WIDEBAND SPEECH AND AUDIO CODING USING GAMMATONE FILTER BANKS

Eliathamby Ambikairajah, Julien Epps, Lee Lin

School of Electrical Engineering and Telecommunications
The University of New South Wales, UNSW Sydney, NSW 2052, Australia
E.Ambikairajah@unsw.edu.au, J.Epps@unsw.edu.au, ll.lin@ee.unsw.edu.au

ABSTRACT

Considerable research attention has been directed towards speech and audio coding algorithms capable of producing high quality coded speech and audio, however few of these use signal representations which account for temporal as well as spectral detail. This paper presents a new technique for 16 kHz wideband speech and audio coding, whereby analysis and synthesis are performed using a linear phase gammatone filter bank. The outputs of these critical band filters are processed to obtain a series of pulse trains that represent neural firing. Auditory masking is then applied to reduce the number of pulses, producing a more compact time-frequency parameterization. The critical band gains and pulse amplitudes and positions are then coded using a combination of non-uniform quantization, arithmetic coding and vector quantization. This coding paradigm produces high quality coded speech and audio, is based upon well-known models of the auditory system, is highly scalable, and has moderate complexity.

1. INTRODUCTION

There is currently no single coding algorithm capable of supporting efficient coding of both speech and music at low bit rates. Shortcomings of the various existing algorithms have recently motivated researchers to seek a model of the auditory periphery for speech coding [5] that is accurate, provides an efficient parameterization of the signal, and can be inverted with relatively low computational effort.

The use of auditory models in speech and audio coding is well established, particularly in MPEG audio compression [1],[3]. Several researchers [9] have investigated the inversion of auditory models using iterative processing, however auditory inversion has only recently been achieved at low computational expense using gammatone filters [5], where the application was restricted to speech coding.

This paper presents a new paradigm in which the analysis, coding and synthesis of both speech and audio signals is performed in the auditory domain. Linear phase 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.

The inclusion of simultaneous and temporal masking models in our algorithm allows the elimination of redundant information in the critical band pulse trains and thus increases the feasibility of the parameterization for coding applications.

Section 2 of this paper describes the process by which speech and audio signals are decomposed into the proposed parameterization. In section 3, the procedure for synthesis from the critical band pulse trains is reported. Coding of the auditory domain parameterization for 16 kHz sampled wideband speech and audio signals is considered in section 4.

2. PARAMETERIZATION OF SPEECH AND AUDIO SIGNALS

Analysis of the input speech or audio signal in the proposed coder comprises linear phase critical band filtering, peak location, masking, and these stages are illustrated in Fig. 1.

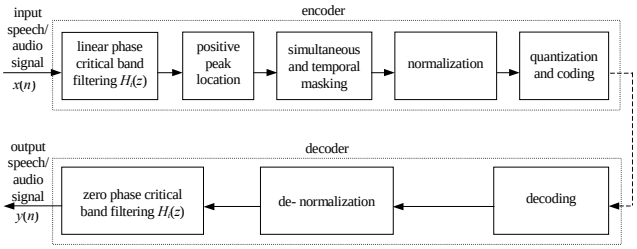

Figure 1. Auditory domain speech/audio analysis, coding and synthesis for the *i*th channel.

2.1 Critical Band Linear Phase Gammatone Auditory Filter Bank

The analysis filter bank employed in this approach contains gammatone filters $H_i(z)$ whose centre frequencies and bandwidths match those of the critical bands [10]. Thus, for an 8 kHz signal bandwidth, 21 filters were used. Existing multi-pulse wideband speech and audio coders have only used up to four sub-bands [7], however these do not allow the application of temporal masking similar to models of the human ear.

Gammatone filters can be implemented using FIR or IIR filters [6] or frequency domain techniques. In this research, FIR filters were employed in order to implement linear phase filters with identical delay in each critical band. The analysis filters had a length of $2N$-1 coefficients, and were obtained by convolving a sampled gammatone impulse response $g(n)$ of length N = 100 with its time reverse, where

$$g(n) = a(nT)^{N-1} e^{-2\pi bERB(f_c)nT} \cos(2\pi f_c nT + \varphi), \qquad (1)$$

f_c is the centre frequency, T is the sampling period, n is the discrete time sample index, a, $b \in \Re$ are constants, and $ERB(f_c)$ is the equivalent rectangular bandwidth of an auditory filter. At a moderate power level, $ERB(f_c) = 24.7 + 0.108 f_c$. Examples of the impulse responses of these filters are shown in Fig. 2.

Filters with a constant delay across all bands are useful in the analysis stage for time-aligning critical band pulses (see Fig. 4) across different bands. A further advantage is that the analysis and synthesis filters have identical coefficients (see section 3).

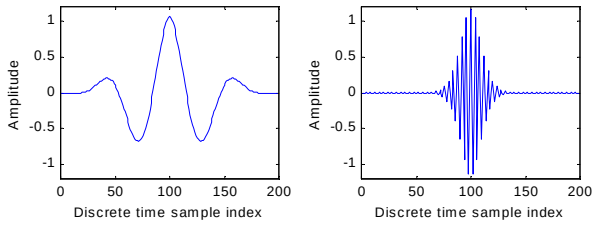

Figure 2. Impulse responses of the (a) 3rd (centre frequency 250 Hz) and (b) 18th (centre frequency 4 kHz) critical band linear phase gammatone filters.

2.2 Auditory Masking

The output of each filter was half-wave rectified, and the positive peaks of the critical band signals were located using a simple peak detector, as described in [5]. 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 lower power components of the critical band signals are rendered inaudible by the presence of larger power components in neighbouring critical bands, a simultaneous masking model was employed. A masking model similar to that of MPEG [3] was utilized to calculate the masking threshold L_i (dB) for the ith critical band in each frame, however the optimal masking model for this scheme has yet to be determined. The simultaneous masked pulse train $x_i'(n)$ for the ith critical band was derived from pulses in the unmasked pulse train $x_i(n)$ whose amplitudes were above the masking threshold calculated for each critical band.

A further masking effect occurs as a result of weaker signal components being rendered inaudible by stronger signal components in the same critical band that precede them in time. The masking threshold $Y_i(n)$ for this temporal post-masking decays approximately exponentially following each pulse, or neural firing [11]. A simple approximation to this masking threshold, introduced in this paper, is

$$Y_i(n) = \begin{cases} x_i'(n), & x_i'(n) > c_0\, Y_i(n-1) \\ c_0\, Y_i(n-1), & otherwise \end{cases}, \qquad (2)$$

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, as seen in Fig. 3.

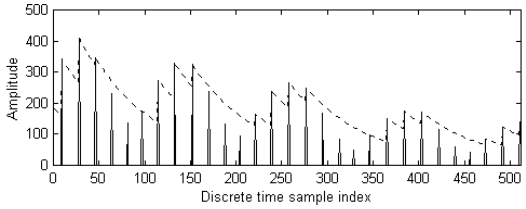

Figure 3. A frame of critical band pulses (solid), and corresponding temporal masking threshold (dashed) for filter number 8 (centre frequency 840 Hz).

The purpose of applying masking is to produce a more efficient and perceptually accurate parameterization of the firing pulses occurring in each band. Experiments revealed that simultaneous masking removed an average of around 10% of the pulses, while the application of temporal post-masking removed an average of 55% of the pulses (on top of the reduction due to simultaneous masking). This was achieved while maintaining transparent quality coded speech and audio.

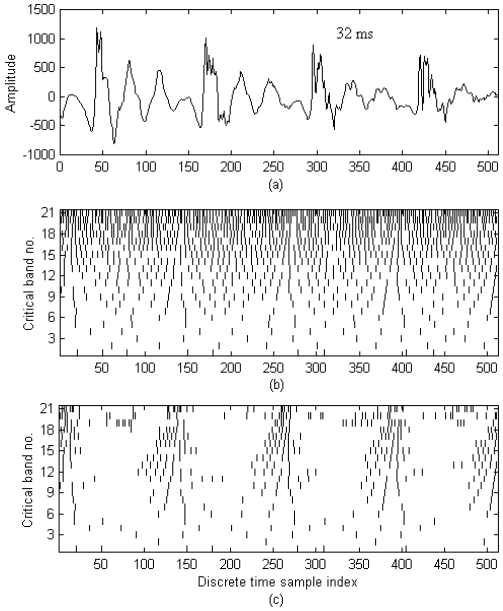

Figure 4. (a) A frame of 16 kHz sampled speech. (b) Critical band pulse train auditory domain representation after peak-picking and normalization. (c) Critical band pulse train auditory domain representation after peak-picking, simultaneous and temporal post-masking, and normalization.

The overall effect of simultaneous and temporal masking in reducing the number of pulses used to represent the input signal is illustrated in Fig. 4. This efficient auditory domain time-frequency representation characterizes both the spectral envelope and the timing information, providing acoustic parameters that are closely related to the speech or audio waveform in addition to the spectrum. It is therefore a promising new parameterization for automatic speech recognition, since Atal's [2] experiments show that spectral envelope parameters contain insufficient information to distinguish between various sounds in speech accurately.

The masked pulse train in each critical band 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.

3. SPEECH AND AUDIO SYNTHESIS

Reconstruction of the speech or audio signal from the critical band gains and pulse trains is achieved simply and at low complexity using the filter bank scheme shown in Fig. 5.

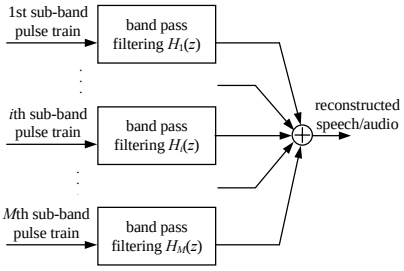

Figure 5. Speech/audio synthesis using linear phase gammatone filters (M = 21).

As described in [5], the critical band signals can be reconstructed from the pulse trains by means of bandpass filtering. Since linear phase analysis filters of equal lengths are employed in the analysis stage, the use of the same filters in the synthesis stage results in the correct overall phase characteristic. There is no requirement for perfect reconstruction in this scheme since the intermediate processing is highly non-linear.

The use of FIR analysis and synthesis filters of length 199 gives an overall delay of 12.4 ms at 16 kHz sampling frequency. A trade-off exists between the delay and the accuracy with which the lowest frequency filter impulse responses are modeled. Other possible analysis/synthesis filter bank implementations are considered in [6].

The inclusion of more than 21 analysis and synthesis filters in the signal bandwidth reduces the ripple in the reconstruction transfer function (across all bands), and therefore increases the output speech quality. In previous research [5], 20 filters were used for a 4 kHz signal bandwidth, producing a ripple of 2 dB. In our implementation, 21 filters were sufficient to ensure a ripple within 1.5 dB over the frequency range 25-7100 Hz, and this ripple was not perceptible during informal listening tests. The number of analysis/synthesis filters can be varied, however larger numbers of filters result in more pulses to code.

4. SPEECH AND AUDIO CODING

For each frame, the speech/audio signal parameters requiring coding are the gain of each critical band, the amplitude of each pulse, and the position of each pulse. The variety of different possibilities for coding each of these fundamental parameters means that scalable coding of any type of input signal can be achieved using a single parameterization. Different signal bandwidths can easily be accommodated by including or omitting the various critical band pulse trains as required. In this section, we consider an implementation of the novel analysis-synthesis techniques of section 2 and 3 for the coding of wideband speech and audio signals.

4.1 Critical Band Gains

Experiments employing non-uniform scalar quantization of the log critical band gains revealed that between 5 and 7 bits per sub-band per frame was sufficient to achieve good decoded signal quality. Higher frequency critical bands tended to require 7 bits to encode the gain, while for lower frequencies 5 or 6 bits per frame were sufficient. The more precise temporal resolution of the auditory system at higher frequencies is an important consideration when determining the frame length, and shorter frames have often been used in higher frequency sub-bands of existing wideband speech and audio coders. Our choice of frame lengths was based upon that of [4] and informal listening tests, and is shown in Table 1.

Table 1. Frame lengths for different critical bands

Critical bands	Centre freqs (Hz)	Frame length (ms)
1-9	50-1000	32
10-15	1170-2500	16
16-19	2900-4800	8
20-21	5800-7000	4

4.2 Pulse Amplitudes

Due to the half-wave rectification, all pulses have non-negative amplitude, and thus there is no need to code the signs of pulses. The pulse amplitudes in each critical band can be normalized by the critical band gain to form a pulse train with amplitudes typically clustered close to one. Kubin and Kleijn [5] found that forcing all normalized pulse amplitudes to have unity amplitude produced clearly audible distortion, preferring instead to code them using 1 bit per pulse.

Our experiments reveal that transparent quality reconstructed speech and audio is obtained by forcing all normalized pulse amplitudes to unity amplitude. We suggest that this is due in part to the removal of perceptually redundant pulses during the masking stage, and also due to an appropriate choice of frame lengths. Practically, this means that the normalized pulse amplitudes can be coded with zero bits per pulse, although their positions must be transmitted.

4.3 Pulse Positions

The quantization of the pulse positions is a significant issue. The application of an M channel filter bank to a signal frame of

length N produces MN samples to be coded. Peak picking reduces this to roughly $0.15 \times MN$, and masking reduces this to around $0.06 \times MN$, with no loss to the perceptual quality of the decoded speech. For $M = 21$ filters, thus the total number of pulses to be coded is approximately $1.26 \times N$. Avenues for further reducing the number of pulses are yet to be fully explored. This compares with previous work [5], in which the number of pulses to be coded was about $3.2 \times N$ for a signal bandwidth of 4 kHz.

In many audio coding applications [3], critical sub-sampling produces N samples to be quantized for a frame of length N. Unlike the scheme proposed in this paper, those applications do not preserve the timing information contained in the critical band signals, since the sub-sampling process effectively performs time averaging of the sub-band signals.

Experimental work on pulse position quantization revealed that in the higher frequency critical bands the pulse trains could be coded less precisely without detrimental effects on the decoded signal quality. Thus, the pulse trains of critical bands 10 to 21 (frequencies above 1.25 kHz) were split into 2 ms sub-frames and their positions were coded using vector quantization. Bit allocation across the critical bands varies monotonically from 6 bits per sub-frame in the 10^{th} band to 2 bits per sub-frame in the 21^{st} band. Vector quantization of this kind introduces a substantial increase in the required computation, however since the 2 ms pulse train vectors consist only of ones and zeros, an XOR operation can be used to simplify the distance measure.

In the lower frequency critical bands, there is a perceptual requirement to encode the pulse positions precisely, and thus arithmetic coding [8] is applied in these bands. This is feasible partly because in these bands, there is a sparse density of pulses, particularly as a result of the simultaneous and temporal masking process.

The overall average bit rate resulting from the quantization techniques outlined in sections 4.1 to 4.3 was 69.7 kbps, and the bit allocation is shown in Table 2.

Table 2. Average bit allocation for each 32 ms frame

Parameter	Bits per frame
Critical band gains	556
Pulse positions (bands 1 to 9)	842
Pulse positions (bands 10 to 21)	832
Total	2230

The coding scheme outlined in this section offers only one possibility for the quantization of the critical band gains and pulse trains. The parameterization described is highly scalable, and offers possibilities for high quality coding of speech at low bit rates through to 44.1 kHz audio. Note that in low bit rate speech coding, the encoding of the pulse positions would be based upon the pitch period rather than being precisely quantized.

5. CONCLUSION

A new speech and audio coding paradigm has been presented which employs a linear phase gammatone filter bank to derive a perceptually accurate, yet compact, time-frequency parameterization. The use of masking in particular is essential to reducing the quantity of information required to parameterize the input signal with this technique. The simplicity of the analysis and synthesis algorithms allows for very computationally efficient realizations, and the ease with which different signal types and bandwidths can be represented offers possibilities for integrated, scalable speech and audio coding. Informal listening tests indicate that this coding paradigm provides transparent reconstructed speech and audio signals. Future research will concentrate on further reduction of the number of pulses required to parameterize speech and audio signals, thereby decreasing the bit rate.

6. ACKNOWLEDGMENTS

The authors wish to thank Dr. P. Mulhern, Head of Research and Development, Athlone Institute of Technology, Ireland, for providing the funding and Dr Andy Davis, British Telecom laboratories, Ipswich, UK for making BT speech files available for the experiments. This research was also partly funded by the Australian Research Council Small Grant Scheme, Australia.

7. REFERENCES

[1] 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.

[2] Atal, B. S., "Automatic speech recognition: A communication perspective", in *Proc. IEEE ICASSP*, 1999, pp. 205-208.

[3] Brandenburg, K. B., and Stoll, G., "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio", *J. Audio Eng. Soc.*, vol. 42, no. 10, Oct. 1994, pp. 780-792.

[4] Krasner, M. A., "The critical band coder – digital encoding of speech signals based on the perceptual requirements of the auditory system", in *Proc. IEEE ICASSP*, 1980, pp. 327-330.

[5] Kubin, G., and Kleijn, W. B., "On speech coding in a perceptual domain", in *Proc. IEEE ICASSP*, 1999, pp. 205-208.

[6] Lin, L., Holmes, W. H., Ambikairajah, E., and "Auditory filter bank inversion", in *Proc. IEEE ISCAS*, May 2001.

[7] Lin, X., Hanzo, L., Steele, R., Webb, W. T., "Subband-multipulse digital Audio broadcasting for mobile receivers", *IEEE Trans. on Broadcasting*, vol. 39, no. 4, December 1993, pp. 373-382.

[8] Rubin, F., "Arithmetic stream coding using fixed precision registers", *IEEE Trans Inf. Theory*, vol. IT-25, no. 6, 1979, pp. 672-675.

[9] Slaney, M., Naar, D., and Lyon, R. F., "Auditory model inversion for sound separation", in *Proc. IEEE ICASSP*, 1994, vol. II, pp. 77-80.

[10] Zwicker, E., and Terhardt, E. "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. Am.*, vol. 68, no. 2, November 1980, pp. 1523-1525.

[11] 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, March 1991, pp. 115-126.