PTC: Perceptual Transform Coding for Bandwidth Reduction of Speech in the Analog Domain, Pt 1

A new method for optimizing the bandwidth of phone signals using auditory psychophysics

By Doug Smith, KF6DX

A revolution is afoot in Amateur Radio: An increasing number of operators are producing high-fidelity audio in the narrow bandwidths available to us on HF SSB. Many of us have grown tired of listening to the same old "communications-quality" signals. We have yearned for a more pleasurable sound from our equipment. Coupled with skills learned in professional recording and broadcast studios, the availability of high-quality HF transceivers in the last few years has enabled some startling accomplishments in narrow-band audio quality.

It is remarkable what can be achieved in a bandwidth (BW) of only 3 kHz. Characteristics of speech processing can be manipulated to allow the perception of much greater BW. Properties of human speech can be further exploited to reduce the occupied BW of phone emissions quite significantly. That is the subject of this paper.

PO Box 4074
Sedona, AZ 86340
kf6dx@arrl.org

Drawing on the extensive audio-coding research of others, I will show how certain human speech and hearing attributes lend themselves to analog BW compression of speech. In Part 2, I will demonstrate how a speech signal of 4-kHz BW is compressed to occupy less than 1 kHz and a full-range signal of 15-kHz BW to less than 4 kHz. I will emphasize the technical tradeoffs that influence sound quality. The goal is to retain the perceived quality of the original, uncompressed signal. First, however, please follow me through a little history and background as I lay down the basis for my invention.

A History of Phone Modes

In the days before SSB became popular on our bands, AMers used a lot of plate-modulated vacuum-tube equipment. It was relatively easy to obtain a broad baseband frequency response with this type of gear—perhaps it was too easy to be too broad! It was also easy to sustain lots of interference and noise, since each information-bearing sideband reaches only about 1/6 of the total output power. Although each sideband is a mirror image of the other, selective fading often makes it difficult to recover all of the energy from both sidebands simultaneously. Carrier fades tend to cause severe distortion. Modern methods of exalted-carrier, synchronous detection have largely solved those problems, but the occupied BW of AM has relegated it to some obscurity on the Amateur Radio bands. It is retained for broadcasting because it is detectable with relatively simple equipment.

SSB is popular because all the output power is dedicated to the information and emissions occupy only the BW necessary for perfect reproduction. SSB also does not suffer from the distortion caused by carrier fading. It does impose constraints, however, that result in loss of fidelity. In the filter method of SSB generation, it is usually essential to "roll off" the low-frequency response to ensure adequate suppression of the carrier and opposite sideband. Even with the phasing method, opposite-sideband suppression may suffer if low audio frequencies are not attenuated. These problems have made it difficult to achieve good low-frequency response in SSB. Operators have been frustrated (until recently) by the limitations of IF

filters in their transceivers. They can seriously attenuate both low- and high-frequency audio content.

SSB experimenters are well aware of certain speech-processing tools, such as AF and RF compressors. Automatic level control (ALC) is found in every modern rig. ALC is just a form of compressor that prevents drive signals from exceeding the PEP limitations of the transmitter. In a peak-limited system, average output power depends heavily on the nature of the modulation. Some voices produce peak-to-average ratios of up to 15 dB; a station running 1500 W PEP might only produce an average output of about 50 W!

Because of the Hilbert-transform or "repeaking" effect of SSB, AF limiting achieves only a modest intelligibility increase even with large compression ratios. IF or RF compression avoids this problem—6 dB or more improvement in average output power is possible.

For audiophiles, the trouble with any compression scheme is that it adds distortion. Naturally, any departure from linearity involves harmonic distortion (HD) and intermodulation distortion (IMD). At high compression ratios, an AF compressor especially suffers from HD effects that reduce clarity. Formant energy and plosive sounds tend to be sacrificed. IF and RF compressors generate HD that falls outside the band of interest; hence it is easily removed by filtration. These compressors still create in-band IMD, though; this distortion ultimately limits their effectiveness.

While on the subject, let's note that distortion caused by our electronics limits the quality level we can finally attain, no matter what we do. Many receivers produce as much in-band IMD as do transmitters. The phase and amplitude of each IMD product are influenced by many variables. Levels can be measured, however, and the transfer function ascertained. Whether these products augment or diminish intelligibility seems to involve another set of variables that depend on the nature of human speech and hearing systems. As I'll highlight later, these cannot be directly measured.

So the question is, How can we produce better audio quality while using a narrow BW? A lot of work has been done on this problem, especially with respect to digital coding of audio.[1, 2, 3, 4, 5, 6] The impetus for this work has been provided by the recording industry, telephone companies and interest in passing audio over Internet connections at low bit rates. Most of the breakthroughs in such coding have focused on characterizing human speech in ways that are efficiently represented by ones and zeros. Progress on BW compression in the analog domain has been frustrated by increasing emphasis on digital modes. Digital methods may have an advantage in error detection and correction, and in signal-to-noise ratio (SNR), but they likely will never be the most BW-efficient techniques for speech coding.

Linear predictive coding (LPC) and other methods[7, 8, 9] have concentrated on passing parameters that describe features of speech production. They are "lossy" in the sense that they sacrifice perfect reproduction of the input waveform for BW reduction. Perceptual audio coders[10, 11, 12] code in such a way that redundancy and irrelevancy in speech are removed, reducing BW. Both approaches take advantage of the fact that only perceived quality matters. I shall adopt this as my sole criterion for the remainder of this discussion.

Evaluating the Human Hearing System

Speech communication is crucial to our society. It conveys the sense of how someone feels, how they are thinking and some idea of who they are more than any other form. Nothing is more comforting than hearing the voice of a loved one in dire times. I postulate, therefore, that this mode of telecommunications will never be replaced.

Because of that suspicion, I can write that the secondary goal of any speech-coding scheme is to preserve those characteristics of speech that allow us to recognize the speaker, along with the nuances that are so important. In other words, we have to conserve certain distinctive qualities of speech so that we can't tell the speech was coded. Let's examine what those qualities are and what it is about human hearing that influences perception.

Perception vs. Measurement

In the study of the human hearing system, it must be clear that there is no objective means of measurement. All information about what someone hears (or doesn't hear) must be learned subjectively through the responses of the listener. All we can do is ask questions of a subject and attempt to infer something about the nature of sounds. Furthermore, we have no guarantee that a particular stimulus will be perceived in the same way by one subject as another. We therefore define our terms for measurement and perception differently and separately.

Sound *intensity* is a physical measure of air pressure level. Two persons equipped with identically calibrated instruments will measure the same intensity for any given sound. *Loudness* is the corresponding perceptual magnitude. It can be defined as "that attribute of auditory sensation in terms of which

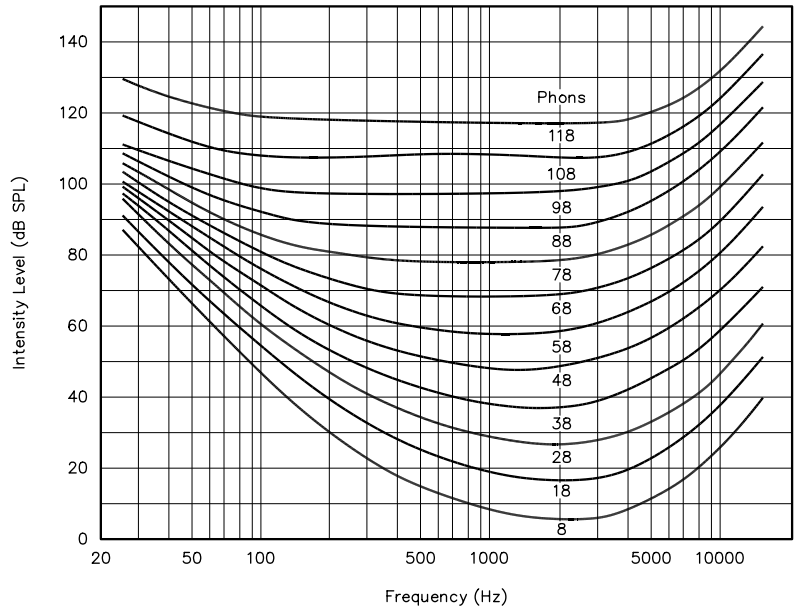

Fig 1—Intensity versus frequency for constant loudness.

[1]Notes appear on page 12.

sounds can be ordered on a scale extending from quiet to loud."[13] The unit of loudness, the *sone*, is defined by subjectively measuring loudness ratios. A stimulus half as loud as a one-sone stimulus has a loudness of 0.5 sones. A 1-kHz tone at 40 dB sound-pressure level (SPL) is arbitrarily defined to have a loudness of one sone.

We might be left to wonder how a unit based solely on individual perceptions can be useful, especially since so much variation exists from person to person. The method of applying stimuli and of obtaining responses from listeners has a large effect on results. Loudness comparison of two equal-frequency tone bursts, however, generally produces reliable and repeatable data. Loudness comparisons between dissimilar stimuli, such as between a pure tone and a polyphonic source, yield unpredictable results because of poorly understood subjective effects. So a quantification of loudness scaling (one sound is half as loud as another) is as good as absolute loudness matching (one sound is the same loudness as another). Additionally, some researchers have observed under many conditions that loudness adds.[14] Binaural presentation of stimuli generally results in loudness doubling and two equally loud sources—if they are far enough apart in frequency—are twice as loud as one alone. Because of other effects described below, this rule must be used with caution, though. There is evidence that loudness addition is far from a perfect description of human perception.[15]

Frequency is a physical measure of a sound's number of cycles per second; each of us can measure frequency identically using similar instruments. We define *pitch* as the perceptual quantity corresponding to frequency. Pitch is to frequency as loudness is to intensity. Note that the relations between loudness/intensity and pitch/frequency are not necessarily linear, nor are the two perceptual measures independent of one another. Under certain conditions, the loudness of a constant-intensity sound can be shown to decrease with decreasing frequency; pitch can be shown to decrease with increasing intensity, even when frequency is held constant.

As ably documented by Fletcher,[16] Stevens and Davis,[17] and others, loudness depends on both frequency and intensity. Fig 1 (after Reference 17) shows some loudness contours. Each curve represents a constant-sone level. These data have been measured countless times, but the basic revelations remain unchanged. The most sensitive frequency region of the ear is between 1.5 and 3.0 kHz and the curves get flatter as the intensity is raised. Further, loudness grows faster with intensity at low frequencies. Finally, the curves reveal the dynamic range of hearing: Single tones below the zero-sone curve are inaudible, while tones above the top line are painful. In fact, we know today that the useful dynamic range of human hearing is substantially less than shown. Extended exposure to sounds well under the top line produces permanent hearing loss in some individuals.[18]

This is Auditory Psychophysics

We're now well into what is called *auditory psychophysics*, or just *psychoacoustics*. Recall that our goal is to exploit the redundancies and irrelevancies in speech to reduce its occupied BW. To identify the irrelevant content, we must discover how well the ear-brain combination discerns differences in intensity and frequency. Moreover, we must try to ascertain the performance of the hearing system in the presence of polyphonic sounds; that is, how certain sonic components tend to dominate others of lesser intensity or of small frequency difference.

I will now expand the discussion to include definitions for various perceptual thresholds, to introduce the idea of *masking*, and to present the concept of *critical bands*.

Thresholds of Hearing

One of the thresholds of hearing, the *intensity threshold*, is defined as the lowest intensity the listener can detect. We cannot directly measure the listener's perception, though; we can only ask whether he or she thinks the sound is audible. This might seem a fine distinction, but the method of measurement determines the threshold as much as the listener's aural gifts.

At or near the intensity threshold, the subject's *criterion level* is in play.[19] He or she might indicate some sound is audible when it *might be* present, or perhaps only when it is *definitely* present. With no incentive to produce correct results (such as large sums of cash), the criterion level is beyond the experimenter's control.

An interesting way of dealing with the uncontrolled criterion-level problem is to use a *criterion-free* experimental model. According to Hall (see Note 19), the simplest of these is the "two-interval, forced-choice" paradigm. In this method, the stimulus is presented at random in one of two observation intervals. The subject is asked to determine in which of the two intervals the stimulus was present. A perfect observer always selects the interval that elicits the larger

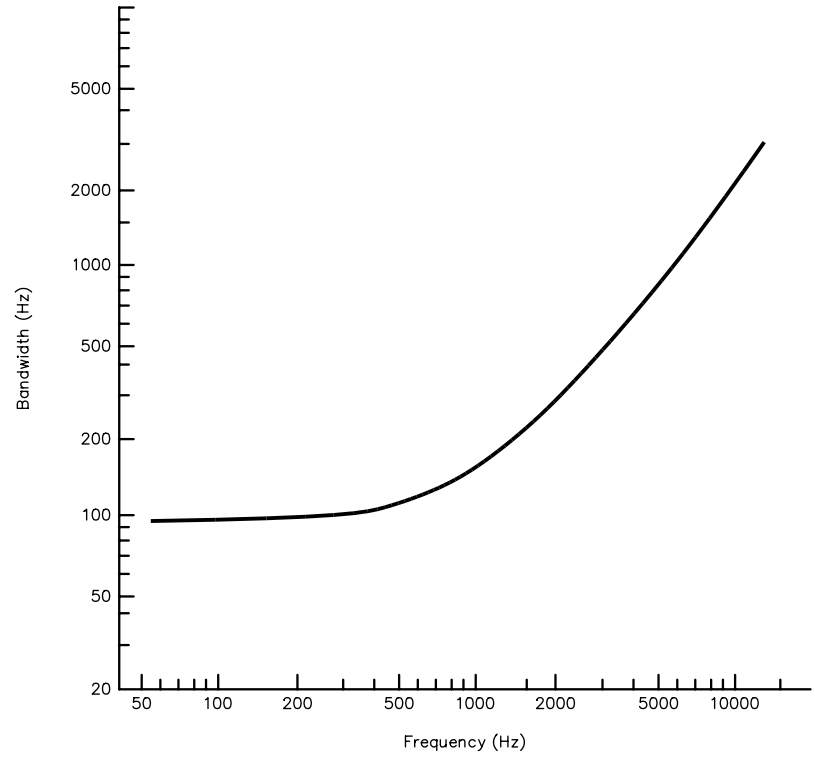

Fig 2—Critical BW versus frequency.

decision variable; thus the criterion level is no longer a factor. He or she has a 50% chance of selecting the correct interval even without actually detecting the stimulus. It can be shown that the *psychometric function* thereby produced solves the criterion-level problem.

I think it interesting to note that all this has a bearing on "A/B" comparisons as commonly done on the air, regardless of the parameter being changed. A measurement criterion such as loudness or signal strength must first be set, then the stimulus presented *at random* to the observer. Any additional information given the observer prior to measurement, such as "A is amplifier off, B is amplifier on," introduces bias in the result. Further discussion of *detection theory* is beyond the scope of this paper.

Getting back to definitions, we may also define *differential intensity threshold* as the ability to detect whether one sound is louder than another. In fact, we may define differential thresholds for other attributes of sounds, such as frequency and duration. A differential threshold is the amount one or more of these attributes must change to allow an observer to detect the change.

In the first half of the last century, German physiologist E. H. Weber gave us the first serious, quantitative depiction of differential thresholds. According to *Weber's Law*, the differential intensity threshold dI is proportional to the stimulus intensity I, or:

$$\frac{dI}{I} = k \qquad \text{(Eq 1)}$$

where k is known as the *Weber fraction*. This alleged constant has also been applied to sensitivity to changes in frequency and BW, as well as nonauditory senses such as color, image sharpness, pain, smell and taste. Very soon after Weber made this "law" known, folks found out it broke down at intensities near absolute thresholds. Physicist G. T. Fechner, also a German, suggested a *modified Weber's Law*:

$$\frac{dI}{(I + I_0)} = k \qquad \text{(Eq 2)}$$

where I_0 is a constant. It's a good approximation, but it apparently doesn't hold exactly.

Masking

Masking is defined as the ability of one sound (the masker) to render another (the desired) inaudible when present simultaneously or closely in time. It is quantified as the difference between the absolute intensity threshold of the desired in the absence of the masker and the elevated intensity threshold of the desired when the masker is present. Fletcher and Munson made a landmark study of the relation between loudness and masking effects.[20] They found that quieter sounds that are close in frequency to dominant sounds are rendered inaudible in proportion to their spectral separation and their relative intensities. They were among the first to use bands of "colored" noise as maskers. An important effect is the relationship between the masker BW and the amount of masking. This relation is most prominent when the desired signal lies within the masker's BW. Noise whose entire BW lies outside the desired signal's frequency does not contribute much to its masking. This is one manifestation of the human hearing system: For many auditory functions, the ear behaves as if it is a set of band-pass filters and energy detectors. These filters are said to occupy critical bands.

Critical Bands and Peripheral Auditory Filters

The above-mentioned relation between BW and masking is only one example of human hearing behavior relevant to the coder I will describe in Part 2. Another example is provided by SSB over HF, where the ear quite often encounters severe phase distortion. The ear seems to tolerate relatively large shifts in the relative phases of speech components without impairing intelligibility, when the components are far enough apart in frequency. Scharf[21] defined the critical bandwidths associated with these theoretical auditory filters as "that bandwidth at which subjective responses rather abruptly change." He measured critical bands using two-tone masking and loudness-summation techniques. Zwicker *et al*[22] measured phase sensitivity using polyphonic sounds. These studies agree fairly well with others performed over the years. Fig 2 is a plot of critical BW versus frequency that averages the Scharf and Zwicker data.

These and other studies support the idea that *differential frequency threshold* increases with frequency. In other words, it is more difficult to discern small frequency differences at high audio frequencies. Since we decided that our perception of things is all that matters, it makes sense to analyze speech signals with a system whose frequency resolution matches that of the human hearing system. It is remarkable that this sort of approach also seems to apply across a broad scale of other things we can classify. The science of image compression and construction, for example, has made extensive use of the methods I will relate in Part 2.

Notes

[1]R. E. Crochiere, S. A. Weber, and J. L. Flanagan, "Digital Coding of Speech in Subbands," *Bell System Technical Journal*, Vol 55, October 1976.

[2]P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, (Englewood Cliffs, NJ: Prentice-Hall, 1992).

[3]M. Vetterli, and J. Kovacevic, *Wavelet and Subband Coding*, (Englewood Cliffs, NJ: Prentice-Hall, 1995).

[4]N. S. Jayant, and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, (Englewood Cliffs, NJ: Prentice-Hall, 1984).

[5]R. D. Jurgen, "Broadcasting with Digital Audio," *IEEE Spectrum*, March 1996.

[6]P. Noll, "MPEG Digital Audio Coding Standards," *The Digital Signal Processing Handbook*, V. K. Madisetti, and D. B. Williams, editors (Boca Raton, FL: CRC Press LLS, 1998).

[7]L. R. Rabiner, and R. W. Schafer, *Digital Processing of Speech Signals*, (Englewood Cliffs, NJ: Prentice-Hall, 1978).

[8]B. S. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Transactions on Communications*, COM-30, April 1982.

[9]A. Gersho, "Advances in Speech and Audio Compression," *Proceedings of the IEEE*, Vol 82, 1994.

[10]J. D. Johnston, and A. J. Ferreira, "Sum-Difference Stereo Transform Coding," *ICASSP-92 Conf. Rec.*, II, 1992.

[11]D. Sinha, J. D. Johnston, S. Dorward, and S. R. Quackenbush, "The Perceptual Audio Coder (PAC)," *The Digital Signal Processing Handbook*, V. K. Madisetti, and D. B. Williams, editors, (Boca Raton, FL: CRC Press LLS, 1998).

[12]R. V. Cox, "Speech Coding," *The Digital Signal Processing Handbook*, V. K. Madisetti, and D. B. Williams, editors, (Boca Raton, FL: CRC Press LLS, 1998).

[13]B. Moore, *An Introduction to the Psychology of Hearing*, (London: Academic Press, 1989).

[14]H. Fletcher, "Loudness, Masking, and Their Relation to the Hearing Process and Problem of Noise Measurement," *Journal of the Acoustic Society of America*, Vol 45, 1969.

[15]B. Scharf, and D. Fishkin, "Binaural Summation of Loudness: Reconsidered," *Journal of Experimental Psychology*, Vol 86, 1970.

[16]H. Fletcher, *Speech and Hearing in Communication*, ASA Edition, J. B. Allen, editor, American Institute of Physics, New York, New York, 1995.

[17]S. S. Stevens, and H. W. Davis, *Hearing*, (New York: John Wiley & Sons, 1938).

[18]C. M. Harris, editor, *Handbook of Acoustical Measurements and Noise Control*, (New York: McGraw-Hill, 1991).

[19]J. L. Hall, "Auditory Psychophysics for Coding Applications," *The Digital Signal Processing Handbook*, V. K. Madisetti, and D. B. Williams, editors, (Boca Raton, FL: CRC Press LLS, 1998).

[20]H. Fletcher, and W. A. Munson, "Relation between Loudness and Masking," *Journal of the Acoustic Society of America*, Vol 9, 1937.

[21]B. Scharf, "Critical Bands," *Foundations of Modern Auditory Theory*, J. V. Tobias, editor, (New York: Academic Press, 1970).

[22]E. Zwicker, G. Flottorp and S. S. Stevens, "Critical Bandwidth in Loudness Summation," *Journal of the Acoustic Society of America*, Vol 29, 1957.