ON SPEECH CODING IN A PERCEPTUAL DOMAIN

*Gernot Kubin**

Vienna University of Technology
Gusshausstrasse 25/389
A-1040 Vienna, Austria

W. Bastiaan Kleijn

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
100 44 Stockholm, Sweden

ABSTRACT

For speech coders which fall within the class of waveform coders, the reconstructed signal approaches the original with increasing bit rate. In such coders, the distortion criterion generally operates on the speech signal or a signal obtained by adaptive linear filtering of the speech signal. To satisfy computational and delay constraints, the distortion criterion must be reduced to a very simple approximation of the auditory system. This drawback of conventional approaches motivates a new speech coding paradigm in which the coding is performed in a domain where the single-letter squared-error criterion forms an accurate representation of perception. The new paradigm requires a model of the auditory periphery which is accurate, can be be inverted with relatively low computational effort, and which represents the signal with relatively few parameters. In this paper we develop such a model of the auditory periphery and discuss its suitability for speech coding. Our results indicate that the new paradigm in general and our auditory model in particular form a promising basis for the coding of both speech and audio at low bit rates.

1. INTRODUCTION

Assuming the standard squared error criterion, it is well-known from source coding theory that scalar quantizers and vector quantizers perform best in terms of distortion versus rate when the scalars or signal blocks to be quantized are independent. As a result, many models which are used for waveform coding of speech and audio signals, are essentially designed to represent the speech signal in terms of one or more sample sequences in which the samples within a sequence as well as between the sequences are close to independence according to some measure. Often, this process takes the form of linear decorrelation by means of methods such as linear prediction and the Karhunen-Loève transform.

The linear-decorrelation strategy is easily motivated if the overall distortion criterion for the reconstructed speech signal can be expressed as a sum of sample distortions, i.e. if the criterion is a *single-letter* criterion. Unfortunately, proper hearing models operating in the speech domain do not satisfy this requirement. Hearing-based distortion criteria have significant co-dependencies between samples, and thus also between adjacent signal blocks. As a result, optimal performance cannot be obtained by quantizing the blocks in their sequential time order. Particularly with the relative small data blocks used in waveform coders such as CELP (which use subframes from 2.5 to 10 ms in duration) this co-dependency of the distortion reduces performance. This has been a major incentive to use simple criteria with less co-dependency in such coders. An example of the significant performance penalty associated therewith is the accurate coding of the phase spectrum of the pitch cycle of a high-pitched speaker in a CELP coder; it is well-known from sinusoidal coding techniques that such an accurate phase description for high pitched speech is irrelevant to a human listener.

In this paper, we propose a straightforward solution to the forementioned problem of criterion co-dependency of variables or vectors which are to be quantized: we first transform the signal to a representation where co-dependency of the distortion criterion on the representation variables vanishes. In other words, we transform the signal into a domain where a single-letter criterion, such as the squared error criterion is an accurate representation of distortion. Of course, a major constraint on the auditory model is that it must be invertible, to allow the reconstruction of the speech signal at the receiver. We use a physiologically-motivated hearing model to obtain such a mapping.

2. AN INVERTIBLE AUDITORY MODEL

The purpose of our auditory model is to obtain a speech signal representation which facilitates speech coding. Thus, we used the following guidelines in the design of our auditory model:

1. The single-letter squared-error distortion criterion must be an accurate description of distortion as perceived by the human auditory system.
2. A perceptually accurate inversion of the model must be possible at low computational complexity.

Auditory models used in speech and audio coding have usually not been aimed at obtaining a single-letter criterion. However, particularly audio coding algorithms have carefully exploited knowledge about the human auditory system. It is common in audio coders to have as a first processing stage a filterbank with a structure motivated by that of the human auditory periphery. Additional auditory knowledge is exploited in the form of distortion criteria which make the distortion co-dependent on the representation variables. Our approach differs from such audio coding algorithms in that we aim to use a more sophisticated transform in combination with a simple single-letter distortion criterion. Thus, our approach should, in principle, have an advantage in the trade-off made between computational effort, delay, and accuracy.

Physiologically plausible invertible auditory model have been used before for purposes other than speech or audio coding [1]. The aim of various invertible auditory models has been to understand perception [2, 3, 4], to test the accuracy of the auditory model [5, 6], and to enhance speech [4]. The inversion process for the more recent models is iterative in nature [3, 4, 6].

Some of the inversion procedures are based on the framework of projections onto convex sets [7], where the constraints on the reconstructed signal are specified in the form of convex sets. Iterative projections onto these sets converge to a point on the intersection

*The first author performed his research as a visiting scientist at KTH.

of the sets. For coding purposes, the method of convex projections suffers from several disadvantages: *i*) the iterative nature of the algorithm leads to a high computational load, and *ii*) if not all constraints can be formulated as convex sets, then convergence cannot be guaranteed. We should also mention the iterative inversion procedures developed for spectrograms [8], which are closely related to the cochleagram auditory representation. These inversion methods are guaranteed to converge to a locally optimal solution, but again require a high computational effort.

2.1. The Auditory Model

Our source coding system is outlined in figure 1. The early processing stages are similar to most other physiologically-motivated auditory models. The first analysis stage is a filterbank which simulates the motion of the basilar membrane. We use the well-known gamma-tone filterbank [9] for this purpose. A 20-channel implementation of this filterbank is shown in figure 2. The filterbank is followed by a half-wave rectifier and a power-law compression, simulating the behavior of the inner hair cell. The input, $x(n)$, and output, $y(n)$ of the power-law compression operator are related by

$$y(n) = x(n)^b. \tag{1}$$

In our work we use $b = 0.4$.

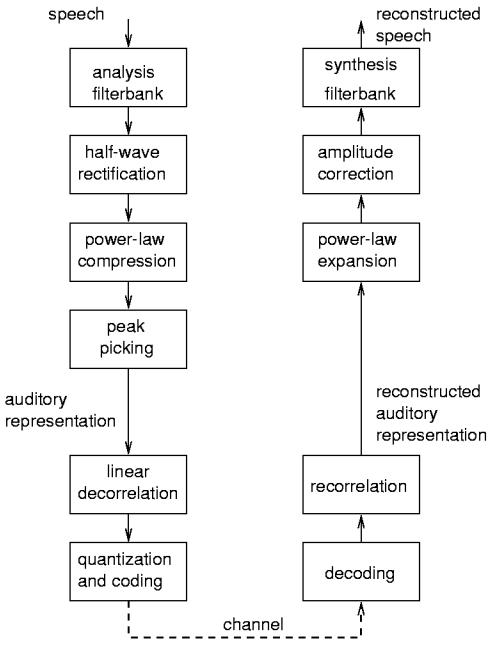

Figure 1: The proposed coding system.

In the processing stage following the power-law compression, we want to simplify the data representation to facilitate coding, without removing information essential to perceptually accurate reconstruction. In particular, it is important that the model preserves information about the local cyclostationarity of the signal, i.e. that the model preserves the differences between voiced speech (nearly periodic) sounds and unvoiced (wide band) sounds. Information about this fine structure of the speech is lost if time-averaging is applied to the processed filterbank outputs, as is common to many auditory models (e.g. [6, 10]). Time-averaging is useful for speech recognition purposes [6, 11], but leads to low reconstruction quality upon inversion [6].

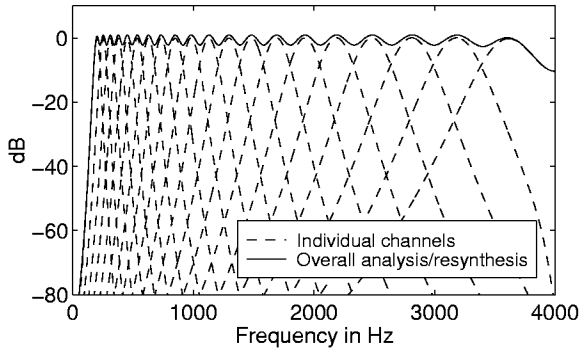

Figure 2: Transfer functions for our 20-channel gamma-tone filterbank. The reconstruction transfer function is also shown, prior to equalization.

In our model, the power-law compressor is followed by an adaptive sampling mechanism simulating the firing behavior of auditory neurons. Our model provides the neurons with a physiologically plausible behavior and simplifies the data structure. Each neuron has a *state* which decays exponentially with a characteristic decay time τ and which is reset when it fires. The reset level is dependent on the input level during firing. The firing probability of a neuron per unit time is a monotonically increasing function of the difference between the neuron input and its state.

The functionality of the neuron model can be explained with a sine-wave input with a period $p << \tau$. Each channel of the filterbank is associated with an ensemble of neurons. The behavior of our neuron model leads to a high rate of neuron firings at the peak of the input sine wave (ignoring processing delays) and low firing rates elsewhere. In other words, clusters of high firing rates are *synchronized* with the signal period. This synchronicity of neuron firings with the input signal is present even for neurons corresponding to filters in the auditory filterbank which have their best frequency (frequency location of largest filter gain) relatively far from the sine wave frequency. The probabilistic nature of the firing of an individual neuron means that the firing of the ensemble has a certain time resolution. Since synchronization is known to occur until about 4000-5000 Hz, it is reasonable to postulate the time localization of the firing clusters of the neuron ensemble to be of the order of 0.2 ms.

To make the representation of the half-wave rectified signal complete, i.e. to make reconstruction possible, we need to preserve not only the location of the peaks, but also their amplitude. In terms of our neuron model, this represents the firing rate, i.e. the number of neurons of the ensemble which fires at the peak location. Thus, our neuron ensembles essentially sample the compressed and half-wave rectified filter outputs at the signal peaks. In the following, we will refer to the firing clusters as *firing pulses*, which have both a *time location* and *amplitude*.

The nature of the neuron firing for the lower frequency bands allows a simple peak-picking implementation for an 8 kHz sampling rate. The peak-picking operator output, $w(n)$, is given by

$$w(n) = \begin{cases} y(n), & y(n) > y(n-1) \wedge y(n) > y(n+1) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The peak-picking procedure resembles the pulse-ribbon model of Patterson [12]. However, in contrast to this earlier model, we preserve an amplitude in addition to a firing cluster location.

Next, we consider the behavior of our model for signal characteristics common to speech. We start with the harmonic set of sine waves, which mimics voiced speech. For each hair cell, the firing cluster occurs when the sinusoid nearest in frequency to the peak of the corresponding filter is at its maximum. Thus, the neuron firings are synchronous (phase-locked) with the pitch period or submultiples of the pitch period. The neuron firings are aligned across the frequency channels, except for the delay of the filters of the gamma-tone filterbank, which decreases with increasing frequency. This is shown in figure 3. For the case of wide-band noise, the neuron firings are nearly regularly spaced in time, a result of the band-limitation of the channel. However, the irregularity of these pulse spacings is very clear upon autocorrelation or Fourier transform of the pulse train of firing pulses for a channel.

It is natural to ask if we achieved the goal of a single-point distortion criterion. Assuming that the auditory model used is correct (and not trivial), our procedure does reach this goal whereas conventional coders do not. Our confidence in the accuracy of our model is based on published experimental results for the components we use (e.g., [9, 12]) and for related models as well as the experimental results described in section 4.

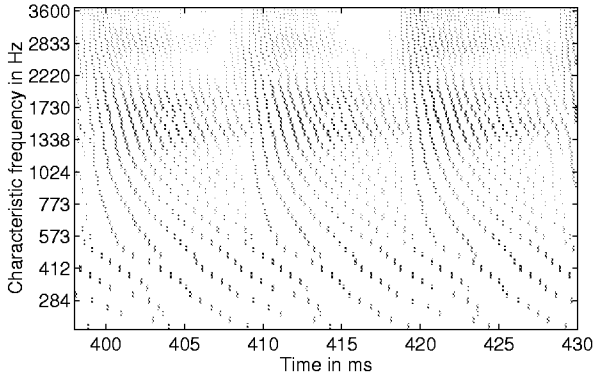

Figure 3: The perceptual domain (50 channels) for the sound [I] in "there **is**", spoken by a male. Three pitch periods are shown with the pulse alignment and formant structure are clearly visible.

2.2. The Inversion Mechanism

The first step in our inversion procedure is a power-law expansion on the firing pulses. We now have, for each channel, the positive peaks of the original signal. Each channel signal consists of mostly zeros and the power-law expanded firing pulses. The channel signals approximate the situation where a signal is downsampled and then upsampled by means of inserting zeros. This insertion of the zeros leads to aliasing which can be removed by means of bandpass filtering. This reasoning immediately suggests that bandpass filtering each (expanded) train of firing pulse should lead to a good approximation of the outputs of the analysis filter bank. Thus, application of the synthesis filterbank to the train of firing pulses and adding these signals should lead to a good reconstruction at low computational complexity. This is confirmed experimentally in section 4. We stress that the *pulse-train character of our auditory representation is fundamental to the fast reconstruction process*, allowing reconstruction without resorting to iterative methods.

For the model inversion to work, we must include proper normalization such that the the signal power in each channel is correct. Thus, prior to filtering, we compensate for *i*) the firing rate and *ii*) the finite sampling rate of the channel signals. The analysis filterbank outputs resemble sinusoids. If the sinusoids have a period of P, then the peak-picking procedure reduces the amplitude, when observed only within the band corresponding to the channel by a factor P. The average per-cycle maximum amplitude of a sampled sinusoid, α, for the case of a unity amplitude sine wave and a unity sampling period is

$$\alpha = \int_{-1/2}^{1/2} \cos(\frac{2\pi t}{P})dt = \frac{P}{\pi}\sin(\frac{\pi}{P}). \tag{3}$$

Thus, compensating for the finite sampling rate and the subsampling effect resulting from the peak picking gives an overall amplitude correction factor of $\pi/\sin(\pi/P)$.

The definition of the inverse filterbank is not entirely straightforward since the analysis filterbank is not orthogonal. Let K be the number of channels, and $H_k(z)$, $k = 1, \cdots, K$ the transfer function of the individual analysis filters. Using a synthesis filterbank with the transfer function

$$G_k(z) = \frac{H_k^*(z)}{\sum_i H_k(z)H_k^*(z)} \tag{4}$$

for the individual filters will lead to exact reconstruction. Thus, for the case that $\sum_i H_k(z)H_k^*(z) = 1$, the synthesis filterbank is just the analysis filterbank with time-reversed impulse response. (A time delay is needed to make the filterbank causal.) In the general case, accurate signal reconstruction can be obtained with a linear-phase equalization filter which accounts for the denominator of equation 4. Synthesis filter banks without equalization are commonly used [4]. We found that for 20 channels this results in a 2 dB ripple (shown in figure 2). Such a ripple is almost inaudible and decreases with increasing number of channels.

3. CODING ASPECTS

Our perceptual representation is a relatively sparse representation. However, it has more pulses than the original speech signal has samples. This increase in data is associated with a strong interdependency of the pulse amplitudes, and the increased effectiveness of a single-letter distortion criterion. To facilitate quantization with scalar or low-dimensional vector quantizers, the strong interdependency of the pulses, i.e. the redundancy, in the perceptual domain must be reduced first. A strong constraint applies on this redundancy reduction operator: it should not change the single-letter character of the distortion criterion.

A significant part of the redundancy in our perceptual representation is the within-channel redundancy. The information residing in each channel, i.e. in each firing pulse train can be separated into two aspects *i*) the amplitude of the pulses (representing the modulation amplitude of the channel) and *ii*) the spacing of the pulses which represents the spectral fine-structure (tone-like or wideband) in the channel. The Karhunen-Loève transform and closed-loop prediction in combination with scalar quantization preserve the single-letter distortion criterion and generally reduce redundancy. The discrete Fourier transform and particularly the discrete-cosine transform are often used as satisfactory approximations to the Karhunen-Loève transform.

The Fourier transform on each pulse train channel results in a two-dimensional representation (channel index versus frequency) which displays the *modulation spectrum* [13]. This is related to the syllabic articulation rate at low modulation frequencies and to the level of signal periodicity at higher modulation frequencies. Further decorrelation can be obtained with transformations along the channel axis. All these transformations preserve the squared-error criterion.

4. EXPERIMENTAL RESULTS

In this section we describe the reconstruction quality of our auditory model with and without quantization noise. For simple analysis/synthesis, we found that the quality of the reconstructed signal increases with increasing number of channels. However, at about 20 channels the perceived quality of the reconstruction has converged to that of the original signal for both speech and audio signals. The only audible difference is that some of the low frequencies are missing, which is a direct result of our choice of the filterbank configuration (see figure 2). A comparison of waveforms is shown in figure 4. Reconstruction from the pulse-train representation obtained with the peak-picking procedure results in a reconstructed signal with a segmental signal-to-noise ratio of 20 to 25 dB.

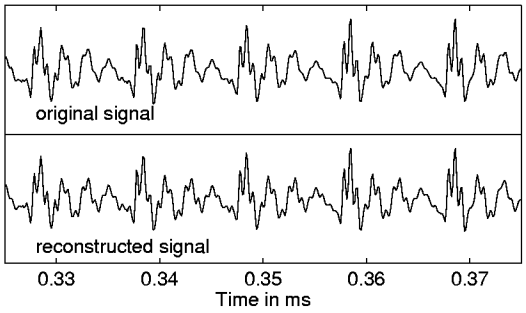

Figure 4: The reconstruction accuracy of the auditory model.

It is useful to examine the overall firing pulse rate prior to the quantization experiment. For both speech and audio signals, our experiments show that the average pulse rate of each of the channels approximates its best frequency. For the 20-channel case and for a real speech signal, the first channel has a pulse rate of about 200 Hz and the 20th channel has a pulse rate of about 3300 Hz. The average pulse rate is about 1300 Hz. Compared to the sampled speech signal, the perceptual-domain representation has about 3.2 times more pulses than there are samples in the original signal.

It seems reasonable to assume that the human auditory system is tolerant of internally generated noise, particularly given the random nature of neuron firings. This implies that an accurate model of the auditory system should be robust against noise introduced in any of the auditory representations. To check the robustness of the firing-pulse representation, we applied a simple quantization scheme to the pulse amplitudes. We estimate the maximum pulse amplitude in each channel once per 20 ms. Using this maximum value, and the fact that the signal is positive, a scalar quantizer is used to quantize the pulse amplitudes in the 20 ms frame. Surprisingly, we found that the reconstructed signal was essentially transparent for both speech and audio when we quantized the firing pulse amplitudes with only 1 bit per pulse (we used amplitudes 0.4 and 0.8 of the local maximum). However, using zero-bit quantization (fixed pulse amplitude within each channel for each frame) results in clearly audible distortion. We conclude that the firing-pulse representation is robust to amplitude modifications.

5. CONCLUSION

We have argued that it is deleterious for coding efficiency to perform quantization of parameters or parameter vectors independently when their distortion criteria are co-dependent. To avoid this problem, the co-dependency of the distortion must be removed first by transforming the signal into a perceptual domain.

While the distortion co-dependency is removed by a transformation into the perceptual domain, the resulting representation generally has a significant amount of redundancy. Much of this redundancy can be removed by means of linear decorrelation techniques such as Fourier transforms which preserve the single-letter distortion criterion. In future work, we plan to code the signal after decorrelation.

We have shown that it is possible to construct physiologically motivated auditory models which allow very fast and perceptually accurate inversion. The robustness of the quality of the reconstructed acoustic signal under distortion in the perceptual domain shows that our model is accurate. The computational speed of the inversion procedure and the sparseness of its firing pulse representation give our model a significant advantage over other invertible auditory models (which use iterative inversion procedures) for speech and audio coding applications.

6. REFERENCES

1. M. Slaney, "Pattern playback from 1950 to 1995," in *IEEE Conf. Syst. Man. Cybernetics*, (Vancouver), pp. 1–6, 1995.
2. F. S. Cooper, "Acoustics in human communication: evolving ideas about the nature of speech," *J. Acoust. Soc. Am.*, vol. 68, no. 1, pp. 18–21, 1980.
3. T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3549–3554, 1993.
4. M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, vol. II, (Adelaide), pp. 77–80, 1994.
5. R. Hukin and R. Damper, "Testing and auditory model by resynthesis," in *Proc. Eurospeech*, (Paris), pp. 243–246, 1989.
6. X. Yang, K. Wang, and S. Shamma, "Auditory representation of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824–839, 1996.
7. P. L. Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, pp. 182–208, 1993.
8. D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 32, pp. 236–243, April 1984.
9. R. D. Patterson, K. Robinson, J. Holdsworth, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception* (Y. Cazals, L. Demany, and K. Horner, eds.), pp. 429–446, Oxford: Pergamon, 1992.
10. S. Seneff, "A joint synchrony/mean rate model of auditory speech processing," *J. Phonet.*, vol. 16, pp. 55–76, 1988.
11. T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trejectories of spectral envelopes," in *Int. Conv. Spoken Lang. Process.*, (Philadelphia), pp. 2490–2493, 1996.
12. R. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, pp. 1560–1586, 1987.
13. T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrowband carriers"," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905, 1997.