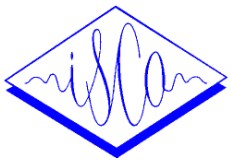

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

The audibility of narrow band noise in flat spectral complex sounds

C. Ma and L.F. Willems
Institute for Perception Research
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

Masking of narrow band noise by periodic pulse trains of different repetition rates and by synthetic vowels was investigated in this paper. Results indicate that the masking process is controlled by limited spectral resolution in the low-frequency part of the masker and by limited temporal resolution in the high-frequency part. The results are also discussed in connection with auditory correlates in speech signal processing.

I. Introduction

Masking experiments have been reported in the psychoacoustic literature for simple maskers such as pure tones, or white and narrow band noises [1]. The attention of these experiments was mainly focused on some specific properties of the human auditory system, and those simple maskers or targets can be used to serve the purpose.

However, the auditory pathway involves very complex nonlinear processing, which is not yet well understood. Generally, the behavior of the human auditory system is very much dependent on the physical properties, such as sound pressure level and spectrum of the sound waves that drive the eardrum. Those results from simple masker-target setup cannot be generalized to complex signals. Furthermore, speech sounds have enormous amounts of variations, are dynamic in nature, and spectrally complex. Therefore, the response of the auditory pathway to complex signals like speech, in general, cannot be predicted from the response to simple sinusoids.

It is the intention of the present masking study to make a contribution to the understanding of the perceptual correlate of speech processing, such as speech coding, speech synthesis, or speech manipulation. The technique of masking is used here, because it provides a threshold value of a target signal in a masker, but also sheds light on how the auditory system processes complex sounds. In speech coding, for example, quantization noise is the target, the speech sounds are the maskers, and the noise should be inaudible in the coded speech. In order to achieve our goal, a series of experiments with periodic pulses and synthetic vowels as masker and narrow band noise as the target were set up. The maskers are still simple compared with speech signals. This makes it possible to systematically study some important aspects of a signal and to facilitate the experiment. The spectrally flat signals are, for instance, also used as the excitations in speech synthesizers for voiced sounds. Based on the experiment, some auditory correlates of the speech signal can also be investigated which are related to the limitation of the auditory system such as phase sensitivity and masking. Hence, in speech signal processing, the auditory correlates of the speech signal have to be carefully dealt with and the irrelevant aspects of the signal can be ignored. We will elaborate those perceptual correlates in the next section.

II. Auditory behavior and speech processing

Digital speech coding is a major part of the speech processing and has been investigated for decades. It makes bit-rate reduction realizable by utilizing the fact that speech has much redundancy, and the auditory system has a limited spectral resolution, which leads to masking effects [1]. All the coding schemes having been explored have taken advantage of these two properties in the information transmitter and receiver to reduce the bit rate as much as possible while keeping quantization noise inaudible. For instance, in the simple PCM coding, quantization noise can be masked by the speech sound when a high bit-rate is used and thus high signal to noise ratio is maintained [2] [3]. The pioneering work by Schroeder, Atal and Hall, which is based on pure tone masking behavior of the ear, has laid the foundation of a noise hiding technique, leading to the introduction of a noise shaping filter into the coding scheme in order to match the spectrum of the noise and the speech signal[2] [10]. Furthermore, several other systems, for example subband coding, have been developed to more accurately implement a particular noise shaping [9][11] [12]. So far, however, this noise hiding technique has been derived from pure tone masking behavior only.

Speech signals are often manipulated or transformed for different purposes without damaging their subjective sound quality, such as in pitch manipulation, phase equalization and dispersion. The long-term spectrum of the speech is usually preserved after these manipulations, but the phase spectrum has often been totally changed for each pitch period of the speech [14][15][17][20]. Such manipulations and transformations take the advantage of the fact that the human auditory system seems rather insensitive to phase manipulations in understanding the utterances. However, the phase spectrum does play a role in judging the sound quality[18] and the pitch of the speech[19], but with limitations. Therefore, it is a very interesting and practical question to what degree and how phase plays a role in determining speech sound quality.

Finally, there also has been the question how to choose the excitation signal in the source-filter model to produce a natural sound. It is a well-known fact that the LPC synthesizer with pulse excitation produces mechanical sounding speech[4]. If an excitation function closer to the glottal pulse shape is chosen, more natural sounds can be obtained. Rosenberg and Holmes have carried out a systematic experiment to test the effect of glottal pulse shape on the quality of natural vowels [21] [22]. All excitations had a flat spectrum and excitations of different shape reflected differences in the phase spectrum. By fitting a function to the real glottal pulse, a perceptually preferable phase spectrum was obtained. This has also been shown in multipulse excitation LPC synthesizer[2]. All this shows that the phase spectrum plays an important role in speech sound quality.

III. Perceptual experiment

A. Stimuli

All stimuli in this investigation were produced by computer, using a sampling frequency of 20kHz and a dynamic range of 16 bits. Signals were lowpass filtered with a cutoff frequency of 7.8kHz, referred to the 3 dB attenuation point, and a attenuation slope of 90 dB/octave.

Maskers were trains of impulses synthesized by adding equal-amplitude cosine-phase harmonics of a certain

fundamental frequency, according to the following formula.

$$s(n) = \sum_{i=1}^{M} A\cos(i\omega_0 n + \psi) \quad (1)$$

where ω_0 is the fundamental frequency and M was chosen such that the masker covered a frequency range of 10 kHz. For zero phase stimuli ψ was equal to zero, and for cosine-sine alternating phase stimuli, ψ was equal to $(i-1)*\pi/2$.

In addition, synthetic vowel maskers were synthesized with impulse excitations, which consisted of equal amplitude zero-phase harmonics, and LPC filters with five formants [4]. The parameters of the LPC filters were analyzed from natural vowel sounds.

The noise target was assumed to be a stationary process, approximated by the following series [6]

$$\hat{x}_t = \sum_i c_i \exp^{j\omega_n it} \quad (2)$$

with uncorrelated coefficients c_i (random variable) and i chosen such that the above formula could produce a particular narrow band noise.

Its power spectrum was

$$S_{\hat{x}}(\omega) = 2\pi \sum_i E\{||c_i||^2\}\delta(\omega - i\omega_n) \quad (3)$$

It was found that enough sinusoids had to be used so that the tonal character of the individual components is replaced by the atonal, diffuse character that is associated with noise [7]. That also means that the fundamental ω_n had to be small. In this experiment, ω_n was equal to 4 Hz, which corresponds to a period of 250 ms. The harmonics could not be resolved in Fourier spectral analysis over such a short duration of 250 ms[8]. Therefore, the noise produced in this way was perceptually and mathematically suitable for the experiment.

It was chosen in this experiment that all c_i were equal to a constant C and i covered a frequency range of a critical bandwidth in which loudness summation applied[23]. The noise targets were located at selected center frequencies over a 5kHz frequency range. In the lower frequency region, the center frequency of the target noise band corresponded approximately to a harmonic or to the center frequency between two successive harmonics. The values of the critical bandwidth are calculated from the formula of Zwicker and Terhardt [16]:

$$CBW = 25 + 75((1.0 + 1.4 * (f_c/1000.0)^2))^{0.69},$$

where CBW is the value of the critical bandwidth and f_c is the center frequency in Hz.

The threshold of the narrow band noise in a specified frequency region can be represented as:

$$TD = 10\log(\frac{C^2\omega_0}{A^2\omega_n}) \quad (4)$$

The TD represents the ratio of the average energy of the noise to that of the signal in a 1 Hz bandwidth. In other words, it is the ratio of the power spectral densities, which is represented by S/M in the plot.

B. Method

A two interval, two alternative, forced choice(2I2AFC), adaptive procedure was used to determine thresholds [5]. Each interval contained either 200 ms of masker alone or 200ms of masker plus target noise, both intervals including 25 ms sinusoidal onset and offset ramps. The order of the two intervals was randomized and the maskers were presented at a sound pressure level of 80 dB to the subjects. The pause between two intervals was 500ms. Three experienced subjects with normal hearing listened diotically in a sound-proof room through ETYMŌTIC RESEARCH ER-2 insert earphones. The level of the target noise was initially well above threshold. Using a two-down one-up procedure, the first run had a step size of 8 dB and the second run had a step size of 4 dB, in order to quickly reach the threshold value. Starting from the third run, two consecutive correct responses made the noise level decrease by 2 dB, while after each incorrect repose it was increased by 2 dB. This procedure estimates the 70.7% correct response point of the psychometric function. The average of the midpoints, excluding the first three points, of every second run was accepted as the threshold level. Fourteen runs were taken for each data point. The response time was controlled by the subjects.

C. Experiments

In the first experiment, flat-spectrum maskers with zero phase were used and fundamental frequencies of 100Hz, 150Hz, 200Hz, 250Hz and 400Hz were chosen.

The second experiment was designed to investigate the effect of the maskers' phase. Maskers with a 100Hz fundamental frequency and cosine-sine alternating phase were used.

Finally, vowel-sound-maskers were investigated. The vowel sounds were synthesized by using impulse trains of fundamental frequencies 100Hz and 200Hz as the input to the LPC filter. Maskers of this kind have both amplitude and phase complexity.

IV. Results and discussion

A. Results

The results from three subjects were similar, and thus the average results of three subjects are presented. The standard deviations are also given. For the first experiment, with maskers of fundamental frequencies of 100Hz, 150Hz, 200Hz, 250Hz and 400Hz, average thresholds are plotted in Figs.1a,c,d,e and f. The x-axis represents the center frequency of the bandpass noise and the y-axis represents the detection threshold. The figures are arranged in this manner for the convenience of later comparisons.

One observes that the detection threshold decreases by approximately 9dB/oct at high center frequency for maskers of fundamental frequencies of 100Hz,150Hz, 200Hz, but this disappears for maskers of fundamental frequencies of 250 and 400Hz. In addition, the higher the fundamental frequency, the higher the rolloff frequency where the detection threshold starts to decrease by 9dB/oct. It is also obvious from Figs. 1c,d,e,f that the threshold of the narrow band noise in the lower frequency region shows dips and peaks.

Results of the second experiment are shown in Fig.1b. It indicates that masking behaviour of the alternating phase masker of fundamental frequency of 100Hz is at low center frequency similar to that of maskers of the same fundamental frequency with zero phase. However, the masking thresholds at high center frequencies are very close to those produced by the zero-phase maskers of fundamental frequency of 200Hz.

The results of the third experiment, plotted in Figs.1g (f_0=100Hz) and 1h (f_0=200Hz) indicate that formant peaks are apparent in the masking patterns. They are marked by arrows. There is no decrease of 9dB per octave however, even in the vowel sound of fundamental frequency of 100Hz.

B. Discussion

In principle, the experimental results show that the masking threshold of the narrow band noise is dependent on both the center frequency of the noise and the phase and the fundamental frequency of the masker.

In Fig. 1a, and 1b there is a distinct difference between the masking patterns of zero-phase and alternating-phase masker of fundamental frequency of 100Hz in high frequencies, but there is no significant difference in low frequencies. Over the lower frequency range the spectral resolution plays a dominant role in determining the masking thresholds and the influences of the phase can be neglected. The auditory system resolves the harmonics in the lower frequency region, which cannot interact and produce complex phase-dependent time patterns. The masking patterns show clear peaks and dips for the pulse trains of high fundamental frequencies in the low frequency region. This is because the critical bandwidth is smaller than the fundamental frequency and the spacing between harmonics, making the masking patterns of individual masker harmonics visible. The number of this resolved harmonics is about three or four. The spectral analysis dominates to very high frequencies in the process of Fig. 1e and 1f.

The phase of the maskers show clear influence on the masking pattern in the higher frequency region because enough harmonics in a wide critical band can result in deeply modulated temporal signal structures. The response of the auditory filters to the pulse maskers in high frequency dies away well within a pitch period, which leaves a "silent" interval. It is assumed that detection happens mainly in the silent intervals. If the bandwidth of the filter is increased, i.e. the center frequency of the noise is increased, the silent interval will be increased. The energy of the target noise is integrated and it leads to a decrease of threshold by -9dB/oct in spectrum. This temporal analysis dominates from very low frequencies in the process of Fig. 1a, 1b and 1c. The spectral resolution degrades towards high center frequencies and the temporal resolution degrades towards lower center frequencies. Masking patterns therefore show plateaux in the middle frequency region in Fig. 1b,1c,1d and 1e, but the plateau does not appear in the Fig. 1a and 1f, corresponding to maskers of fundamental frequency of 100Hz and 400Hz.

Vowel sounds have a complex spectrum and their masking patterns can vary in many ways. In Fig. 1g and 1h the masking pattern shows peaks and valleys that correspond to the formant spectrum. This reflects the auditory representation of speech [26]. However, the masking pattern does not have the -9dB per octave decrease, compared with pulse train maskers. This is because the spectra of the vowel sounds generally have a -6dB roll off and a dispersed phase spectrum. The temporal masking effect are therefore reduced[13].

These results agree with the idea that ear is sensitive to the phase relationships among spectral components within a critical band and can detect the temporal fine structures in high frequencies where the impulse responses of the auditory filter have fast decay, in other words, the ringing has died out well within the pitch period of the masker pulses [18][24][25]. Therefore, this temporal analysis ability strongly depends on the pitch period of the masker. Duifhuis has also investigated this temporal behavior and found the same phenomenon [24]. The flat-spectrum vowel experiment by Schroeder and a more systematic phase-vowel experiment by Traunmüller have also demonstrated that our ears have the ability to perform a temporal structure analysis [27][28]. The important aspects for the temporal analysis in recognizing vowel sounds is that, at least, the phase information of the second formant must not be masked by other vowel components, i.e., vowel sounds of this kind should have a low fundamental frequency and a high level of the second formant. In speech manipulation and synthesis, the phase spectrum in high frequencies have to be carefully considered.

The auditory system has often been described by five successive stages: a bank of linear critical bandpass filters, each followed by a memoryless nonlinear element, an adaptation process, a low pass filter, and a decision device [29]. There are many different versions of this model. However, the bandpass filters are decisive to the temporal resolution and the low pass filter determines the integration time. The data obtained from the present experiments can be qualitatively explained by this model. But the temporal behavior also involves some nonlinear adaptation processes [13] [29] and thus needs further investigations.

V. Conclusion

The results have shown that the masking patterns of complex signals strongly depend on both amplitude and phase spectrum of the maskers. However, in the low frequency region of the maskers, the amplitude spectrum is a determinant for the masking pattern and phase effects are negligible. The upper frequency of this region is dependent on the fundamental frequency of the masker. In speech coding systems, the quantization noise can be hidden by taking advantage of this auditory masking property. But the noise shaping technique [10], which is based on pure tone masking patterns and thus considers only the amplitude spectrum cannot be easily justified by the results. In the high frequency region, the time pattern or temporal structure of the signal becomes more decisive in determining masking threshold. Therefore, in phase and pitch manipulations and speech coding, the high frequency region of the sounds have much more influence on sound quality. Pitch manipulations, for example, change the waveform very much, which can be shown that the short-time spectrum of the speech is greatly changed because the short spectrum analysis window have good time resolution, but the long-term spectrum envelope is almost unchanged[20]. This changes of the temporal structure is only prominent in the high frequency region. However, the -6 dB/ oct roll off of the spectrum of vowel sounds makes these quality change of the sounds less audible. Care has to be taken for the transients in speech, which generally have much energy in high frequency region. Further investigation will be carried out towards the temporal factor in the maskers.

VI. Acknowledgement

The constructive criticism and the helpful comments of Prof. Dr. A.J.M. Houtsma and Dr. Armin Kohlrausch of IPO are gratefully acknowledged.

References

[1] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, Springer-Verlag Berlin Heidelberg, 1990.

[2] B. S. Atal, "Speech coding and Human Speech Perception", In: Working Models of Human Perception Ed. by B.A.G.Elsendoorn and H. Bouma, London:Academic Press. 1988.

[3] N.S. Jayant and P. Noll, Digital coding of the waveforms, Prentice Hall, Englewood Cliffs, New Jersey, 1984.

[4] J. D. Markel and A. H. Gray, Linear Prediction, 2nd ed. The Hague, The Netherlands:Mouton, 1970.

[5] H. Levitt, "Transformed up-down method in psychoacoustics," J. Acoust. Soc. Am. 49,467-477, 1971.

[6] A.Papoulis, Probability, Random Variables, and Stochastic processes, 461-465, New York: McGraw-Hill.

[7] T.H. Schafer,R.S. Gales, C.A. Shewmaker, and P.O. Thompson, "The frequency selectivity of the ear as determined by masking experiments," J. Acoust. Soc. Am. 22,490-496, 1950.

[8] F. J. Harris, "On the Use of Windows for Harmonic Analysis With the Discrete Fourier Transform", Proc. IEEE, Vol.66, No.1, 51-83, 1978.

[9] M. A. Krasner, "The critical band coder–digital encoding of speech signals based on the perceptual requirements of the auditory system," Proc. ICASSP, 327-331, 1980.

[10] M.R. Schroeder, B.S.Atal and J.L. Hall, "Optimizing Digital Speech Coder by Exploiting Masking Properties of the Human Ear", J. Acoust. Soc. Am. 66(6), 1647-1652, 1979.

[11] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE Journal on selected areas in communication, Vol. 6, No.2, 1988.

[12] R.N.J. Veldhuis, M. Breeuwer and R. V. D. Waal, "Subband coding of digital audio signals without loss of quality", Proc. ICASSP, 2009-2012, 1989

[13] W. Jesteadt, S.P. Bacon and J.R. Lehman "Forward masking as a function of frequency, masker level, and signal delay," J. Acoust. Soc. Am. 71, 950-962, 1982.

[14] T. Moriya and M. Honda, " Speech coder phase equalization and vector quantization", Proc. ICASSP, 1701-1704, 1986.

[15] H.W. Strube, "How to make an all-pass filter with a desired impulse response", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-30, No. 2, 1982.

[16] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am. 68(5), 1523-1525, 1980.

[17] T.F. Quatieri, J.T. Lynch, M.L. Malpass, R.J. McAulay and C.J. Weinstein, "The VISTA speech enhancement system for AM radio broadcasting", Final Technical Report, Lincoln Lab., MIT, 29, 1990.

[18] J.L. Goldstein, "Auditory spectral filtering and monaural phase perception," J. Acoust. Soc. Am. 41, 458-, 1967.

[19] A.J.M. Houtsma and J. Smurzynski, "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. 81(1), 305-310, 1990.

[20] F. Charpentier and E. Moulines, "Pitch-synchronous wave form processing techniques for text-to-speech synthesis using diphones," Proceedings EUROSPEECH-89, vol. 2, 13–19, 1989.

[21] J. N. Holmes, " The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No.3, 1973.

[22] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. 49, 1971.

[23] E. Zwicker, G. Flottorp and S.S. Stevens, " Critical band width in loudness summation", J. Acoust. Soc. Am. 29, 1957.

[24] H. Duifhuis, "Audibility of high harmonics in a periodic pulse", J. Acoust. Soc. Am. 48, 888-893, 1970.

[25] R. D. Patterson, "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am. 82, 1560-1586, 1987.

[26] B.C.J. Moore and B.R. Glasberg, "Masking patters for synthetic vowels in simultaneous and forward masking," J. Acoust. Soc. Am. 73, 906-917, 1983.

[27] M.R. Schroeder and H.W. Strube, "Flat-spectrum speech", J. Acoust. Soc. Am. 79, 1580-1583, 1986.

[28] H. Traunmüller,"Phase vowel", in The psychophysics of speech perception, ed. by M.E.H. Schouten, Dordrecht, 1987.

[29] M.R. Schroeder, "Models of Hearing", Proc. of The IEEE, Vol. 63, No. 9, 1975.

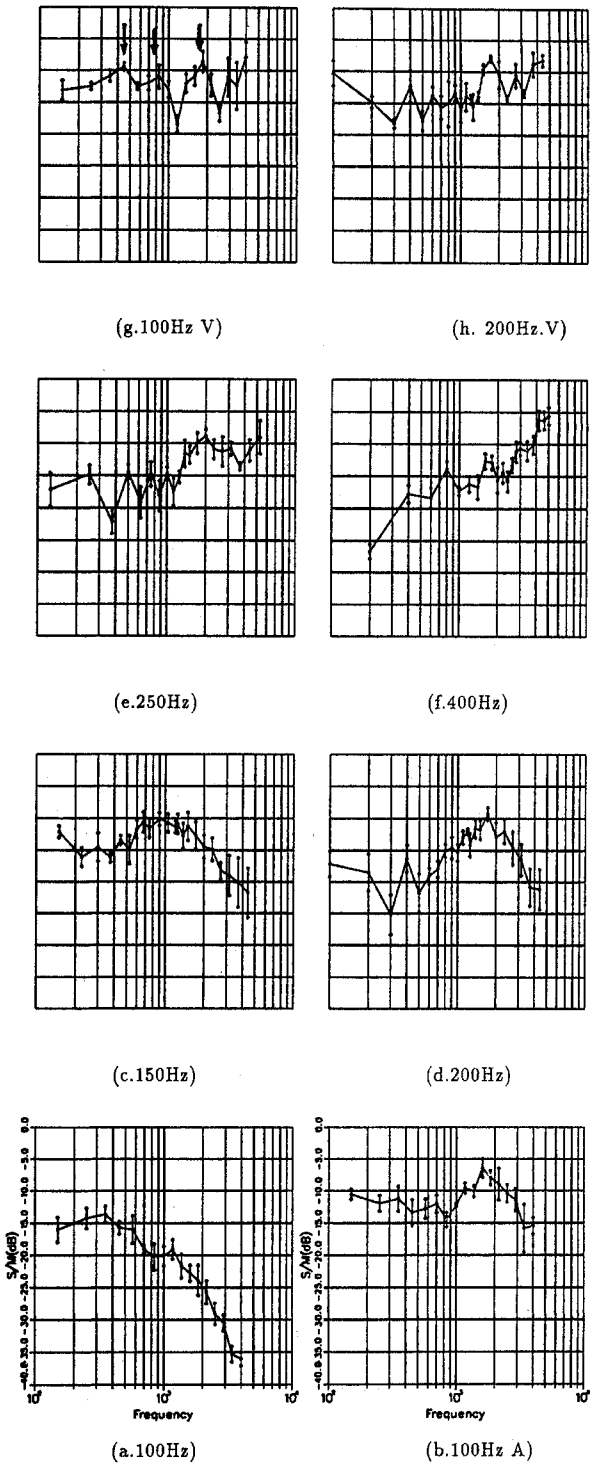

Figure 1: (a),100Hz zero-phase masker. (b),100Hz sin-cos phase masker. (c),150Hz zero-phase masker. (d),200Hz zero-phase masker. (e),250Hz zero-phase masker. (f),400Hz zero-phase masker. (g), vowel masker, f_0=100Hz. (h), voew masker, f_0=200Hz.