PERCEPTUAL DOMAIN BASED SPEECH AND AUDIO CODER

L. Lin, E. Ambikairajah and W. H. Holmes
School of Electrical Engineering and Telecommunications
The University of New South Wales, UNSW SYDNEY NSW 2052, Australia
ll.lin@ee.unsw.edu.au, Ambi@ee.unsw.edu.au, H.Holmes@unsw.edu.au

Abstract

This paper applies a new auditory filterbank to wide band speech and audio coding. The coding algorithm is capable of producing high quality coded speech and audio, which account for temporal as well as spectral details. The analysis and synthesis are performed using a critical-band-rate auditory filterbank with superior auditory masking properties. The outputs of the analysis filters are processed to obtain a series of pulse trains that represent neural firing. Post and pre temporal masking models are applied to reduce the number of pulses in order to produce a compact time-frequency parameterization. The pulse amplitudes and positions are then coded using run-length coding algorithm.

1. Introduction

The Gammatone auditory filters were first proposed by Flanagan [5] to model basilar membrane motion, and were subsequently used by Patterson *et al.* [12] as a reasonably accurate alternative for auditory filtering, and have since become very popular. This parallel auditory filterbank outperforms the conventional transmission line auditory model [1][11] in terms of computational simplicity and has its applications in various types of signal processing required to model human auditory filtering. For example, Robert and Eriksson [4] applied them to produce a non-linear, active model of the auditory periphery. Kubin and Kleijn [6] applied them to speech coding.

Adequate modelling of the principal behavior of the peripheral auditory systems is still a difficult problem. One shortcoming of the Gammatone filter is that it does not provide an accurate frequency domain description of the tuning curves, because of its flatter upper-frequency slope. To overcome this problem, a new critical band auditory filterbank [10] is obtained based on the well-known masking curves [14][15]. Critical band filters designed using this method achieve high frequency domain accuracy and computational efficiency.

Ambikairajah *et al.*[3] proposed a new scheme for 16 kHz wide band speech and audio coding, in which the analysis and synthesis of both speech and audio signals is performed in the auditory domain. Gammatone filters are applied to the input signal to obtain an auditory-based time-frequency parameterization that comprises critical band pulse trains. This parameterization approximates the patterns of neural firing generated by the auditory nerve, and preserves the temporal information present in speech and music. An advantage of the parameterization is its ability to scale easily between different sampling rates, bit rates and signal types.

In this work, we apply the new filterbank in [10] to wideband speech and audio coding under the same paradigm as in [3]. To reduce computational complexity, 8th order IIR filters are used as analysis filters rather than FIR linear phase filters [3] with at least 100 coefficients for each filter. Temporal masking is applied to eliminate redundant information in the critical band pulse trains. A simple technique to code the pulse positions and amplitudes is proposed.

The paper is organized as follows: Section 2 of this paper describes the auditory filterbank used in the coding scheme. Section 3 presents techniques to reduce redundancy in the pulse trains, and also the quantization and coding scheme.

2. Critical Band Auditory Filterbank

The auditory filterbank proposed in [10] models the psychoacoustical tuning curves in Critical-band scale thus resulting in a parallel auditory filterbank with minimum-phase digital filters. The tuning curves are generated based on the relation between the psychoacoustical masking curves and the tuning curves. Masking is usually described as the sound-pressure level of a test sound necessary to be barely audible in the presence of a masker. Using narrow-band noise of a given centre frequency and bandwidth as maskers and a pure tone as test sound, masking patterns have been obtained by Zwicker and Fastl [14][15]. It is also observed that the shapes of the masking patterns at different centre frequencies are very similar when plotted using the critical band rate scale. In measuring the psychoacoustical tuning curves the level of the test tone is fixed, while the level of the masker is increased so that the test tone just becomes inaudible. A set of tuning curves in critical band scale is generated based on the similarity of the masking curves in critical band scale [10]. The tuning curves are consistent with the measurement of nerve tuning curves [7] and the basilar membrane response [13]. These tuning curves are modeled by a critical band auditory filterbank. The transfer function of the auditory filter model is expressed in z-domain as [10]:

$$G(z) = \frac{(1 - r_0 z^{-1})\,(1 - 2r_B \cos(2\pi f_B / f_s) z^{-1} + r_B^2 z^{-2})}{(1 - 2r_A \cos(2\pi f_A / f_s) z^{-1} + r_A^2 z^{-2})^4} \qquad (2.1)$$

The parameters in Equation 2.1 are described below: The parameters r_A and f_A are calculated by

$$f_A = \sqrt{f_c^2 + B_w^2}\,, \qquad r_A = e^{-2\pi B_w / f_s} \qquad (2.2)$$

where f_s is the sampling frequency. The bandwidth B_w and f_c in Equations 2.2 are calculated from the following equations [14][15]

$$Z_c = 13\tan^{-1}(0.76 f_c / 1000) + 3.5\tan^{-1}(f_c / 7500)^2$$
$$B_w = 25 + 75[1 + 1.4(f_c / 1000)^2]^{0.69} \qquad (2.3)$$

where Z_c is the critical band rate in Bark corresponding to f_c. The spacing of Z_c is linear in critical band rate.

The parameter r_0 is chosen as $r_0 = 0.955$. The term $(1 - 2r_B\cos(2\pi f_B / f_s)z^{-1} + r_B^2 z^{-2})$ produces a notch filter with a sharp dip at a point to the right of the centre frequency f_c so that the upper-frequency slope of the overall filter is steep enough. The parameter r_B is chosen as $r_B = 0.985$. To ensure the notch happens at a frequency location about 60 dB lower than the centre frequency f_c, the empirical formula that we obtained can be used to choose f_B:

$$f_B = 117.5(f_c / 1000)^2 + 1135.5(f_c / 1000) + 277.0 \qquad (2.4)$$

where f_c should be in Hz.

The frequency responses of five filters at critical bands 4, 7, 10, 13 and 16 are plotted in Figure 1 together with the corresponding tuning curves. The auditory filterbank used in speech and audio coding consists of 21 such filters.

Attempts to fit Gammatone filters to the tuning curves have been made. It is found that the upper-frequency slopes of the Gammatone filters are not steep enough to model the tuning curves.

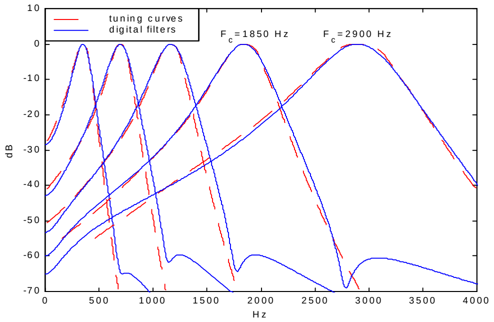


Figure 1. Comparison of auditory filters with auditory tuning curves (solid line: digital filters; dashed: tuning curves)

3. Speech and Audio Coding

3.1 Speech/audio Coding Using an Auditory Filterbank

The speech and audio coding system implemented in this work is an IIR/FIR analysis/synthesis scheme [8], which is shown in Fig. 2. Other possible analysis/synthesis filterbank implementations are considered in [8].

Each IIR analysis filter has 8 poles and 3 zeros. Analysis filterbank can also be implemented in FIR form [3][6], but at least 100 coefficients are required for each FIR filter to approximate the impulse response of the IIR filter with reasonable accuracy. The auditory filterbank is also power-complementary, which can be described as

$$\sum_{i=1}^{M} |G_i(e^{j\omega})|^2 \approx C \qquad (3.1)$$

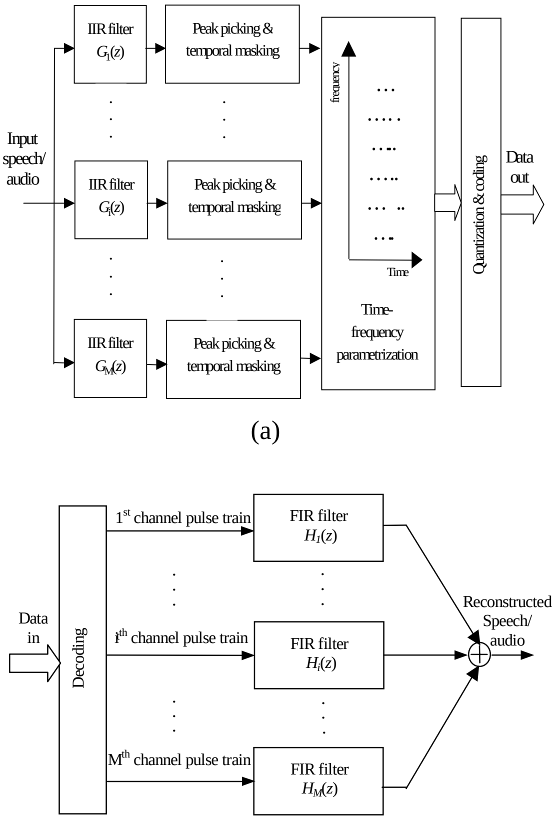
where $G_i(e^{j\omega})$ is the frequency response of the analysis filter at the ith channel and M is the total number of channels. If we choose the synthesis filters as

$$h_i(n) = g_i(-n) \quad \text{f or} \quad i = 1 \ldots M \qquad (3.2)$$

i.e., the synthesis filterbank is implemented as FIR filters obtained from the time-reversed impulse response of the analysis filters in the corresponding channel, then the reconstruction is nearly perfect and the following property holds

$$\sum_{i=1}^{M} g_i(n) * h_i(n) \approx C\delta(n)\,. \qquad (3.3)$$

Each FIR synthesis filter has 128 coefficients hence 8 ms is required to make the filter causal for f_s =16 kHz.



(a)

(b)

Figure 2. Speech/audio using an auditory filterbank (a) Analysis; (b) Synthesis

3.2 Temporal Masking

The output of each filter was half-wave rectified, and the positive peaks of the critical band signals were located. Physically, the half-wave rectification process corresponds to the action of the inner hair cells, which respond to movement of the basilar membrane in one direction only. Peaks correspond to higher rates of neural firing at larger displacements of the inner hair cell from its position at rest. This process results in a series of critical band pulse trains, where the pulses retain the amplitudes of the critical band signals from which they were derived.

In recognition of the fact that weaker signal components become inaudible by stronger signal components in the same critical band that precede them in time, temporal post masking and pre masking models are employed. When the signal precedes the masker in time, it is called pre-masking; when the signal follows the masker in time, the condition is post-masking. A strong signal can mask a weaker signal that occurs after it and a weaker signal that occur before it [2] [14] [15].

3.2.1 Post-masking

The masking threshold $y_i(n)$ for this temporal post-masking decays approximately exponentially following each pulse, or neural firing. A simple approximation to this masking threshold, introduced in [3], is

$$y_i(n) = \begin{cases} x_i'(n), & x_i'(n) > c_0\, y_i(n-1) \\ c_0\, y_i(n-1), & otherwise \end{cases} \qquad (3.4)$$

where $x_i'(n)$ is the ith of M = 21 simultaneous masked critical band pulse train signals, $c_0 = \exp(-\tau_i)$, and n is the discrete time sample index. The time constants τ_i, $1 \le i \le M$, were determined empirically by listening to the quality of the reconstructed speech, and values between 0.008 (i = 1) and 0.03 (i = 21) were chosen. All pulses with amplitude less than the masking threshold $y_i(n)$ were discarded. The threshold is shown by the dashed lines in Figure 3a.

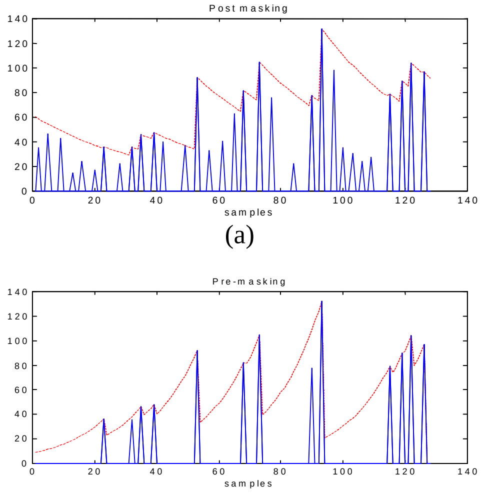
3.2.2 Pre-masking

Pre-masking is added in this work. The masking threshold $z_i(n)$ for this temporal pre-masking is chosen as

$$z_i(n-1) = \begin{cases} x_i''(n-1), & x_i''(n-1) > c_1 z_i(n) \\ c_1 z_i(n), & \textit{otherwise} \end{cases} \qquad (3.5)$$

where $x_i''(n)$ is the ith critical band pulse train after post-masking, c_1 is chosen as $c_1 = \exp(-3\tau_i)$ to simulate the fast exponential decay of pre-masking. All pulses with amplitude less than the masking threshold $z_i(n)$ were discarded (Figure 3b). A reduction rate of 10% can be achieved by pre-masking on the pulses obtained after post-masking.

The purpose of applying masking is to produce a more efficient and perceptually accurate parameterization of the firing pulses occurring in each band. Experiments show that the application of temporal masking reduces the pulse number to 0.70*N* (*N* is the frame size) while maintaining transparent quality of the coded speech and audio. This is a significant improvement upon the reduction to 1.26*N* in the previous application [3], where the Gammatone filters were used as the front end. The improvement is mainly due to the spectral shape of the new auditory filters used in this work.



(a)

(b)

Figure 3. (a) Pulse reduction using post-masking; (b) further pulse reduction using pre-masking (solid lines: pulses; dashed lines: thresholds)

3.2.3 Thresholding

The pulses in the silent frames obtained after auditory filtering and peak picking are most likely due to background and quantization noise. These pulses are at random positions and their magnitudes are very small hence the sound synthesized from these pulses are inaudible. By thresholding, these pulses can be eliminated without affecting the quality of the synthesized signal. The threshold can be found from the silent frames at the beginning of the coding process.

3.3 Quantization and Coding

The pulse train in each critical band after redundancy reduction was finally normalized by the mean of its non-zero pulse amplitudes across the frame. Thus, the parameterization consists of the critical band gains (incorporating the normalization factors), and a series of critical band pulse trains with normalized amplitudes. For each frame, the signal parameters requiring for coding are the gain of each critical band, the amplitude and the position of each pulse.

3.3.1 Pulse Amplitudes

Each critical band gain is quantized to 6 bits and the amplitude of each pulse is quantized to 1 bits, which will not result in any perceivable deterioration in the quality of the reconstructed speech and audio signal. Alternatively, vector quantization can be adapted to reduce the bits required for coding the amplitude [3].

3.3.2 Pulse Positions

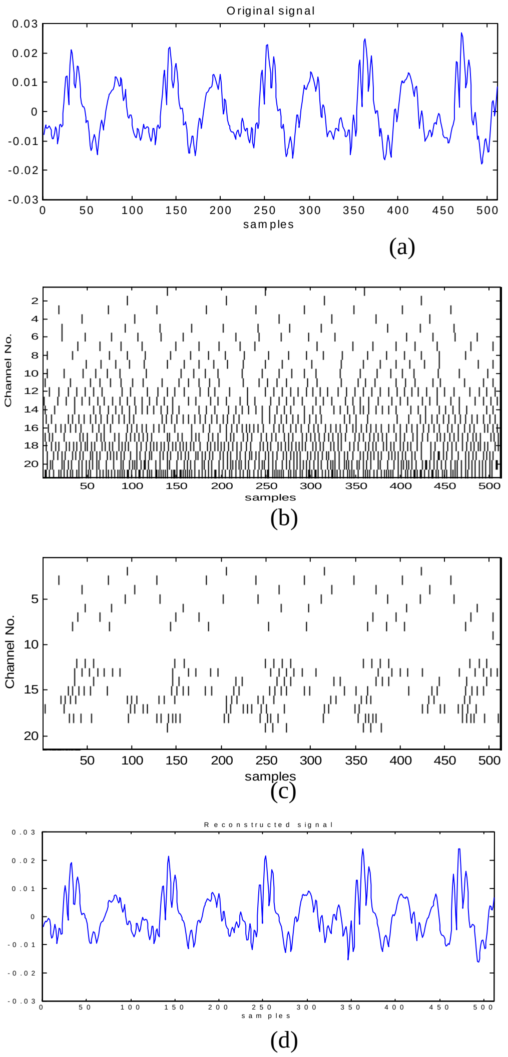
The pulse positions are coded using a new run-length coding technique. After temporal masking and thresholding, most locations on the time-frequency map have zero pulses. This suggests that we can just code the relative position of the neighboring pulses or the number of zeros between pulses. Specifically, the data in all channels with one frame (8 ms) is concatenated into one large vector and is scanned for pulses. Then the number of zeros preceding each pulse is coded using 7 bits. An example is shown below

$$\ldots 0\;0\;1.2\;\underbrace{0\;0\ldots0\;0}_{\substack{128\text{ zeros}\\(0000000)}}\;\underbrace{0\;0\ldots0\;0}_{\substack{120\text{ zeros}\\(1111000)}}\;1.5\;\underbrace{0\;0\ldots0\;0}_{\substack{89\text{ zeros}\\(1011001)}}\;0.8\;0\;0\ldots$$

If the number of zeros is over 128, a code word of 0000000 is generated and the counting of zeros restarts after the 128 zeros. If during the decoding process, seven consecutive zeros are encountered, then no pulse will be generated and the decoding carries on to the next code word. This coding strategy can also be called run-length coding and is lossless.

The overall average bit rate resulting from this coding scheme is 58 kbps. This is an improvement upon the 69.7 kbps in the previous work [3]. By exploring the statistical correlation and redundancy among the pulses, Huffman coding or arithmetic coding can be applied to further reduce the bit rate.

The synthesis process starts with decoding to obtain the pulse train for each channel and then filtering the pulse train by the corresponding FIR synthesis filter. Summing up the output from all filters results in the reconstructed speech and audio signal (Figure 4), which is perceptually the same as the original.



(a)

(b)

(c)

(d)

Figure 4. (a) Original speech; (b) Pulses obtained from peak-picking; (c) Pulses retained after temporal masking; (d) Reconstructed speech

4. Conclusions

A new critical band auditory filterbank is applied to speech and audio coding. The auditory filterbank has been designed to model the psychoacoustical tuning curves. It is shown that this filterbank has superior masking properties when applied to speech and audio coding. The filterbank is implemented as IIR/FIR analysis/synthesis scheme hence computationally more efficient. Simple run-length coding algorithm is used to code the positions of the pulses. The reconstructed speech and audio signals are perceptually transparent. This auditory-system-based coding paradigm produces high quality coded speech and audio, is highly scalable, and has moderate complexity.

5. References

[1] Ambikairajah, E., Black, N. D. and Linggard, R., “Digital filter simulation of the basilar membrane”, *Computer Speech and Language*, 1989, vol. 3, pp. 105-118.

[2] Ambikairajah, E., Davis, A. G., and Wong, W. T. K., “Auditory masking and MPEG-1 audio compression”, *Electr. & Commun. Eng. Journal*, vol. 9, no. 4, August 1997, pp. 165-197.

[3] Ambikairajah, E., Epps, J. and Lin, L., “Wideband speech and audio coding using Gammatone filter banks”, *Proc. ICASSP* (Salt Lake City), 2001, pp. 773-776.

[4] Robert, A. and Eriksson, J., “A composite model of the auditory periphery for simulating responses to complex sounds”, *J. Acoust. Soc. Am.*, vol. 106, 1999, pp. 1852-1864.

[5] Flanagan, J.L., “Models for approximating basilar membrane displacement”, *Bell Sys. Tech. J*, 1960, vol. 39, pp. 1163-1191.

[6] Kubin, G. and Kleijn, W.B., “On speech coding in a perceptual domain”, *Proc. ICASSP* (Phoenix, USA), 1999, pp. 205-208.

[7] Liberman, M.C. “Auditory-nerve response from cats raised in a low-noise chamber”, *J. Acoust. Soc. Am.*, vol. 63, 1978, pp. 442-455.

[8] Lin, L., Holmes, W. H. and Ambikairajah, E., “Auditory filter bank inversion”, *Proc. ISCAS 2001* (Sydney), May 6-9, 2001. Vol. 2 pp: 537 – 540.

[9] Lin, L., Ambikairajah, E. and Holmes, W. H., "Log-magnitude modelling of auditory tuning curves", *Proc. ICASSP* (Salt Lake City), 2001, pp. 3293-3296.

[10] Lin, L., Ambikairajah, E. and Holmes, W. H., "Auditory filterbank design using masking curves", *Proc. EUROSPEECH 2001-Scandinavia* (7th European Conference on Speech Communication and Technology), pp. 411-414.

[11] Lyon, R.F., "A computational model of filtering detection and compression in the cochlea", *Proc. ICASSP*, 1982, pp. 1282-1285.

[12] Patterson, R.D., Allerhand, M., and Giguere, C., "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform", *J. Acoust. Soc. Am.*, vol. 98, 1995, pp. 1890-1894.

[13] Rhode, W.S., "Observation of the vibration of the basilar membrane of the squirrel monkey using the Mossbauer technique", *J. Acoust. Soc. Am.*, vol. 49, 1971, pp. 1218-1231.

[14] Zwicker, E. and Zwicker, U. T., "Audio engineering and psychoacoustics: matching signals to the final receiver, the human auditory system", *J. Audio Eng. Soc.*, vol. 39, No. 3, 1991, pp. 115-125.

[15] Zwicker, E. and Fastl H., *Psychoacoustics: Facts and models*, Springer-Verlag, 1999.