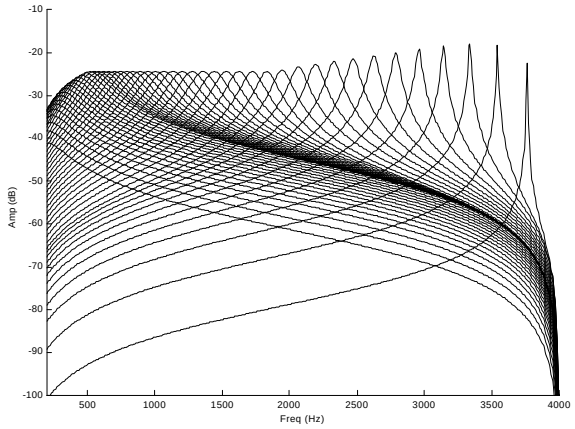

Figure 1. Amplitude-frequency characteristic of transfer function for SDCM

The SDCM inputs the speech and distributes it along the BM with the low frequency components on the apex and high frequency components on the base. Each cochlear filter corresponding to one position of BM analyzes the speech that is distributed within its range and outputs the results to the PANPM. The PANPM simulates the strength effect of each auditory nerve and ensemble effect of all the auditory nerves, and it selects the most significant frequency components using some rules based on the outputs of cochlear filter bank. It has the following four functions: 1) pick out the maximal value at the output of each cochlear filter for a frame of speech in the frequency domain and take down the corresponding frequency as dominant frequency. 2) After picking out all the maximal values for all cochlear filters and taking down all the corresponding dominant frequencies, count the number for each dominant frequency from low frequency domain to high frequency domain. Because one cochlear filter only outputs one dominant frequency, the number of each dominant frequency is in fact the number of cochlear filters which output this dominant frequency, that is to say, all these cochlear filters output the maximal value in the same frequency. 3) Each dominant frequency whose number is larger than or equal to the threshold M will be selected to be one of the initial auditory spectrum components. 4) According to the auditory perception and masking properties, the initial auditory spectrum will be optimized and those components that have smaller amplitudes and smaller frequency intervals will be removed, and finally a formal auditory spectrum is obtained.

Auditory spectrum only shows the positions of those frequency components that are considered to be most significant to human ear. The real amplitudes and phases in these frequency positions are not yet known. The amplitude and phase parameters can be found in so called auditory spectrum-based speech feature that will be described below.

3. SPEECH FEATURE AND ITS QUANTIZATION

In order to extract speech feature, those frequency components in speech FFT spectrum which are equal to those in auditory spectrum are selected, and their corresponding amplitudes and phases in FFT spectrum are also selected. So the auditory spectrum-based speech feature includes three kinds of parameters, i.e. frequencies, amplitudes and phases. Before quantizing these parameters, the number of auditory spectrum lines is to be limited in order to meet the requirements in both bit-rate and quality of synthetic speech. The maximal number of auditory spectrum lines is set to be no larger than 8, so there will be at most 24 parameters for a frame of speech. Experiments show that there are approximately 5 auditory spectrum lines at average for a frame of speech, so there will be at average 15 parameters for a frame of speech.

Limitation of auditory spectrum lines is only made for those speech frames that have more than 8 auditory spectrum lines. Those spectrum lines that have smaller amplitudes in FFT spectrum are removed in order to maintain the maximal value of 8.

According to the properties of auditory model, there will be more auditory spectrum lines for those speech frames in the range with high energy, and less auditory spectrum lines for those frames in the range with low energy. So the number of auditory spectrum lines for each speech frame is different, but it is no bigger than 8 for all frames as there is a limitation. For this reason, it is obviously not suitable to take all the parameters in one speech frame as one vector, because there will be different dimensions for different vector, which is very difficult to deal with. In addition, three parameters, i.e. frequency, amplitude and phase have different statistical properties. Frequency and amplitude have Gaussian distribution, but phase doesn't have. So frequency and amplitude for each spectrum line are constructed to be a two-dimensional vector, and will be quantized using vector quantization (VQ). Phase will be quantized using scalar quantization(SQ). In order to reduce the dynamic range, amplitudes are used in its logarithm form. Experiments show that it is good enough to use the code book size of 256 for VQ, i.e. 8 bits are used for VQ. Another 2 bits are used to correct the frequency whose bias will drastically affect the synthetic speech quality. 3 bits are used to quantize the phase because it only changes within the area $[-\pi, \pi]$. Table 1 shows the coding scheme for one spectrum line in a speech frame. Figure 2 shows the principles of speech feature extraction and parameter coding, and Figure 3 shows the

principles of the parameter decoding and speech synthesis.

Table 1. Coding scheme for one spectrum line in a frame of speech

Parameter type	Bits
Frequency & amplitude vector	8
Frequency correction	2
Phase	3
Total bits	13

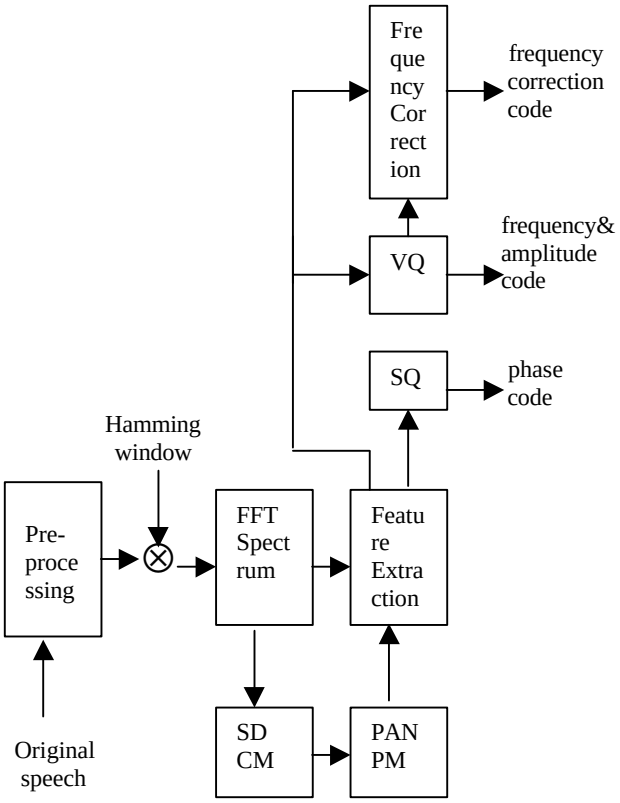

Figure 2. Principles of speech feature extraction and parameter coding

4. EXPERIMENTAL RESULTS

In Figure 2, speech is sampled using 8kHz sampling frequency and pre-emphasized using a first-order difference equation with coefficient 0.95 and windowed using Hamming window. Frame length is set to be 30ms with 10ms overlapping period, so the speech is analyzed with speed of 50Hz. 256 point FFT and 50 cochlear filters are used to analyze speech. According to the coding scheme in Table 1, the maximal coding bit-rate

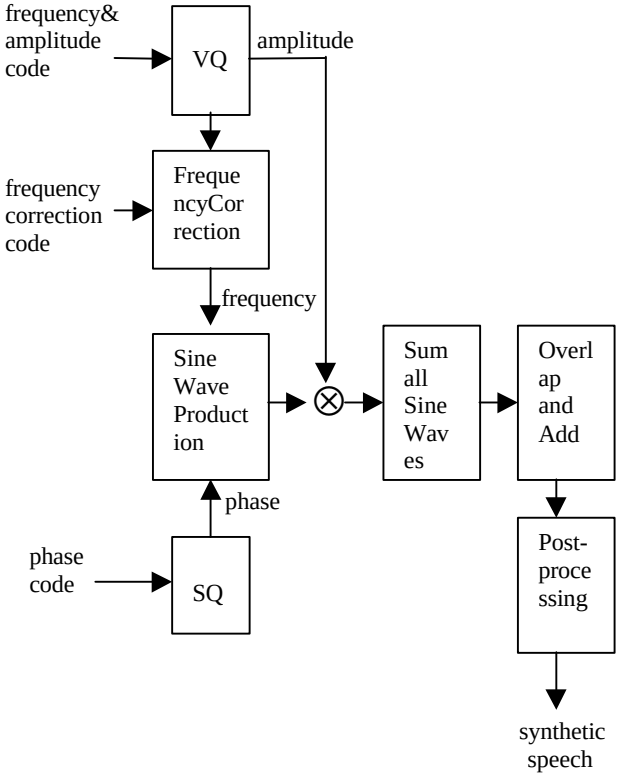

Figure 3. Principles of parameter decoding and speech synthesis

will be 5.2kbps if the number of auditory spectrum lines is set to be 8 at its maximal value , and if there are 5 spectrum lines at average for each frame of speech, the bit-rate will be 3.25kbps at average.

In Figure 3, each sine wave is produced using frequency and phase parameters and weighted by amplitude, and then summed to produce a frame of synthetic speech. The synthetic speech is produced by overlapping and adding each frame of speech, and then de-emphasized before output. Figure 4 shows a frame of original clean speech and Figure 5 shows its corresponding synthetic speech. Figure 6 shows a frame of noisy speech corresponding to Figure 4 with SNR=15dB, and Figure 7 shows the synthetic speech corresponding to Figure 6.

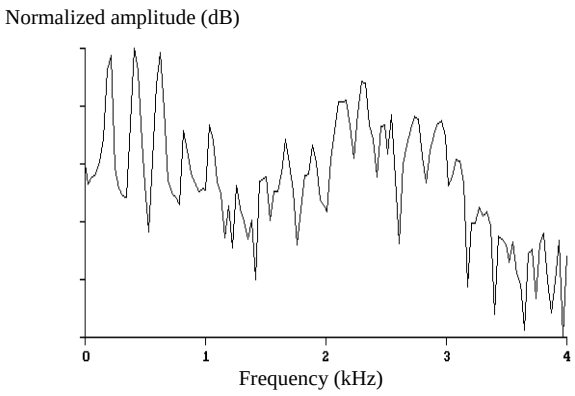

Figure 4. Spectrum of a frame of original clean speech

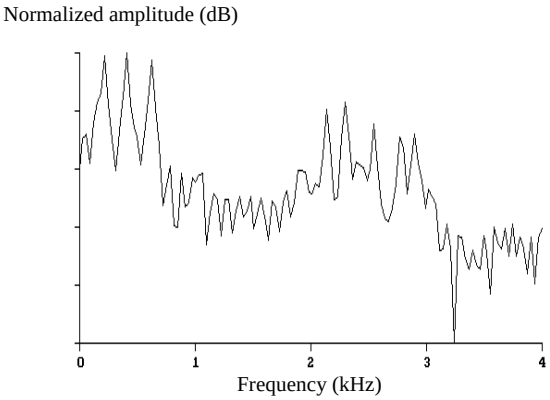

Figure 5. Spectrum of a frame of synthetic clean speech

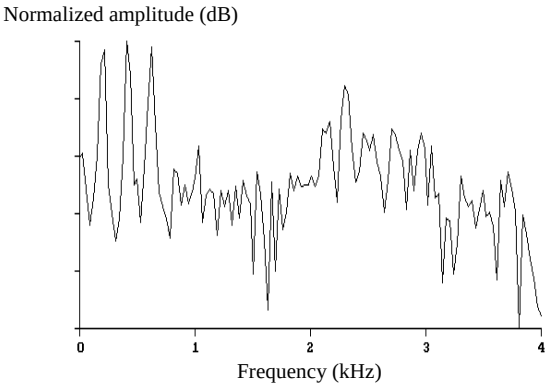

Figure 6. Spectrum of a frame of original noisy speech with SNR=15dB

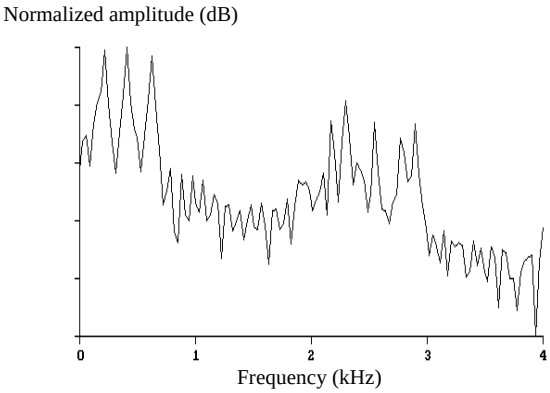

Figure 7. Spectrum of a frame of synthetic noisy speech corresponding to Figure 6

Comparing Figure 5 with Figure 7, it seems that there is no much difference. Experimental results show that the synthetic speech retains most of the intelligibility and clearness of the articulation of the original speech, except with some artifact which is produced by synthesizing unvoiced speech. The speakers can be easily identified from the synthetic speech, the quality of which is found to be better than that of LPC coding with same bit-rate and close to that of MBE 4.8kbps speech coder.

5. CONCLUSIONS

In this paper, an auditory spectrum-based speech feature is proposed and the corresponding algorithm for feature extraction is developed. After optimized with properties of auditory perception and auditory masking, the speech feature is quantized and a new speech-coding algorithm with average bit-rate of 3.25kbps is developed. After synthesized with "overlap and add" method, the synthetic speech is quite intelligible and clear. Its quality is found to be better than that of LPC coder with the same bit-rate and close to that of MBE 4.8kbps coder. Compared with the conventional algorithm, the new algorithm reduces its complexity because there is no need to make voiced/unvoiced decision and pitch estimation when analyzing speech. It is also not so much sensitive to the background noise and speaker variation as the conventional algorithms. The synthetic speech is almost unchanged with SNR=15dB. It seems that the new algorithm has more strong robustness and adaptation than the conventional ones.

Acknowledgments

We acknowledge the collaboration of Yuan Jingxian of Shanghai University and Keunglui and Yim Chiho of Hong Kong University of Science and Technology.

6. REFERENCES

[1] A. M. Kodoz, Digital speech (coding for low bit rate communications systems), Ch8, p.239-270, John Wiley & Sons, 1994.

[2] J. Campbell et al, "The proposed federal standard 1016 4800 bit/s voice coder: CELP", Speech Technology, p.58-64, April 1990.

[3] L. B. Almeida and J. M. Tribolet, "Non-stationary spectral modeling of voiced speech", IEEE Trans. Acoust., Speech and Signal Processing, Vol.31, p.374-390, June 1983.

[4] P. Hedelin, "A tone-oriented voice-excited vocoder", Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, p.205-208, March 1981.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. Acoust., Speech and Signal Processing, Vol.34, p.744-754, August 1986.

[6] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation", IEEE Trans. Acoust., Speech and Signal Processing, Vol.34, p.1449-1460, December 1986.

[7] O. Ghitza, "Auditory nerve representation criteria for speech analysis/ synthesis", IEEE Trans. Acoust., Speech and Signal Processing, Vol.35, No.6, p.736-740, 1987.

[8] Wan Wanggen, Yu Xiaoqing, "A second-order difference cochlear model", Acta Electronica Sinica, Vol.23, No.7, p.6-9, 1995(in Chinese).