Auditory modelling for speech processing in the perceptual domain

L. Lin* E. Ambikairajah† W. H. Holmes‡

(Received 8 August 2003; revised 28 January 2004)

Abstract

The human hearing system is the most robust speech processor despite noisy environments. This work presents a new computational model for our auditory system by exploring the psychoacoustical masking properties. The model is then applied to speech coding in the perceptual domain. The coding algorithm is capable of producing high quality coded speech and audio, which account for temporal as well as spectral details. The proposed filterbank is also applied to speech denoising in the perceptual domain. The enhanced speech is of good perceptual quality.

*School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, AUSTRALIA. mailto:ll.lin@ee.unsw.edu.au

†University of New South Wales

‡University of New South Wales

See http://anziamj.austms.org.au/V45/CTAC2003/Lin2/home.html for this article, Published September 1, 2004. ISSN 1446-8735

Contents

1 Introduction

When our ear is excited by an input stimulus, different regions of the basilar membrane respond maximally to different frequencies, that is, a frequency tuning occurs along the membrane. We can therefore think of the response patterns as due to a bank of cochlea filters along the basilar membrane. Adequate modelling of the principal behaviour of the peripheral auditory systems is a very difficult problem. Earlier models used transmission line representations to simulate basilar motion [6]. Recently parallel auditory filterbanks such as the Gammatone filters [7], have become very popular as a reasonably accurate alternative for auditory filtering. A parallel auditory filterbank is easily inverted and hence has applications in auditory-based speech and audio processing. In this work we present a new parallel auditory filterbank on the critical band scale. The filterbank models psychoacoustic tuning curves obtained from the well known masking curves.

Current applications of speech and audio coding algorithms include cellular and personal communications, teleconferencing, secure communications.

Low bit rate speech coders provide impressive performance above 4 kbps for speech signals. But do not perform well on musical signals. Similarly, transform coders perform well for music signals, but not for speech signals at lower bit rates. There is therefore a need for high quality coders that work equally well with either speech or general audio signals. In this work we propose a scheme for a universal coder based on an auditory filterbank model that handles both wide band speech and audio signals.

Speech noise reduction is a very important research field with applications in many areas such as voice communication and automatic speech recognition. The most popular methods, with many variants, are Wiener filtering and spectral subtraction [4]. Although these methods reduce the noise, they also reduces speech power and hence introduce speech distortion. In this work we propose a denoising technique based on an auditory filterbank and a new perceptual modification of Wiener filtering. Speech distortion is reduced and speech intelligibility is improved.

2 A critical band scale auditory filterbank

This section presents a parallel auditory filterbank model that matches psychoacoustical tuning curves. The tuning curves are obtained by exploring the relation between auditory masking and tuning curves and the similarity of the masking curves in the critical band scale. Details are described by Lin, Ambikairajah and Holmes [5]. The transfer function of the critical-band auditory filterbank that models the psychoacoustical tuning curves is developed in the z-domain [5]:

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

where $f_s = 16\,\text{kHz}$ is the sampling frequency, and the parameters

$$f_A = \sqrt{f_c^2 + B_w^2} \quad \text{and} \quad r_A = e^{-2\pi B_w/f_s}\,.$$

The parameter B_w is calculated using the formula in [8]:

$$B_w = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69}\,,$$

$$Z_c = 13\arctan(0.76 f_c/1000) + 3.5\arctan(f_c/7500)^2\,,$$

where Z_c is the corresponding critical band rate of f_c. The parameters r_0 and r_B are chosen as $r_0 = 0.955$ and $r_B = 0.985$. We use the following empirical formula to choose f_B:

$$f_B = 117.5(f_c/1000)^2 + 1135.5(f_c/1000) + 277.0\,.$$

The frequency response of the 21 critical band auditory filters in the frequency range of 0 to 8 kHz is shown in Figure 1 by the dashed lines.

The proposed critical-band auditory filterbank is also approximately power-complementary. That is,

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

where C is a constant and $G_i(e^{j\omega})$ 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 filter as

$$h_i(n) = g_i(-n) \quad \text{for} \quad i = 1, \ldots, M\,, \tag{3}$$

then the synthesis filterbank is implemented using FIR filters obtained by time-reversal of the impulse responses of the corresponding analysis filters. The signal reconstruction is nearly perfect, that is, $\sum_{i=1}^{M} g_i(n) * h_i(n) \approx C\delta(n)$. Figure 1 shows the overall analysis/synthesis frequency response by the solid line. It resembles the frequency response of an all-pass filter. The implementation of the analysis/synthesis filterbank scheme is shown in Figure 2. Each analysis filter is implemented as an IIR filter with 8 poles and 3 zeros. Each synthesis filter is implemented as a FIR filter with 128 coefficients. An 8 ms delay is required to make the filter causal if $f_s = 16$ kHz. Between the analysis and synthesis sections is the processing block that carries out speech coding or denoising algorithms, which is described next.

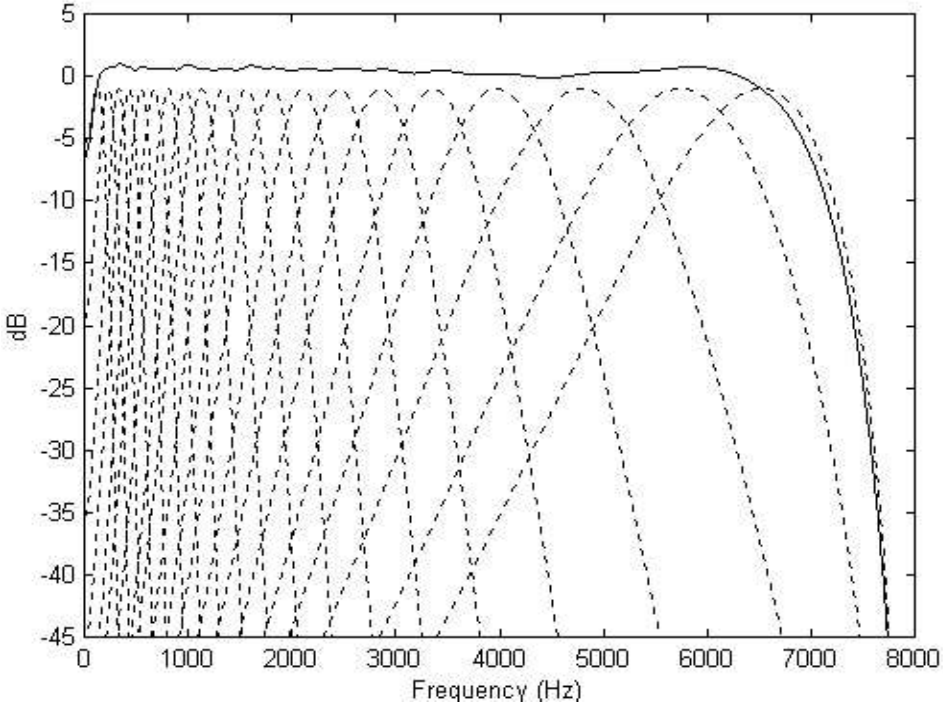

FIGURE 1: Frequency response of the auditory filterbank; dashed: analysis filters, solid: overall analysis/synthesis response.

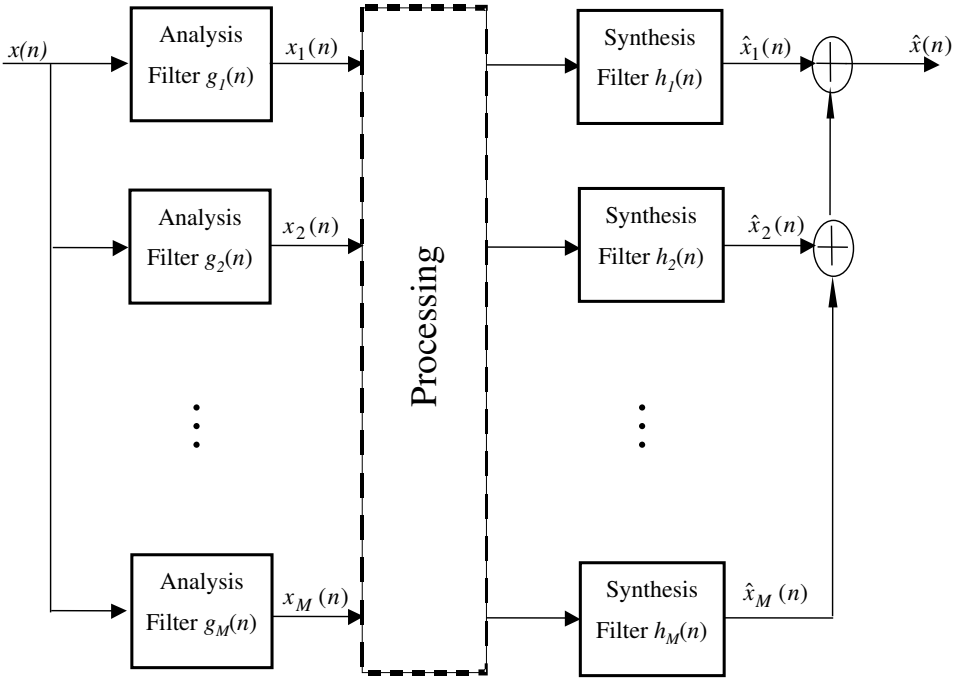

FIGURE 2: Speech processing based on an auditory filterbank.

3 Application of an auditory filterbank to speech processing

3.1 Speech coding using an auditory filterbank

The first step of the coding scheme is to filter the speech/audio signal by the critical-band analysis filters $g_i(n)$. The output of each filter, $x_i(n)$, is then half-wave rectified, and the positive peaks of the critical band signals are 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 [2, 3]. 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. Figure 3 shows, using spikes, a sequence of such pulses for the critical band centred at 1 kHz.

The masking properties of human auditory system are applied to eliminate redundant pulses. Because 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 is employed. Weak signal components become inaudible by the presence of stronger signal components in the same critical band that precede or follow them in time, and this is called temporal masking. When the signal precedes the masker in time, it is called pre-masking; when the signal follows the masker in time, the condition is called post-masking [1, 9, 10]. A strong signal can mask a weaker signal that occurs after it and a weaker signal that occurs before it. Both temporal pre-masking and temporal post-masking are employed in this work to reduce the number of pulses. Figure 3 shows an example of post-masking with the masking thresholds shown using the dashed line. All pulses with amplitudes less than the masking threshold are discarded. The darkened spikes are the pulses to be kept after applying post-masking.

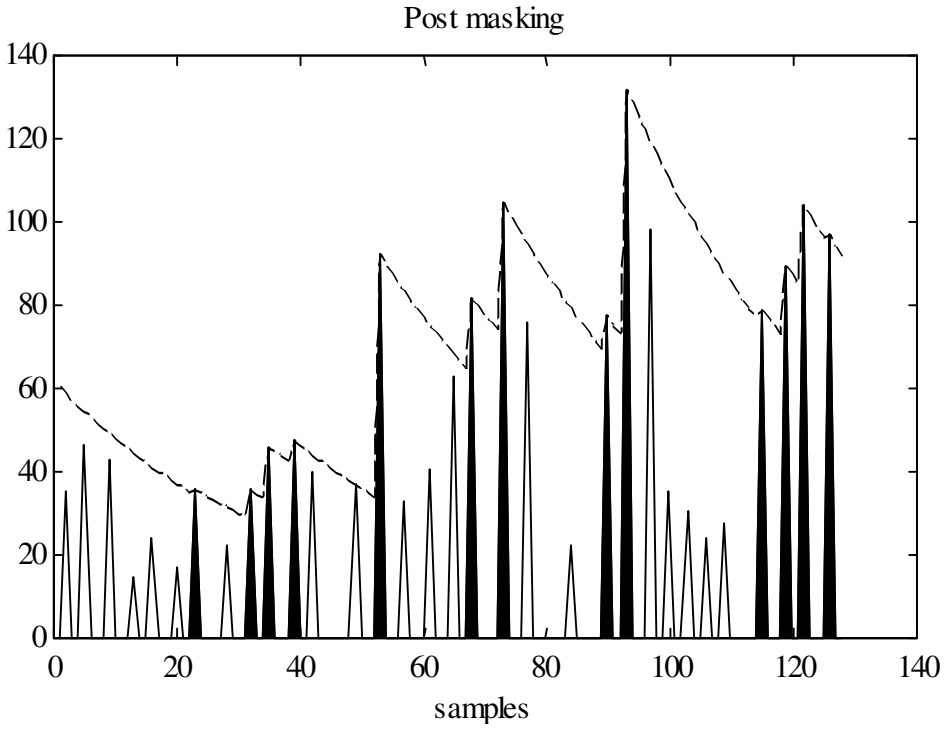

FIGURE 3: Pulse reduction using post-masking; solid lines: pulses, dashed lines: thresholds (centre frequency 1 kHz).

The upper panel in Figure 4 shows the pulses locations of 21 channels obtained at the stage of peak-picking. The lower panel in Figure 4 shows the pulses retained after applying auditory masking. The purpose of applying masking is to produce a more efficient and perceptually accurate parameterization of the firing pulses occurring in each band.

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. For each frame, the signal parameters requiring for coding are the gains of the critical bands and the amplitudes and positions of the pulses. Each critical band gain is quantized to 6 bits and the amplitude of each pulse is quantized to 1 bit. The pulse positions are coded using a new run-length coding technique. The overall average bit rate resulting from this coding scheme is 58 kbps.

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 $h_i(n)$. Summing the outputs from all filters results in the reconstructed speech or audio signal, which is perceptually the same as the original. The lower panel in Figure 5 shows one frame of the resynthesised speech based on the decoded pulse trains. The corresponding original speech is shown in the upper panel of Figure 5. The duration of the speech frame is 32 ms (512 samples for $f_s = 16$ kHz).

The advantage of this coder is that it works equally well with either speech or general audio signals, is highly scalable, and is of moderate complexity. Further research is required to examine the statistical correlation and redundancy among the pulses, and investigate the use of Huffman coding or arithmetic coding techniques to reduce the bit rate further.

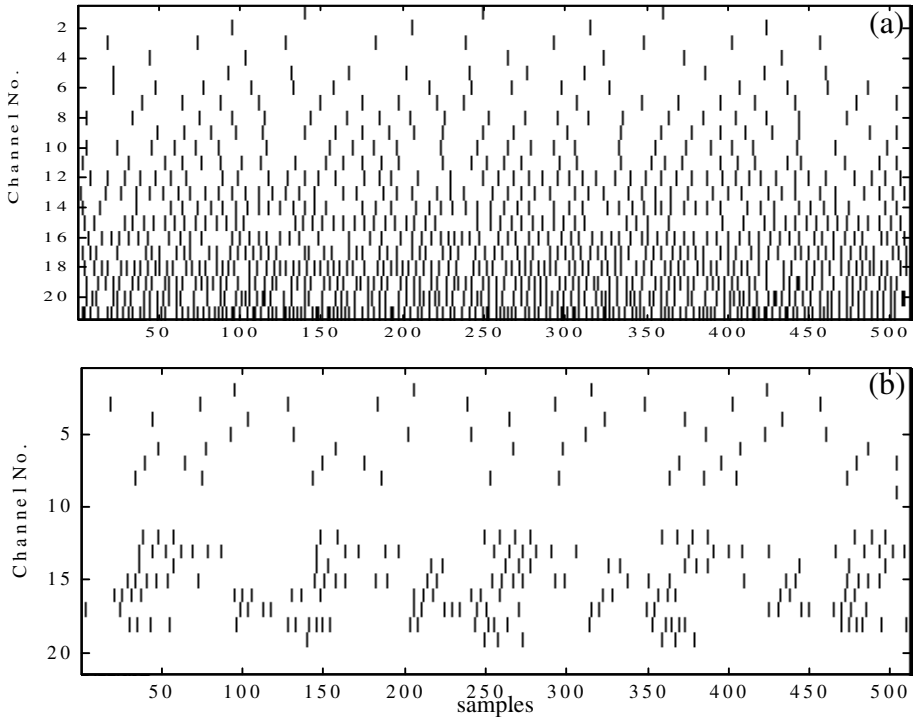

FIGURE 4: Pulse trains of 21 critical bands; (a) before auditory masking, (b) after auditory masking.

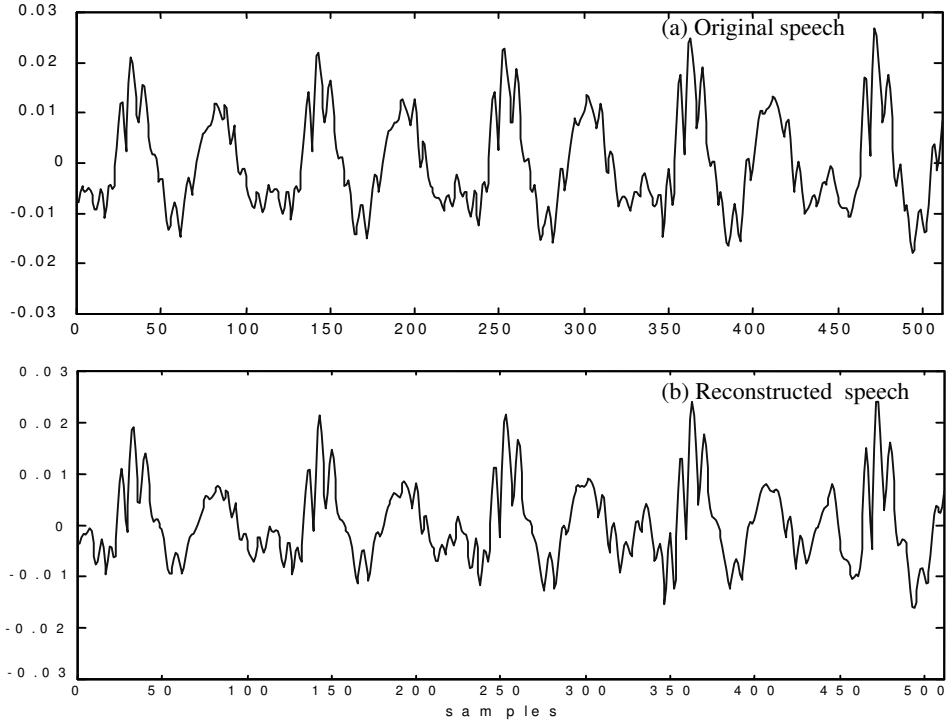

FIGURE 5: A frame of the original speech and its reconstruction.

3.2 Speech denoising using an auditory filterbank

Assume that the input speech to the filterbank is corrupted by additive noise; that is, $x(n) = s(n) + w(n)$, where $s(n)$ is the clean speech and $w(n)$ is the additive noise. Both $s(n)$ and $w(n)$ are assumed zero-mean and uncorrelated.

The first part of our speech denoising scheme is to decompose the noisy speech $x(n)$ into noisy critical band signal (Figure 2):

$$x_i(n) = g_i(n) * x(n) = s_i(n) + w_i(n)\,, \tag{4}$$

where $s_i(n) = g_i(n) * s(n)$ is the output from the ith critical band filter when the input to the filterbank is the clean speech only, and $w_i(n) = h_i(n) * w(n)$ is the corresponding output when the input is the noise only. Both signals, $s_i(n)$ and $w_i(n)$, are zero-mean and uncorrelated, since each auditory filter is a narrow bandpass filter and the clean speech $s(n)$ and the noise $w(n)$ are uncorrelated. Then the denoised subband signal is

$$\hat{s}_i = K_i x_i(n)\,, \tag{5}$$

where the $K_i (i = 1, \ldots, M)$ are the denoising gains to be determined.

Define $\sigma_{s_i}^2 = E\{s_i^2(n)\}$ and $\sigma_{w_i}^2 = E\{w_i^2(n)\}$. The denoising gain K_i is obtained by minimising

$$J_i = (K_i - 1)^2 \sigma_{s_i}^2 + \mu K_i^2 \max\{\sigma_{w_i}^2 - T_i, 0\}\,. \tag{6}$$

The first part of the above equation $(K_i - 1)^2 \sigma_{s_i}^2$ represents the speech distortion due to denoising; the second part $K_i^2 \max\{\sigma_{w_i}^2 - T_i, 0\}$ represents the noise residual. The parameter μ allows a trade-off between signal distortion and noise: if μ is large the noise is reduced, but there is greater signal distortion. T_i is the estimated masking threshold due to the speech signal. The noise is included in this perceptual criterion only if it exceeds the masking threshold. The denoising gain is then

$$K_i = \frac{\sigma_{s_i}^2}{\sigma_{s_i}^2 + \mu \max\{\sigma_{w_i}^2 - T_i, 0\}}\,. \tag{7}$$

When the noise $\sigma_{w_i}^2$ is under the masking threshold T_i, the gain K_i will always be 1. The gain decreases as the noise exceeds this level, but it will always be larger than the optimum solutions to the conventional Wiener problems [4]. The speech distortion is always smaller than achieved with the Wiener solution (that is, if masking is not allowed for). The noise residual is always larger than with the Wiener solution, but the difference will not be audible due to auditory masking effects.

The synthesis process starts with filtering $\hat{s}_i(n)$ by the corresponding FIR synthesis filter $h_i(n)$. Summing the outputs from all filters results in the denoised speech.

The proposed denoising technique is tested on a variety of noises including pink noise, car noise and tank noise. Informal listening demonstrates that the perceptually modified Wiener filter gives denoised speech with more intelligibility than the traditional Wiener filter. An example of speech denoising with car noise of signal-to-noise ratio of 5 dB is shown in Figures 6 and 7. See the clean, noisy and denoised sentences plotted in Figure 6. The denoising gains obtained using the perceptual Wiener filtering in two channels are shown by the solid lines and the conventional Wiener filtering gains are shown by the dashed lines in Figure 7. See that the gain resulted from the proposed denoising approach is always higher than the gain from the conventional Wiener filter and hence speech distortion is reduced.

4 Conclusions

We present a new parallel auditory filterbank that models the psychoacoustical tuning curves. The model is applied to speech coding and speech denoising in the perceptual domain. The decomposition of speech signal into critical band signals enables easy application of auditory masking properties to reduce bit rate in coding and speech distortion in denoising. The auditory-system-based coding paradigm produces high quality coded speech or audio,

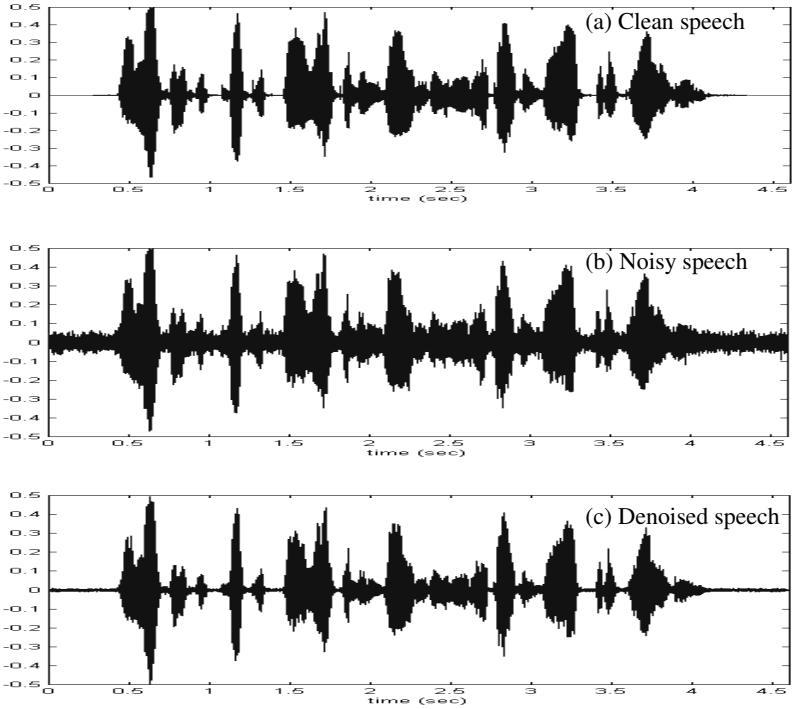

FIGURE 6: Clean, noisy and denoised speech sentences.

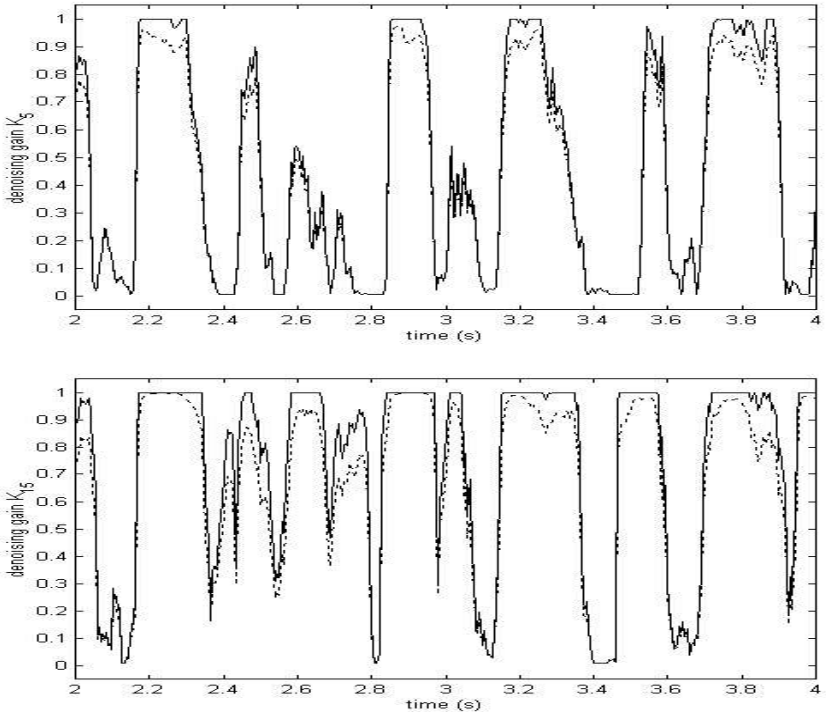

FIGURE 7: Denoising gains for channels 5 and 15; solid: perceptual Wiener filtering, dotted: conventional Wiener filtering.

is highly scalable, and is of moderate complexity. The perceptually modified Wiener filter results in denoised speech with more improved intelligibility and less speech distortion than the conventional Wiener filter.

References

[1] E. Ambikairajah, A. G. Davis and W. T. K. Wong. Auditory masking and MPEG-1 audio compression. *Electr. & Commun. Eng. Journal*, 9(4):165–197. C970

[2] E. Ambikairajah, J. Epps and L. Lin. Wideband speech and audio coding using Gammatone filter banks. *Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing*, pages 773–776, 2001. C970

[3] G. Kubin and W. B. Kleijn. On speech coding in a perceptual domain. *Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing*, pages 205–208, 1999. C970

[4] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, 1979. C966, C976

[5] L. Lin, E. Ambikairajah and W. H. Holmes. Auditory filterbank design using masking curves. *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 411–414, 2001. C966

[6] R. F. Lyon. A computational model of filtering detection and compression in the cochlea. *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, pages 1282–1285, 1982. C965

[7] R. D. Patterson, M. Allerhand and C. Giguere. Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform. *J. Acoust. Soc. Am.*, 98:1890–1894, 1995. C965

[8] E. Zwicker and E. Terhardt. Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, 68:1523–1525, 1980. C967

[9] E. Zwicker and U. T. Zwicker. Audio engineering and psychoacoustics: matching signals to the final receiver, the human auditory system. *J. Audio Eng. Soc.*, 39(3):115–125, 1991. C970

[10] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models.* Springer-Verlag, 1999. C970