Optimizing digital speech coders by exploiting masking properties of the human ear

M. R. Schroeder

Drittes Physikalisches Institut, Universität Göttingen, Federal Republic of Germany and Bell Laboratories, Murray Hill, New Jersey 07974

B. S. Atal and J. L. Hall

Bell Laboratories, Murray Hill, New Jersey 07974
(Received 8 June 1979; accepted for publication 13 August 1979)

In any speech coding system that adds noise to the speech signal, the primary goal should not be to reduce the noise power as much as possible, but to make the noise *inaudible* or to minimize its subjective loudness. "Hiding" the noise under the signal spectrum is feasible because of human auditory masking: sounds whose spectrum falls near the masking threshold of another sound are either completely masked by the other sound or reduced in loudness. In speech coding applications, the "other sound" is, of course, the speech signal itself. In this paper we report new results of masking and loudness reduction of noise and describe the design principles of speech coding systems exploiting auditory masking.

PACS numbers: 43.70.Lw, 43.70.Dn

INTRODUCTION

A question of great importance in the digital coding of speech signals is the problem of an *objective quality measure*. Speech quality measures based on signal waveform differences are limited in applicability. As an extreme illustration, the signal, $s(t)$, and its negative, $-s(t)$, are indistinguishable (except by the highly trained expert)—yet the "signal-to-noise ratio" based on the difference between the two waveforms is −6 dB! More generally, any phase distortion with a group delay variation limited to a few milliseconds has such a small effect on signal quality that it can be disregarded in the context of most synthetic speech quality considerations (Schroeder, 1975).

Thus, we are led to the *short-time amplitude* spectrum as a basis for objective speech quality measures. In the case of voiced speech sounds, the spectrum shows a "fine structure"—the harmonics of the fundamental frequency, which cause the signal to have a distinct pitch (high, low, monotone, quavery, etc.). It has been established that, for synthetic vowels, pitch errors should be smaller than 0.5% (Flanagan, 1972, p. 281) and that it is desirable to reproduce partial aperiodicities in the excitation, e.g., for voiced fricatives such as /v/ and /z/ and semivowels such as /j/ (Fujimura, 1972, Makhoul *et al.*, 1978).

The other attribute of the short-time spectrum is its *envelope* (Schroeder, 1966). The spectral envelope of processed speech is typically modified by three major contaminants:

(a) Additive noise,

(b) Nonlinear signal distortion,

(c) "Linear" distortion caused by spectral imbalance or poor spectral resolution.

In the following, we describe an approach to objective measures of spectral envelope distortion by incoherent processes, such as additive noise, based on detailed characteristics of human auditory perception.

Other objective measures of speech quality are described by Tribolet *et al.* (1978).

I. THE PERCEPTION OF NOISE IN THE PRESENCE OF SPEECH SIGNALS

In any speech coding system that adds noise to the speech signal, the subjective loudness of the noise can serve as a basis for an objective measure of speech quality. The loudness of the noise is determined not just by its total power but also by the distribution of the noise and signal powers along the basilar membrane. The noise can be reduced in loudness or even made completely inaudible by the signal, a phenomenon known as auditory masking.

In speech coding applications, this "concealment" of noise under the signal spectrum can be used to advantage to minimize the degradation of a speech signal by noise.

II. A SUMMARY OF THE METHOD

Our method of calculating speech signal degradation is based on the following mathematical formulation of the physiological process of hearing:

The sound pressure at the eardrums is transformed into the stapes velocity that "drives" the inner ear. This is accomplished by a middle-ear transfer function.

The inner ear then performs a running short-time spectral analysis in which the frequency coordinate f is represented by a spatial coordinate x—along the length of the basilar membrane. We approximate this process by short-time Fourier transformations over successive 20-ms time windows. The frequency variable is then converted to the physiologically based and perceptually decisive "critical-band" scale, the just-mentioned x scale (Zwicker and Feldtkeller, 1967).

The resulting critical-band densities are then converted to *basilar-membrane-excitation functions* by con-

 0001-4966/79/121647-06$00.80

volution with a basilar-membrane spreading function. These excitation functions, related to the envelopes of von Békésy's traveling waves on the basilar membrane, excite the mechanical-to-neural transducers, embodied in the hair cells (Schroeder and Hall, 1974), and thus form the basis of neural processing in higher centers of the auditory pathways.

Specifically, in our applications, we are interested in the perceptual degradation of a speech signal by quantizing noise. We measure this degradation by the relative *loudness* of the noise in the presence of the speech signal. Because the published data on loudness reduction of noise by speech signals are insufficient, we made our own measurements and have included some of the results in this paper (see Sec. IV).

The results of these measurements are used to compute the (reduced) loudness of each noise component and the total noise loudness. Then the loudness of the speech signal is calculated and the ratio of noise-to-speech loudness is taken as the objective measure of speech signal degradation.

III. COMPUTATION OF NOISE LOUDNESS

The *loudness* of a noise in the presence of a speech signal is computed as follows:

(1) The power spectra of the speech signal, $\tilde{S}(f)$, and of the noise, $\tilde{N}(f)$, are computed over time windows of approximately 20-ms duration, characteristic of human speech perception (Flanagan, 1972, pp. 148–149).

(2) The signal and noise power spectra are multiplied by the "middle-ear transmission function," essentially a low-pass filter with a cutoff at 5 kHz and with a small bump around 3.5 kHz, representing the sound transmission from the eardrum to the entrance of the inner ear (Schroeder, 1977, p. 323). For applications in any frequency range below 5 kHz, the middle-ear correction can be omitted—provided the signal and noise power spectra are set to zero above 5 kHz. The notation in the following formulas represents that simplification.

As a further refinement, the coupling of the eardrum to the external sound field could be taken into consideration. This depends on such factors as earphone fit or—in the case of loudspeaker presentations—on head orientation. For details of these "free-field" coupling effects, probably not crucial in most speech quality assessments, we refer to Blauert (1974).

(3) Our sound signal has now reached the inner ear (or cochlea), where a frequency-to-place transformation along the basilar membrane occurs. Here, the power spectra, which are functions of frequency f, are represented as "critical-band densities"—i.e., functions of the critical-band number x. Auditory perception is based on critical-band analysis in the inner ear. In fact, the peripheral auditory analyzer can be roughly thought of as a bank of bandpass filters, numbered $x=1$ to $x=24$, with bandwidths of about 100 Hz below 500 Hz and bandwidths of one-sixth of the center frequency above 1 kHz (Greenwood, 1961; Zwicker and Feldtkeller, 1967).

The relationship between frequency f and critical band number x, which is essentially linear below 500 Hz and exponential above 1 kHz, can be approximated by the formula (Schroeder, 1977)

$$f = 650 \sinh(x/7). \tag{1}$$

The match with measured critical bands is particularly good for frequencies up to 5 kHz (corresponding to $x<20$).

The x scale is not a mathematical figment; it is rooted in the anatomy of the inner ear. One critical band, i.e., an interval $\Delta x=1$, corresponds to a distance of 1.5 mm on the basilar membrane (the ear's primary frequency analyzer) and, more importantly, to a constant 1200 primary auditory nerve fibers—*regardless of location or frequency*.

The critical-band densities for speech and noise, called $S(x)$ and $N(x)$ respectively, are thus given by

$$S(x) = \tilde{S}(f(x)) \frac{df}{dx}. \tag{2}$$

Similarly,

$$N(x) = \tilde{N}(f(x)) \frac{df}{dx}. \tag{3}$$

Thus, the critical-band densities are computed from the spectra by the substitution $f(x)$ and multiplication with the density conversion factor.

(4) Next an "excitation pattern" $E(x)$ is computed from the speech critical-band density $S(x)$. Excitation patterns can be thought of as energy distributions along the basilar membrane. The computation of the excitation pattern is accomplished by convolution of the critical-band densities with a basilar membrane "spreading" function $B(x)$ (Mehrgardt and Schroeder, 1978):

$$E(x) = S(x) * B(x). \tag{4}$$

The spreading function $B(x)$, at intermediate speech levels, has lower and upper skirts with slopes of +25 dB and −10 dB per critical band. A convenient analytical expression for $B(x)$, derived from Zwicker's data (1963, Fig. 1), is (Schroeder, 1977):

$$10 \log_{10} B(x) = 15.81 + 7.5(x+0.474) - 17.5(1+(x+0.474)^2)^{1/2} \text{ dB}. \tag{5}$$

Similarly, the noise critical-band density $N(x)$ is converted to a "noise excitation pattern"

$$Q(x) = N(x) * B(x). \tag{6}$$

(The letter Q was chosen as a reminder that the noise in most applications of digital coding will be a *quantizing* noise.)

(5) Next the loudness L_S of the speech signal is computed as follows

$$L_S = c \int_{0.5}^{24.5} [E(x)]^{0.25} dx \text{ sone}, \tag{7}$$

where the constant c is usually chosen such that $L_S = 1$ for a 1-kHz tone at 40 dB SPL. The resulting loudness unit is called a "sone". The integral over x in Eq. (7) may be replaced by a sum for greater simplicity.

An optimum exponent in the integrand, according to Zwicker (1963), is 0.23. We prefer the "round" value 0.25. This exponent leads to an exponent of 0.3 in the relation between loudness and intensity (i.e., an increase of 10 dB in the intensity causes the loudness to double).

At the threshold of hearing, the loudness contribution is, by definition, zero. Thus, a formula more accurate at low levels is

$$L_S = c \int_{0.5}^{24.5} \mathrm{Max}\{[E(x) - \theta(x)]^{0.25}; 0\}\, dx, \tag{8}$$

where $\theta(x)$ is the threshold of hearing, expressed as a critical-band density (Zwicker and Feldtkeller, 1967).

The loudness of the speech signal will serve as a reference for the noise loudness to be computed next.

(6) The loudness of the noise is reduced by the presence of a masking signal. For the conditions presented by Hellman (1972), a critical band of noise masked by an in-band tone, this reduction can be described analytically as follows:

$$N' = N/[1 + (S/N)^p], \tag{9}$$

where N and S are noise and tone powers, respectively, and N' is the power of a comparison unmasked noise matched to have the same loudness as the masked noise. For $N = S$, $N' = N/2$, i.e., there is a loudness reduction equivalent to 3 dB.

The published results (Zwicker, 1963, Fig. 16; Hellman, 1972) on the loudness of noise in the presence of masking tones are rather meager—partly because of the extreme difficulty of making such measurements (Zwicker, 1963, p. 203). However, we have been able to overcome these difficulties by using a *pulsating* noise in the loudness comparison. According to our preliminary measurements, illustrated in Fig. 2, a good value for the exponent p in Eq. (9) is $p = 2$.

The total loudness L_N of a noise $Q(x)$ in the presence of a speech signal $E(x)$ is then, on the basis of Eq. (9) and in analogy to Eq. (7), postulated to be

$$L_N = c \int_{0.5}^{24.5} \left(\frac{Q(x)}{1 + [E(x)/Q(x)]^p} \right)^{0.25} dx. \tag{10}$$

Equation (10) does not include threshold effects. By definition, the loudness is *zero* (not just small) when the noise is inaudible because it falls below the masked threshold or absolute threshold of hearing. A refined formula including threshold effects is given below.

For our purposes it is convenient to express the masked threshold $M(x)$ as a product of the signal excitation function $E(x)$ and a "sensitivity function" $w(x)$:

$$M(x) = w(x) E(x), \tag{11}$$

where the sensitivity function $w(x)$ defines the threshold of masking. Any noise whose excitation pattern $Q(x)$ falls below the masking pattern $M(x)$ is *inaudible*.

A convenient analytic expression for $w(x)$ from our measurements is

$$10 \log_{10} w(x) \simeq -(15.5 + x)\,\mathrm{dB}. \tag{12}$$

Thus, for $f = 1$ kHz (i.e., $x = 8.5$) the masked threshold $M(x)$ is (approximately) 24 dB below the speech excitation pattern $E(x)$.

Because the noise is inaudible below the absolute threshold of hearing $\theta(x)$ and the masked threshold $M(x)$, the integrand in Eq. (10) should be set equal to zero if $Q(x)$ falls below either $\theta(x)$ or $M(x)$. An appropriate formula for the loudness of the noise, including these threshold effects, reads

$$L_N = c \int_{0.5}^{24.5} \left(\frac{\mathrm{Max}[Q - \mathrm{Max}(M; \theta); 0]}{[1 + (E/Q)^p]} \right)^{0.25} dx. \tag{13}$$

Of the several ways of including the threshold effects in the loudness formula, Eq. (13) has been designed to match observed behavior near threshold without unwarranted mathematical complexity.

(7) Our objective measure of speech signal degradation D is now given by the loudness of the noise divided by the loudness of the signal:

$$D = L_N / L_S. \tag{14}$$

The value $D \simeq 1$ signifies a very inferior speech quality. In fact, for a very poor signal-to-noise ratio of 0 dB and identically shaped signal and noise spectra, (i.e., $N(x) = S(x)$, the distortion measure will be $D = 0.84$.

At the other end of our quality scale, the noise will be *inaudible* if it is about 30 dB below the signal level. Thus, $D \simeq 0$ for $N(x) = 0.001\, S(x)$.

IV. PSYCHOPHYSICAL MEASUREMENTS OF MASKING

Existing data on the masking of noise by tones are limited to the case of the tone at the center frequency of the noise. Since this is too restrictive for our purposes, we are now in the process of making more general measurements. So far we have measured the masked threshold and the suprathreshold loudness of masked noise bursts as a function of the frequency of the masking tone. These measurements are being made for noise bursts of various center frequencies and bandwidths.

The masked threshold of a noise burst was measured with a form of Békésy audiometry, with a swept-frequency tone. For the results shown in Fig. 1, a continuous tone at an intensity of 80 dB SPL masked a noise burst centered at 1 kHz with a bandwidth of one critical band. As the frequency of the masking tone was swept slowly from 500 Hz to 1 kHz, and in a separate experiment from 2 to 1 kHz, the subject pressed a button while the noise burst was audible and released it while the burst was inaudible. The intensity of the noise burst decreased while the button was depressed and increased while the button was released so that it always remained near threshold. The resulting data were smoothed to give the threshold intensity versus

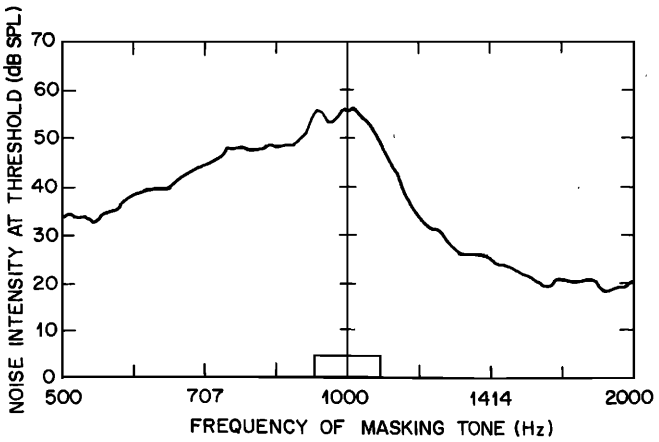

FIG. 1. Auditory threshold for a critical-band noise burst centered at 1 kHz masked by a tone of intensity 80 dB SPL. The frequency band occupied by the noise is indicated by the rectangular shaded area. Note that for a tone frequency of 1 kHz the noise intensity at threshold is 24 dB below the tone intensity. The masked threshold drops more steeply when the tone frequency is raised than when it is lowered, corresponding to the usual frequency asymmetry of auditory masking. Subject: JLH.

frequency curve shown in Fig. 1.

This curve is in approximate agreement with the predictions of Eqs. (1)–(9). Threshold intensity of the noise burst masked by a 1-kHz tone is approximately 56 dB SPL (24 dB below the tone). Threshold intensity drops off sharply, at between 20 and 30 dB per critical band, as the tone frequency increases above 1 kHz, and less sharply, at approximately 10 dB per critical band, as the tone frequency decreases below 1 kHz.

We have obtained similar results with two subjects for noise bursts centered at 400 Hz, 1 kHz, and 2.5 kHz, with bandwidths of 0.25 and 1.0 critical band. When the noise bandwidth is increased to four critical bands, threshold intensity for a burst masked by a tone at the center of the noise band decreases and the threshold noise intensity changes less rapidly as the frequency of the masking tone moves away from the noise center frequency, in agreement with the above equations.

As mentioned above, we have been able to facilitate suprathreshold (above the threshold of hearing) loudness comparisons by using a pulsating noise. The noise burst pulsates continuously (typically 1 s on, 200 ms off) while the masking tone alternates on for some number of bursts (typically three) and off for the same number. The intensity of the masking tone and of the masked noise burst is fixed, and the subject adjusts the intensity of the unmasked noise burst so that it matches the loudness of the masked noise burst. Subjects find the task easy and give reproducible results.

Figure 2 shows matching intensity of the unmasked noise burst versus intensity of the masked noise burst, for various masking-tone frequencies. Conditions are the same as in Fig. 1: The noise burst is one critical-band wide centered at 1 kHz, and the intensity of the masking tone is 80 dB SPL. When the masking tone is

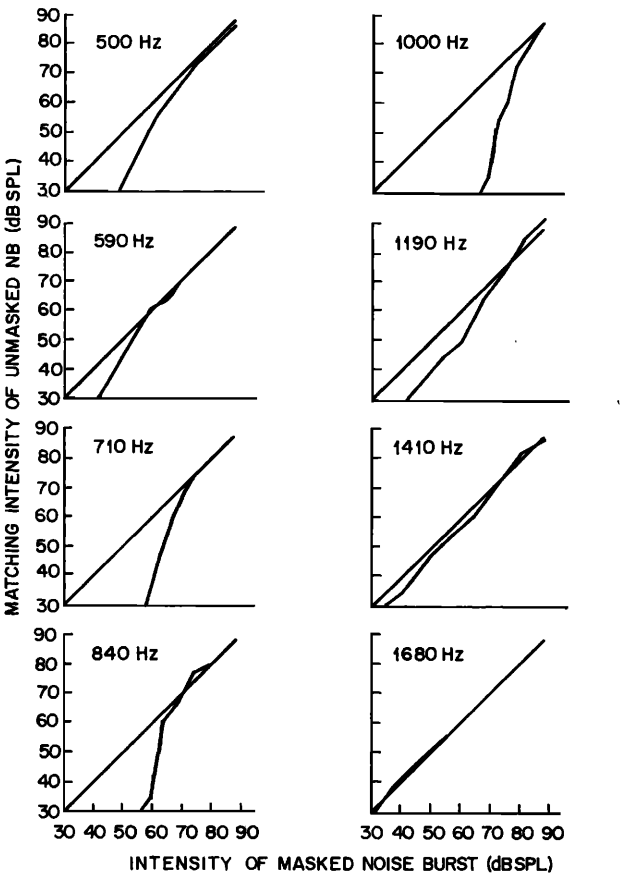

FIG. 2. Suprathreshold loudness measurements for a critical-band noise burst centered at 1 kHz masked by a tone of intensity 80 dB SPL and frequency as indicated in each panel. Note that for a tone frequency of 1 kHz the matching intensity of the unmasked noise burst decreases with a slope of 3 dB/dB when the intensity of the masked noise burst is reduced below 80 dB SPL, in agreement with Eq. (9) with $p = 2$. Subject: JLH.

at or below 1 kHz, matching intensity of the unmasked noise burst drops off sharply as intensity of the masked noise burst is decreased, with a slope of about 3 dB/dB, and the corner intensity below which the noise is reduced in loudness by the masking tone decreases as the frequency of the masking tone decreases. These results are generally consistent with Eq. (9).

When the masking tone is above 1 kHz the matching intensity of the unmasked noise burst drops off less sharply, and there is some indication that the corner intensity does not decrease substantially. The treatment of step (6) would have to be modified to account for these results. However, at the present time we have these masking data from only one subject. The experiment will have to be repeated with other subjects before any modification of Eq. (9) is justified.

V. APPLICATION TO THE DESIGN OF DIGITAL SPEECH CODERS

Adaptive predictive coding of speech signals (Atal and Schroeder 1967, 1970, 1979b) is a powerful method of reducing the information rate (in bits/s) of a speech signal while maintaining high quality. The information rate is reduced by quantizing only that part of the signal which cannot be predicted from the already coded signal. The

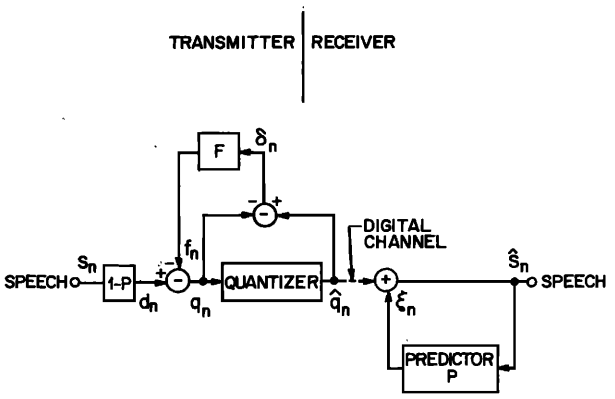

FIG. 3. Block diagram of a generalized predictive coder.

distortion engendered by the quantization in the coding process is a prime example of a signal degradation amenable to minimization by the procedure described above. In fact, for sufficiently high bit rates it will be possible to make the quantizing noise completely inaudible. This is accomplished by shaping the noise spectrum such that for all frequencies of interest the inequality

$$Q(x) \leq M(x) \tag{15}$$

is fulfilled. The noise spectrum shaping can be realized by quantizing noise feedback filters (around the quantizer) (Atal and Schroeder, 1979a).

For lower bit rates, it will not be possible to make the quantizing noise completely inaudible. In this case the loudness L_N of the quantizing noise relative to the loudness of the speech signal L_S should be minimized. The minimization of L_N/L_S is realized by noise spectrum shaping through quantizing error feedback, as illustrated in Fig. 3.

Figure 3 shows a modified version of the original adaptive predictive coder (APC) as redrawn to bring into evidence the two functions of APC:

(1) Spectral flattening of the speech signal, realized by the "inverse" filter $1-P$. [Thus, the signal samples to be quantized are (nearly) uncorrelated.]

(2) Quantizing noise feedback realized by the filter F.

In the circuit equivalent to the original APC, the error feedback filter F was equal to the predictor filter P, to minimize quantizing noise *power*. The resulting quantizing noise spectrum was flat but did not have minimum possible loudness. (Especially in the interformant frequency ranges, the quantizing noise can be quite audible.)

The shape of the noise spectrum for a general feedback filter F is approximately proportional to

$$|1-F|^2/|1-P|^2.$$

Details of computing the filter coefficients of the finite-impulse-response filter F can be found in Atal and Schroeder (1979a). Figure 4 shows the resulting minimum-loudness noise spectrum for a vowel sound.

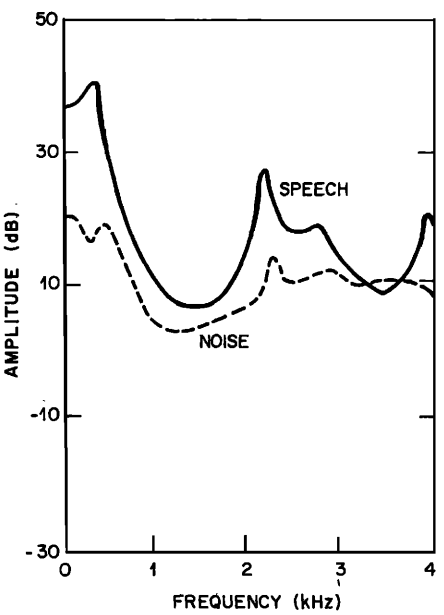

FIG. 4. An example of speech and quantizing noise spectra.

VI. CONCLUDING REMARKS

We have described a method of calculating an objective measure of signal degradation based on known or measurable properties of auditory perception. This method has already been used successfully in the design of adaptive predictive coders yielding audibly improved speech quality at bit rates of 16 000 and 9600 bits/s.

Atal, B. S., and Schroeder, M. R. (1967). "Predictive coding of speech signals," Proc. 1967 Conf. on Communication and Processing, Cambridge, MA, pp. 360–361.

Atal, B. S. and Schroeder, M. R. (1970). "Adaptive predictive coding of speech signals," Bell System Tech. J. **49**, 1973–1986.

Atal, B. S., and Schroeder, M. R. (1979a). "Optimizing predictive coders for minimum audible noise," Conference Record IEEE Intl. Conf. on Acoustics, Speech & Signal Processing, Washington, DC, pp. 453–455.

Atal, B. S., and Schroeder, M. R. (1979b). "Predictive coding of speech signals and subjective error criteria," IEEE Trans. Acoustics, Speech Signal Proc. ASSP-27, 247–254.

Blauert, J. (1974). *Raumliches Hören* (Hirzel, Stuttgart).

Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer, New York).

Fujimura, O. (1972). "An approximation to voice aperiodicity," IEEE Trans. Audio Electroacoust. **AU-16**, 66–72.

Greenwood, D. D. (1961). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**, 1344–1356.

Hellman, R. P. (1972). "Asymmetry in masking between noise and tone," Perception and Psychophys. **11**, 241–246.

Makhoul, J., Schwartz, R., and A. W. F. Huggins (1978). "A mixed-source model for speech compression and synthesis," Proc. 1978 IEEE Internatl. Conf. on Acoustics, Speech and Signal Processing, Tulsa, OK, pp. 163–166.

Mehrgardt, S., and Schroeder, M. R. (1978). "Die Wahrehmbarkeitsschwelle von Unterschieden beim Paarvergleich komplexer Signale," in *Fortschritte der Akustik*, DAGA 78 (VDE-Verlag, Duesseldorf), pp. 515–518.

Schroeder, M. R. (1966). "Vocoders: Analysis and Synthesis of Speech—A review of 30 years of applied speech research" Proc. IEEE **54**, 720–734.
Schroeder, M. R. (1975). "Models of Hearing," Proc. IEEE **63**, 1332–1350.
Schroeder, M. R. (1977). In *Recognition of Complex Acoustic Signals*. Life Sciences Research Report 5, edited by T. H. Bullock (Dahlem Konferenzen) (Abakon Verlag, Berlin) pp. 323–328.
Schroeder, M. R. and Hall, J. L. (1974). "A model for mechanical to neural transduction in the auditory receptor," J. Acoust. Soc. Am. **55**, 1055–1080.
Tribolet, J. M., Noll, P. McDermott, B. J., and Crochiere, R. E. (1978). "A Study of complexity and quality of speech waveform coders," Proc. 1978 IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Tulsa, OK, pp. 586–590.
Zwicker, E. (1963). "Ueber die Lautheit von ungedrosselten und gedrosselten Schallen," Acustica **13**, 194–211.
Zwicker, E., and Feldtkeller, R. (1967). *Das Ohr als Nachrichtenempfänger* (Hirzel, Stuttgart).