DICHOTIC PRESENTATION OF INTERLEAVING CRITICAL-BAND ENVELOPES: AN APPLICATION TO MULTI-DESCRIPTIVE CODING

Oded Ghitza and Peter Kroon

Bell Labs, Lucent Technologies
Multimedia Communications Research Laboratory
Murray Hill, New Jersey 07974, USA
http://www.bell-labs.com/user/{og,kroon}

ABSTRACT

A coding paradigm is proposed which is based solely on the properties of the human auditory system and does not assume any specific source properties. Hence, its performance is equally good for speech, noisy speech, and music signals. The signal decomposition in the proposed paradigm takes advantage of binaural properties of the human auditory system. This also leads to a natural multi-descriptive signal representation.

1. INTRODUCTION

For narrow-band speech (4 kHz bandwidth), the most effective coders are based on CELP techniques operating in the 4–16 kb/s range. Because of the inherent assumptions made about the source, these coders perform poorly with music-like signals. On the other hand, audio coders such as PAC [1] work well at higher rates and wider bandwidth audio signals. At bit rates below 24 kb/s, these coders do not work well for speech-like signals. Hybrid coders such as MTPC [2] attempt to combine both coding paradigms and work reasonable well at ranges between 16 and 24 kb/s for both speech and audio signals (8 kHz bandwidth).

Another problem that occurs with any of the above techniques is that when used in a packet switched network, the coders have to be robust against packet losses. Although error mitigation techniques are effective they lose their efficiency at packet loss rates above 3%. The traditional approach is to increase the receive buffer size, which reduces the impact of late arriving packets. However, in a communications scenario the increased delay will impair two-way communications and requires more sophisticated echo control. Multi-descriptive techniques (e.g., [3, 4]), which allow a source coder to split its information into two (or more) equally relevant bit streams, have been suggested as a solution for this problem. For a two stream scenario each stream is encoded such that when used independently, it provides a reasonable quality x. When both streams are received the decoding provides a quality y, which is better than x. Assuming that both streams can be transmitted such that the packet loss probabilities are independent, much higher loss rates become tolerable with only a small degradation in quality.

The proposed coding paradigm, described in Section 2, addresses both issues. It is based solely on the properties of the human auditory system, and does not assume any specific source properties. Hence, its performance will be equally good for both speech and music signals. The signal decomposition in the proposed method takes advantage of the binaural properties of the human auditory system, leading to a natural multi-descriptive decomposition as well.

2. PRINCIPLE

It is widely accepted that a decomposition of the output of a cochlear filter into a temporal envelope and a "carrier" may be used to quantify the role of auditory mechanisms in speech perception (e.g., [5]). This is supported by our current understanding of the way the auditory system (the periphery, in particular) operates.

By analogy between measured auditory nerve (AN) responses of the cat (e.g., [6, 7]) and possible AN responses of a human, we expect a significant difference between the properties of the firing patterns of low CF and high CF fibers[1]. At low CFs, neural discharges of AN fibers are phase locked to the underlying driving cochlear signal (i.e., synchrony is maintained). At high CFs, the synchrony of neural discharges is greatly reduced. At these CFs, temporal information is preserved by the instantaneous average rate of the neural firings, which is related to the temporal *envelope* of the underlying driving cochlear signal. Obviously, there is no distinct boundary between these AN regions. Rather, the change in properties is gradual. However, our working hypothesis is that the region of transition is around 1200 Hz.

Currently, we lack understanding of the post-AN mechanisms that are active at the low frequency range (and are sensitive to synchrony). Hence, in the current coding paradigm we encode the *baseband* signal (up to 1200 Hz), without decomposition.

For the frequency band above 1200 Hz, we take advantage of the physiological limitations of the Inner Hair Cell (IHC) to follow the carrier information (as reflected by the loss of synchrony in the AN neural firings). Let:

$$s_i(t) = s(t) * h_i(t) = a_i(t)\cos\phi_i(t) \qquad (1)$$

where $s(t)$ is the input signal, $h_i(t)$ is the impulse response of the cochlear filter centered at frequency ω_i, the operator $*$ represents convolution, and $a_i(t)$ and $\cos\phi_i(t)$ are, respectively, the envelope and the carrier information of the cochlear signal $s_i(t)$. Because of the IHC limitations, neural firings of AN nerve fibers originating at ω_i exhibit only the envelope information $a_i(t)$, while the carrier information is lost. Let us synthesize the signal:

$$\hat{s}_i(t) = a_i(t)\cos\omega_i t \qquad (2)$$

that is, the original carrier $\cos\phi_i(t)$ is replaced by a cosine carrier $\cos\omega_i t$. For a band-limited envelope $a_i(t)$, $\hat{s}_i(t)$ is a band-limited signal centered at frequency ω_i. If $\hat{s}_i(t)$ is presented to the listener's ear, the resulting envelope signal at the appropriate place

[1] CF, for **Characteristic Frequency**, indicates the place of origin of a nerve fiber along the basilar membrane in frequency units

along the cochlear partition, which corresponds to frequency ω_i, will be $a_i(t)$. Let:

$$\hat{s}(t) = \sum_{i=1}^{N} \hat{s}_i(t) = \sum_{i=1}^{N} a_i(t) \cos \omega_i t \tag{3}$$

where $a_i(t), i = 1, \ldots, N$ are the envelope signals of N cochlear filters equally spaced along the critical-band scale, with a spacing of one critical band. (For an input signal with 4 kHz bandwidth, the number of critical bands above 1200 Hz is N=10. For a bandwidth of 8 kHz, N=17.) Recalling that information is conveyed to the AN by a large, finite, number of highly overlapped cochlear filters (determined by the discrete distribution of the IHCs along the continuous cochlear partition), for the original signal $s(t)$ the overall envelope information at the AN level is represented with a frequency resolution determined by the density of the IHCs. Hence, the envelope signals $a_i(t), i = 1, \ldots, N$, in Eq. (3) represent only a sparse sample of the overall envelope information at the AN level.

Let $\hat{s}(t)$ of Eq. (3) be presented to the listener's ear. The envelope at the output of the listener's cochlear filter located at frequency ω_i is (ideally) $a_i(t)$, for each $i, i = 1, \ldots, N$. However, the output of a cochlear filter located in between two successive cosine carrier frequencies, ω_i and ω_{i+1}, will reflect "beating" of the two modulated cosine carrier signals passing through the filter. This results in an undesired distortion.

To reduce the amount of distortion due to beating, a dichotic[2] synthesis with interleaving critical bands is proposed. Let $\hat{s}_{odd}(t)$ and $\hat{s}_{even}(t)$ be the summation of the odd components and even components of $\hat{s}(t)$, respectively, i.e.,

$$\hat{s}_{odd}(t) = \sum_{i \in odd}^{N-1} a_i(t) \cos \omega_i t \tag{4}$$

$$\hat{s}_{even}(t) = \sum_{i \in even}^{N} a_i(t) \cos \omega_i t \tag{5}$$

The distance between two successive cosine carriers in each of these signals is larger, resulting in a reduction of distortion due to carrier beating. When $\hat{s}_{odd}(t)$ and $\hat{s}_{even}(t)$ are presented to the left and the right ears, respectively, the auditory system produces a single fused image.

Two points are noteworthy. First, $h_i(t)$ of Eq. (1) is a *cochlear* filter (realized, for example, as a Gammatone filter, [8]). This implies that $h(t) = \sum_{i=1}^{N} h_i(t)$ is *not* an all-pass filter (i.e., the signal $\sum_{i=1}^{N} s_i(t)$, where $s_i(t)$ are the untampered cochlear signals of Eq. (1), is different from the original signal $s(t)$ of Eq. (1)).[3] Such a behavior, however, is legitimate here because we do not aim at reproducing the original signal. Rather, our purpose is to synthesize a signal that will stimulate a neural activity at the listener's AN, which is in correspondence with the cochlear envelope information that would be generated by the original signal.

Second, our signal processing technique (i.e., the usage of pure cosine carriers to place the sampled envelope signals at the appropriate place along the basilar membrane) produces an inherent, undesired, distortion due to the following reason. When the original signal $s(t)$ is passed through the highly overlapped, full resolution cochlear filter-bank, the resulting envelope information gradually evolves as we move across the filter-bank array. In contrast, passing $\hat{s}_{odd}(t)$ or $\hat{s}_{even}(t)$ of Eqs. (4) and (5), respectively, through the same filter-bank will result in much coarser evolution of the envelope information. (This is so because of the sparse sampling of the envelope information by the filters $h_i(t)$ of Eq. (1)).

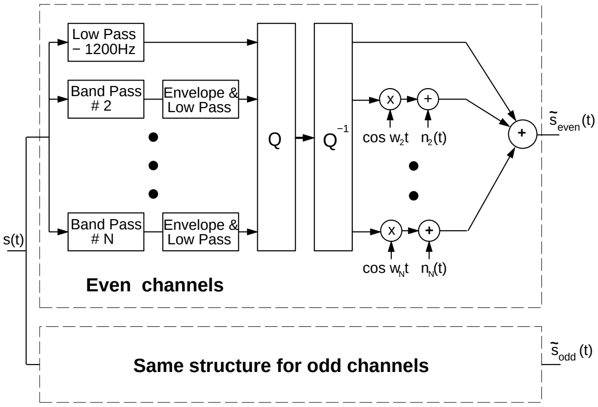

Figure 1: Block diagram of the proposed system.

3. APPLICATION TO CODING

3.1. Information reduction based on perception

One source of information reduction is the IHCs physiological limitations to follow the carrier information, which allows the usage of pure cosine carriers (Eqs. (4) and (5)) whose frequencies are *known* to the receiver. Further information reduction can be achieved by replacing the cochlear envelope $a_i(t)$ by a *smoothed* (via low-pass filtering) envelope $\tilde{a}_i(t)$, i.e.,

$$\tilde{s}_{odd}(t) = \sum_{i \in odd}^{N-1} \tilde{a}_i(t) \cos \omega_i t \tag{6}$$

$$\tilde{s}_{even}(t) = \sum_{i \in even}^{N} \tilde{a}_i(t) \cos \omega_i t \tag{7}$$

Recent psychophysical experiments ([9]) show that if the cutoff frequency of the low-pass filter is about 250 Hz, a speech signal synthesized with these smoothed cochlear envelopes (Eqs. (6) and (7)) is perceptually indistinguishable from the speech signal synthesized with the original envelopes (Eqs. (4) and (5)).

The proposed system is shown in Fig. 1. Note that the information to be transmitted is comprised of the baseband signal and the smoothed critical-band envelopes.

3.2. Application to multi-descriptive coding

By using dichotic synthesis with interleaving channels, the signals $\tilde{s}_{odd}(t)$ and $\tilde{s}_{even}(t)$ of Eqs. (6) and (7) are uncorrelated above 1200 Hz. Hence, the following multi-descriptive synthesis is proposed. At the receiver, depending on the measured channel losses, the left ear (L) and the right ear (R) may be fed by: (1) $\tilde{s}_{odd}(t)$ to L and $\tilde{s}_{even}(t)$ to R, (2) $\tilde{s}_{odd}(t)$ to both L and R, or (3) $\tilde{s}_{even}(t)$ to both L and R.

The baseline system introduces two types of artifacts. One, the usage of pure cosine carriers causes perceivable distortions in $\hat{s}_{odd}(t)$ and $\hat{s}_{even}(t)$ (Eqs. (4) and (5)), the amount of which depends on the interaction between spectral contents and carrier frequencies, and the listener experience.

[2] Signals presented to left and right ears are different.

[3] Note that in traditional sub-band coding systems, the filter bank is designed with a "perfect" reconstruction requirement.

Second, a dichotic presentation creates a spatial image that is different from one created by a diotic (i.e., the same signal is presented to both ears) presentation. When the proposed method is used as a multi-descriptive system, a switch from the dichotic to the diotic mode results in a switch in the spatial location of the image. This problem is usually mitigated in real-world applications where the two-channel delivery is usually accomplished via two loudspeakers (e.g., desktop application), instead of stereophonic headphones.

3.3. Complexity, Delay and Quantization

To make an actual coder based on the described paradigm it is important to constrain the overall complexity and delay. In the description below we give an example how this can be accomplished. Note that more sophisticated methods could be devised to obtain better coding efficiency at the expense of an increase in delay. The bandpass filter-bank of Fig. 1 is implemented with 128 tap FIR filters, introducing an 8 ms delay (8000 samp/sec). The baseband signal (1200 Hz bandwidth) and the low-pass filtered envelope signals (250 Hz bandwidth) are down-sampled by a ratio of 1/3 (2666 samp/sec) and 1/15 (533 samp/sec), respectively, to maintain a simple time relationship between the various signals. Any coding delay at the down-sampled frequencies will increase delay by its respective down-sampling factor. Hence, we concentrate on coding schemes that work on a sample-by-sample basis, such as Delta-Modulation or ADPCM. The down-sampled envelope signals are quite robust against quantization noise and it was found that a simple ADPCM structure with a 2 bits/sample quantizer provides excellent results. In contrast, the baseband signal is more sensitive to quantization errors. Using ADPCM, it was found that at least 3 bits/sample are needed for an acceptable quality. Although the baseband coder could in principle be a multi-descriptive coder [10], for simplicity we use the baseband information for both streams. The total bit rate is 2 channels × (baseband: 8 kb/s + envelopes: 5 × 1.066 kb/s) = 26.66 kb/s, and the total coding complexity is approximately (baseband: 1.3 MIPS + envelopes: 10 × 0.3 MIPS) = 4.3 MIPS.

Backward adaptive prediction with a limited size VQ can further reduce the number of bits per sample, without introducing large algorithmic delays. The predictor order needs not to be very high to produce an accurate description of the spectrum. The specific choice of down-sampling rates allows the use of a 5 dimensional VQ without introducing additional coding delay. We used a modified version of LD-CELP [11] using only a 16-th order predictor on the down-sampled signal, and without post-filter in the decoder, resulting in 2 bits/sample (2666 samp/sec) without noticeable audible degradations and a complexity of 8 MIPS. At the decoder the signals were up-sampled and interpolated with 32-tap FIR filters, resulting in an additional 2 ms delay. The overall complexity of this scheme is about 14 MIPS for the filtering and upsampling, and 11 MIPS for the quantization. The total end-to-end delay is 10 ms due to filtering and 2 ms due to coding. The total bit rate is 2 channels × (baseband: 5.33 kb/s + envelopes: 5 × 1.066 kb/s) = 21.332 kb/s. (For wide-band signals (8 kHz), the net increase in bit rate will only be 3 to 4 kb/s for each bit stream.)

The 21 kb/s system was used in an informal listening tests using two loudspeakers. We defined the packet-size to be equivalent to 15 ms. Without packet losses, most listeners could not perceive any degradations comparing the non-quantized and quantized versions of the baseband and smoothed envelopes ($\tilde{s}_{odd}(t)$ and $\tilde{s}_{even}(t)$ of Eqs. (6) and (7)). We randomly erased packets up to 10% loss rates. The error probability for each of the streams was independent. When a packet-loss was present in both channels we used the concealment technique outlined in [12]. Based on limited informal listening it was found that most listeners could not tell the difference between the 10% loss condition and the no-loss condition. Degradations became more noticeable once the loss rate was increased to 20%. Note that at this loss rate the probability of losing packets from both streams simultaneously becomes 4% and some of the degradations are caused by the inefficiencies of the concealment algorithm.

4. SUMMARY

In this paper we introduced the notion of dichotic synthesis of interleaving critical bands, which provides a framework for (1) a substantial information reduction based on perception (fixed cosine carriers, modulate by smoothed critical-band envelopes), and (2) a natural multi-descriptive signal representation.

REFERENCES

[1] D.Sinha, J. Johnston, S. Dorward, and S. Quackenbush, “The perceptual audio coder (PAC),” in *The digital signal processing handbook* (V. Madisetti and D. Williams, eds.), pp. 42–1:42–17, CRC Press, 1998.

[2] S. A. Ramprashad, “A multimode transform predictive coder (MTPC) for speech and audio,” *IEEE Speech Coding Workshop*, pp. 10–12, 1999.

[3] K. Wolf, A. Wyner, and J.Ziv, “Source coding for multiple descriptions,” *Bell System Technical Journal*, vol. 59, no. 9, pp. 1417–1426, 1980.

[4] L. Ozarow, “On a source-coding problem with two channels and three receivers,” *Bell System Technical Journal*, vol. 59, no. 10, pp. 1909–1921, 1980.

[5] J. L. Flanagan, “Parametric coding of speech spectra,” *Journal of the Acoustical Society of America*, vol. 68, no. 2, pp. 412–430, 1980.

[6] B. Delgutte and N. Y. S. Kiang, “Speech coding in the auditory nerve: I. vowel-like sounds,” *Journal of the Acoustical Society of America*, vol. 75, no. 3, pp. 866–878, 1984.

[7] M. B. Sachs and E. D. Young, “Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate,” *Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 470–479, 1979.

[8] M. Slaney, “An efficient implementation of the Patterson-Holdsworth auditory filter bank,” Tech. Rep. 33, Apple Computer, 1993.

[9] O. Ghitza, “On the bandwidth of the auditory critical-band envelope detectors,” *Journal of the Acoustical Society of America*, Submitted for publication, 2000.

[10] C.-C. Lee, “Diversity control among multiple coders: A simple approach to multiple descriptions,” *IEEE Speech Coding Workshop*, Submitted for publication, 2000.

[11] J.-H. Chen and R. V. Cox, “The creation and evolution of 16 kbit/s LD-CELP: from concept to standard,” *Speech Communication*, pp. 103–112, 1993.

[12] D.Kapilow and R. V. Cox, “A high quality low-complexity algorithm for frame erasure concealment with G.711,” *ITU-T Study Group 16 Contribution*, May 1999.