Speech and Hearing: Some Important Interactions

Manfred R. Schroeder
Göttingen, FRG and Murray Hill, USA

Abstract

I will discuss the following three topics at the interface of speech and hearing:

1. The importance of auditory masking by the speech signal on the perception of speech itself. This 'self-masking' of speech is also important in the synthesis of speech and its digital encoding. In particular, the sizable quantizing noise at the low bit rates desirable for efficient digital transmission and storage of speech signals can be made nearly inaudible by exploiting the masking properties of the human ear. (Collaborators: B.S. Atal and J.L. Hall)
2. Work at Göttingen on processing of speech signals to enhance their intelligibility for the hard-of-hearing, particularly those suffering from sensorineural hearing loss and recruitment. (Collaborators: W.H. Strube, T. Langhans and T. Lewien)
3. Effects of phase on the perception of speech, including the possibility of creating intelligible speech from signals with time-invariant flat power spectra solely by controlling monaural phase. These results point to the importance of temporal cues in the perception of speech beyond presently accepted limits. (Collaborator: S. Mehrgardt).

1. Introduction

I was raised as a physicist and thus, I feel a bit out of place at a Phonetics Congress - but not entirely. In 1953 - 30 years ago - Prof. Werner Meyer-Eppler from the Phonetics Institute in Bonn gave a colloquium talk at the University of Göttingen and I (a young student then) was much impressed by what he had to say about speech and hearing, about linguistics and information theory. My physicist friends pretended that they did not understand what Meyer-Eppler was talking about, but I was so taken with his message that a year later - upon joining Bell Laboratories - I changed my primary research field from microwave physics and physical acoustics to speech.

I remember building the first voice-excited vocoder (VEV) - not from integrated circuits, but from individual components, soldering iron in hand. We discovered that the cepstrum technique - originally suggested to distinguish earthquakes from nuclear explosions - was an ideal tool for the detection of the fundamental frequency of speech.

We manipulated formant frequencies, segmental durations and pitch of natural speech to create the most unnatural speech signals anyone had ever heard. In fact, taking a leave from the great Creator, we tried to change a male voice into a female voice and discovered there is more to the male-female dichotomy than pitch and formant frequencies. (Even after getting formant bandwidths right, our female creations did not sound very inviting - say for sharing an evening at the opera.)

In our work on analyzing and synthesizing speech we were constantly reminded how important hearing - or more generally: auditory perception - is. Let me give you just one small example. At some point in my speech career it occurred to me that the signal-to-noise power ratio (SNR) of voiced speech contaminated by white noise could be enhanced a hundredfold by a pitch-tracking comb-filter. And the physical measurements on a noisy speech signal so processed showed that, indeed, the SNR was improved by more than 20 dB. But when it came to assessing the improvement in speech intelligibility - a subjective measure - the result was nil: the processed speech, although sounding less noisy, was not a bit more intelligible than the unfiltered signal. In other words, whatever my pitch-tracking comb-filter did, the listener's brain could do just as well, using its own biological 'software'.

2. Auditory Masking by Speech Signals

While this early lesson on the interaction of speech and hearing was negative from the point of view of a practical application (enhancement of intelligibility) another interplay between speech and hearing has paid off very handsomely: the exploitation of auditory masking to reduce the audibility of quantizing noise in digital speech. When we first started to apply the principle of linear predictive coding (LPC), everybody was gratified by the resulting high speech quality. Only at very low bit rates some distortion, caused by quantizing noise, could be heard. Then it occurred to me that even this distortion could be eliminated by sweeping the quantizing noise under the 'formant rug', so to speak. Quantizing noise in linear predictive coding typically has a flat spectrum as shown in Fig. 1. But by computing the loudness of the noise in the presence of the speech signal and then minimizing it, resulting in a noise spectrum as shown in Fig. 2, we can make the noise practically inaudible. Even at 1 bit/sample for the prediction residual signal, the quantizing noise is inaudible and input and output speech signals are virtually indistinguishable. Fig. 3 shows some of the steps of this computation. Most of this work was done in collaboration with J.L. Hall and B.S. Atal of Bell Laboratories and P. Calavrytinos at Göttingen (Schroeder et al., 1979*a*, 1979*b*, 1982).

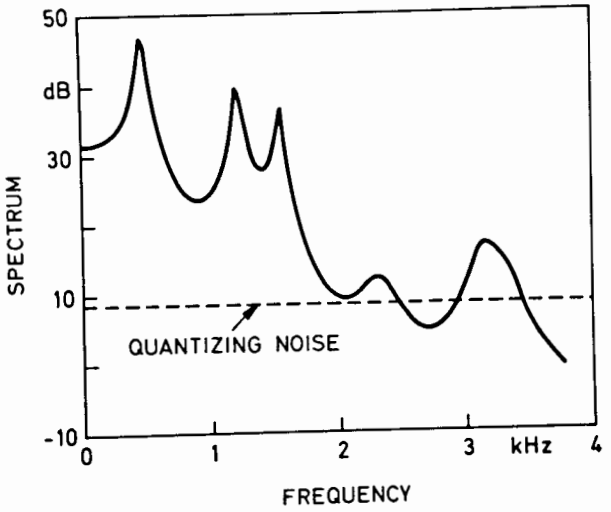

Fig. 1. Quantizing noise in LPC.

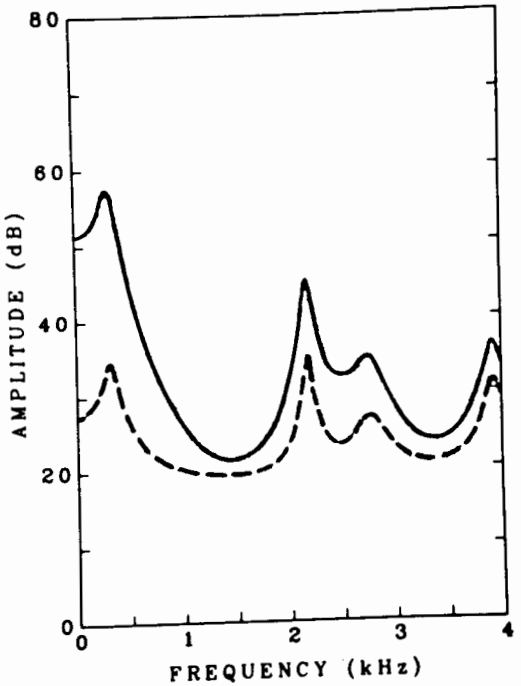

Fig. 2. Minimized loudness of noise in the presence of the speech signal.

SPECTRAL ANALYSIS	$S(t) \longrightarrow \tilde{S}(f)$
RELATION BETWEEN FREQUENCY AND TONALNESS	$f = 650 \sinh\left(\frac{x}{7}\right)$
TRANSFORMATION FROM Hz to Bark	$\tilde{S}(f) \longrightarrow S(x)$
NEURAL EXCITATION FOR SPEECH SIGNAL	$E(x) = S(x) * B(x)$
NEURAL EXCITATION FOR QUANTIZING NOISE	$Q(x) = N(x) * B(x)$

Fig. 3. Some of the steps in the computation.

3. Enhancement of Temporal Cues

I mentioned Meyer-Eppler before and will mention him again. When I visited his Institute at the University of Bonn in 1958, he showed me - among other things - an old German patent of his: a kind of vowel-consonant switch that would increase the power level of consonants in connected speech to prevent them from being masked by preceding vowel sounds in a reverberant auditorium. A beautiful idea - except it did not work. The constant switching made the original speech signal rather unintelligible even before it was subjected to reverberation.

Later, when I learned about the modulation transfer function (MTF), I thought I could outwit reverberation by *increasing* the modulation envelope of speech signals - but to no avail. Nevertheless, manipulating the envelope of speech signals has led to success in another application - improving speech intelligibility for the neurally deaf - more specifically, people who suffer from recruitment. Such patients have only a very narrow level range between the thresholds of hearing and discomfort. If speech is to be intelligible to them, its amplitude has to be fitted into the narrow level 'window'. This would call for amplitude compression. However, such compression would decrease an already marginal intelligibility, because much speech articulation is reflected not in stationary amplitudes and spectra, but in *temporal* cues and fast *transitions* between stationary states. Thus, a more sophisticated kind of modulation processing is called for as illustrated in Fig. 4, which shows different modulation transfers (M), versus modulation frequency Ω. Panel 1 corresponds to an unmodified MTF and panel 2 illustrates ordinary dynamic compression, in which all modulation frequencies are equally attenuated (by about 12 dB in the illustration). The third panel shows a reduction of the

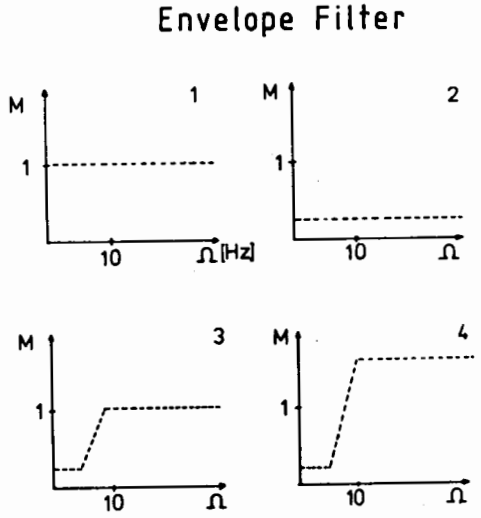

Fig. 4. Different modulation transfers (M) vs. modulation frequency.

modulation at low modulation frequencies (below 5 Hz). The fourth panel shows both this reduction *and* an increase in the modulation at high modulation frequencies, thereby both compressing the dynamic range (determined mostly by the low modulation frequencies) and enhancing the transients. (The corresponding operation in vision is called *edge enhancement*.)

The modification of the modulation spectrum of speech is done separately for each critical band of hearing. The success of this method in enhancing speech intelligibility, both in noisy environments and for recruitment patients, was demonstrated by H.W. Strube, T. Lewien, and T. Langhans and described at the 1982 FASE/DAGA Congress in Göttingen.

4. The Importance of the Time Waveform of Speech Signals

In the 1950s, when I first became interested in speech synthesis, I was almost immediately intrigued by the problems of subjective quality of synthetic speech. Vocoders had a reedy, 'electronic' accent and I thought that the excitation waveform, consisting of sharp pulses for voiced sounds, was perhaps to blame. To investigate this question more deeply, I built a generator for 31 coherent harmonics of variable fundamental frequency. The phase of each harmonic could be chosen to be either 0 or π - a total of $2^{30} = 1{,}073{,}741{,}824$ different waveforms, each of which appeared to have its own intrinsic timbre - their identical power spectra notwithstanding. (I wish Seebeck, Ohm and Helmholtz had had a chance to listen to these stimuli!)

For all phase angles set equal to 0, one obtains a periodic cosine-pulse, see Fig. 5. When this waveform is used as an excitation signal for a speech synthesizer, the result is the reedy quality already mentioned. By contrast, if

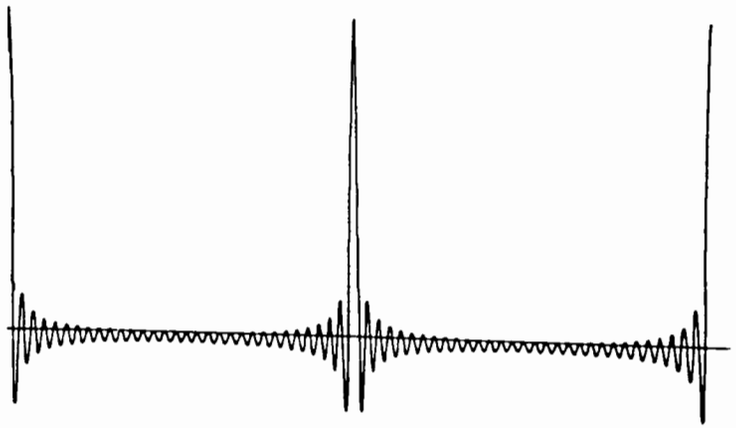

Fig. 5. A periodic cosine-pulse, with phase angles set equal to 0.

one randomizes the phase angles, one gets a less peaky waveform, see Fig. 6, and a mellower sound (Schroeder, 1959). A better-than random choice for the phase angles (one that gives an even less peaky waveform) is given by the formula

$$\varphi_n = \pi n^2/N$$

where n is the harmonic number and N the total number of harmonics in the flat spectrum stimulus. More general formulae, for arbitrary and phase angles restricted to 0 or π, are given in Schroeder, 1970.

Many of the waveforms generated by phase manipulation, although they had smooth spectra without formant structure, had a *vowel*-like quality; I shall return to this astounding observation in a moment. Let me first describe an auditory masking experiment performed by Sönke Mehrgardt at our Institute, which gave an astonishing monaural phase effect.

Mehrgardt took a harmonic tone complex with a speech-like power spectrum. The phase angles of the individual harmonics were either constant or random. The two resulting signals, one reedy sounding, the other with a vowel-like quality, were used to mask a 1200-Hz pure tone. The masked thresholds are shown in Fig. 7. For the random-phase masker, the masked threshold increases with decreasing fundamental frequency of the masker -

Fig. 6. A periodic cosine-pulse, with randomized phase angles.

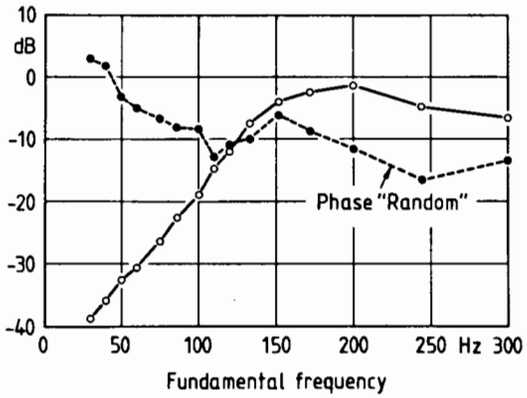

Fig. 7. Direct masking threshold of test tone (1200Hz) masked by a harmonic complex.

as expected, because more masker harmonics fall into the critical band around the maskee (the test tone). By contrast for the constant-phase masker (open circles) the masked threshold *decreases* with decreasing fundamental frequency. A paradox, a 40-dB monaural phase effect!

How is this possible, considering that G.S. Ohm and H.v. Helmholtz argued about whether there are *any* monaural phase effect at all? A possible explanation is illustrated in Figs. 8 and 9, which show the effects on the time waveforms of doubling the 7th harmonic. While this is clearly visible in the case of the constant-phase spectrum (Fig. 8) no such obvious extra signal at the 7th harmonic is detectable in the case of the 'random'-phase masker (Fig. 9).

But it might be argued that higher centers in the auditory pathway do not 'see' the waveforms shown in these two figures. The inner ear performs a critical-band analysis, resulting in signals as illustrated in Fig. 10.

However, after this bandpass filtering by the ear there remain pronounced waveform differences between constant and 'random'-phase signals. For the constant-phase masker, there are still time gaps in the waveform during which the presence of a test-tone could be be detected, especially at very low fundamental frequency, confirming the result of Fig. 7. If this explanation is to hold water, then a short tone pulse used as a signal should show characteristic variations of the masked threshold with time reflecting the *envelope of the masker*. This is indeed observed as shown in Fig. 11 (Mehrgardt, 1982).

I will not pursue here the possibility of using this experimental paradigm to explore the temporal response characteristics of the inner ear. Rather, I want to stress the importance of temporal detail in speech-like waveforms. These results suggest the existence of short time-windows during which the human ear can 'look' at waveforms and perceive important differences even for given amplitude spectra. If this is true, it should be possible to create speech-like percepts for signals having smooth spectra *without formant structure*.

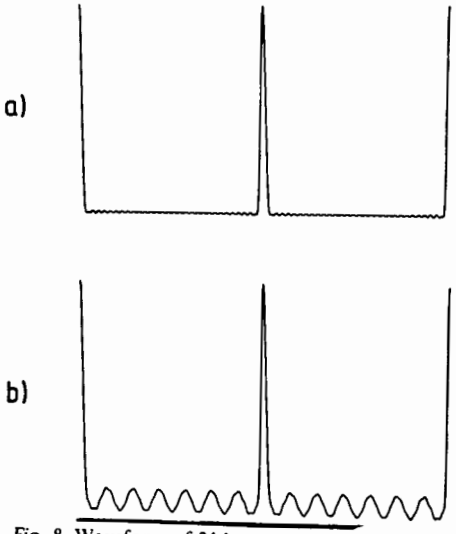

Fig. 8. Waveform of 31 harmonics in zero phase. a) amplitudes Hamming-weighed; b) amplitudes as in a), but amplitude of seventh harmonic doubled.

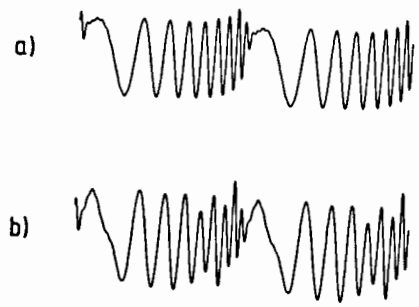

Fig. 9. Waveform of 31 harmonics with "random" phase. a) amplitudes Hamming-weighed; b) amplitudes as in a), but amplitude of seventh harmonic doubled.

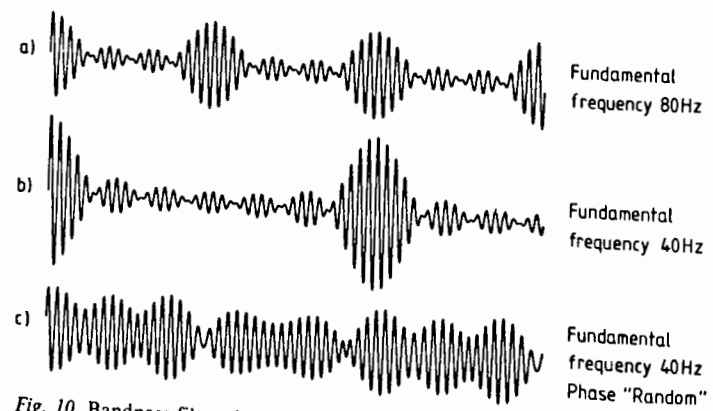

Fig. 10. Bandpass filtered harmonic complex (1/3 octave bandpass, centered at 1500Hz).

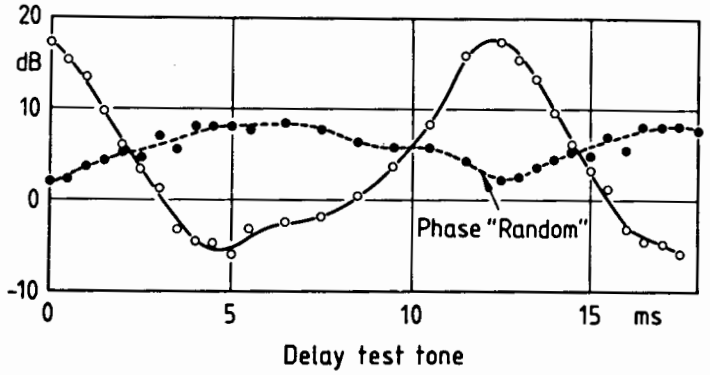

Fig. 11. Direct masking threshold of test tone burst (1600Hz, 5ms) masked by a harmonic complex (fundamental frequency 80Hz).

Such an experiment was performed by Dr. Strube at our institute who synthesized *flat*-spectrum signals corresponding to the three diphthongs /au/, /oI/, and /aI/. This is accomplished by adjusting phase angles below and above the drawn-in vowel spectrum as shown in Fig. 12. All three signals sound clearly speech-like, although there is no formant structure. In fact, they clearly sound like the diphthongs underlying the synthesis. As expected, the diphthongs-like percept is more pronounced at lower fundamental frequencies, consistent with the time-window model.

But is this time-window a complete explanation? More recent results by Mehrgardt hint that it is not. The different maskers illustrated in Fig. 13 all should give the same threshold for the test-tone pulse; but they do not, as shown in Fig. 14. Rather, as the random frequency shifts of the masker components are increased, the threshold increases by about 18 dB although the time gap during which the test tone occurs is clearly defined for all conditions. No such increase of threshold is observed if all frequency components are shifted by the same amount. These results suggest that the ear needs

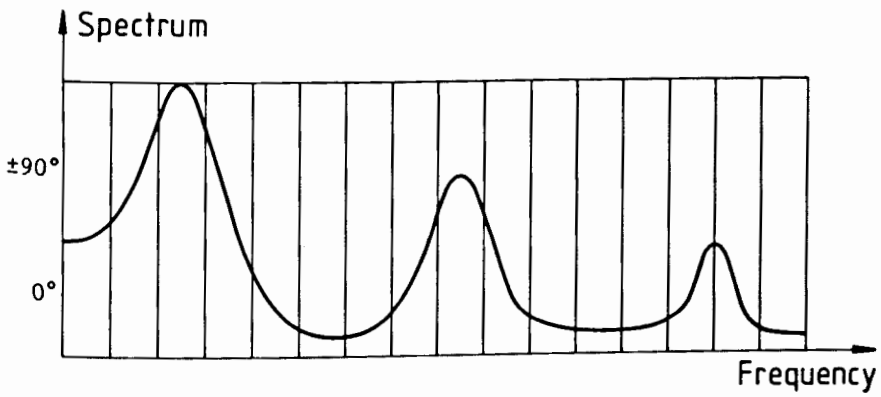

Fig. 12. Flat spectrum signal with adjusted phase angles.

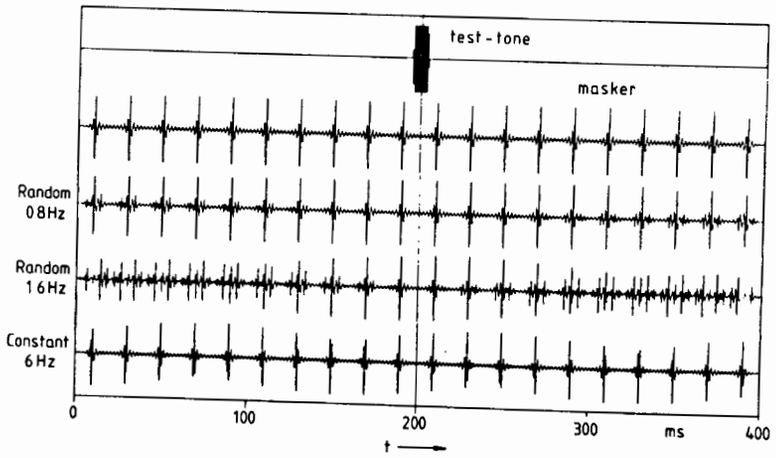

Fig. 13. Different maskers with test tone pulse.

many time gaps (and not just a single one) to optimally detect a single test-tone pulse!

It looks as if *waveform* memory is involved. Is this possible, or even thinkable? We have reached a point where only further experiments can help us. We know quite a bit about the auditory periphery, but woefully little about the higher interpretive functions of hearing. Why do two voices sound like two voices under most conditions? How does our brain integrate the separate harmonic components of a given speech signal into a *single* vowel percept? Precisely under what conditions does such integration take place? Here are some of the most interesting questions for future research.

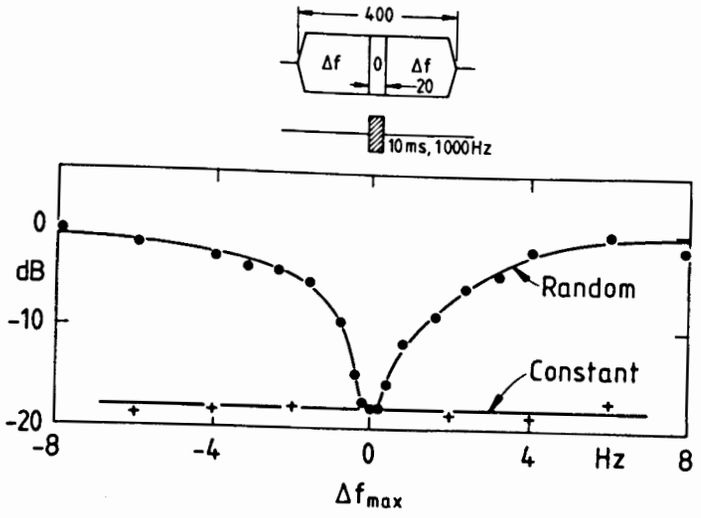

Fig. 14. Increase of threshold with increased random frequency shift of the masker.

What, in speech perception do we hear as 'figure' and what as 'background', to use terms familiar from visual perception. Let me illustrate from an example from language comprehension: There are hundreds of words spelled alike in English and German but having different meanings. I once wrote a little German story using only such words and showed it to a German-speaking friend in the United States and asked him what he thought of it. Answer: 'Nothing, a random collection of words'. Half a year later in *Germany* I showed the same friend the same story and asked him again what he thought of it. The answer this time: 'Quite an interesting story, who wrote it? When I asked him whether he saw anything unusual in the words, the answer was: 'No, I don't see anything.'

As an example of a possibly interesting figure background experiment in speech perception let me sketch Fig. 15. Again we have a flat spectrum with a vowel contour drawn. The harmonic frequency components below and above the vowel line are altered either randomly or more or less coherently along the frequency axis, either by fixed amounts or in a time-varying manner (jitter). At what degree of coherence or jitter do we hear a vowel-like sound? Or will we hear two vowels - or none?

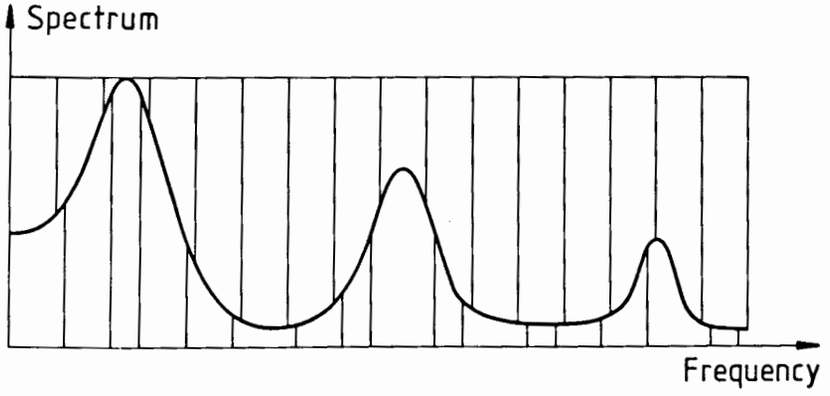

Fig. 15. Flat spectrum signal with random alternations below and above vowel line.

5. Conclusion

What is our main conclusion? We have reaped nice benefits on the basis of our present knowledge. We have made quantizing noise in digital speech practically inaudible even at very low bit rates. We have manipulated the modulation of speech intelligibility in noise and for certain kinds of hearing disorders. And we have shown the importance of temporal cues, beyond prior expectation. But there are still large blank areas on our map of understanding and our most interesting insights into speech and auditory perception are yet to come.

References

Mehrgardt, S. (1982). Kanteneffekte der Mithoerschwelle bei Hochpass- und Tiefpass-Maskierern aus harmonischen Tonkomplexen. In: Fortschritte der Akustik - FASE/DAGA '82, Göttingen: Drittes Physik. Inst., 1223-1234. Göttingen.

Schroeder, M.R. (1959). New results concerning monaural phase sensitivity. *Journal of the Acoustical Society of America,* **31**, 1579(A). See also: Duifhuis, H. (1970). Audibility of high harmonics in a periodic pulse. *Journal of the Acoustical Society of America,* **48**, 888 and Duifhuis, H. (1971). Audibility of high harmonics in a periodic pulse. II. Time Effect. *Journal of the Acoustical Society of America,* **49**, 1155.

Schroeder, M.R. (1970). Synthesis of low-peak-factor signals and binary sequences with low autocorrelation. *IEEE Transactions on Information Theory*, IT-16, No. 1, January 1970.

Schroeder M.R., Atal, B.S., and Hall, J.L. (1979*a*). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In: *Frontiers of Speech Communication Research.* B. Lindblom and S. Öhman, eds., London: Academic Press, no. 217-229.

Schroeder. M.R.. Atal. B.S.. and Hall, J.L. (1979*b*). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America,* **66**, 1647-1652.

Schroeder, M.R. and Atal, B.S. (1982). Speech coding using efficient block codes. *IEEE Proceedings*, International Congress on Acoustics, Speech and Signal Processing, May 1982, Paris. **3**, 1668-1671.