ON THE SIGNIFICANCE OF TEMPORAL MASKING IN SPEECH CODING

Jan Skoglund

Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
jans@s2.chalmers.se

W. Bastiaan Kleijn

Department of Speech, Music and Hearing
Royal Institute of Technology
SE-100 44 Stockholm, Sweden
bastiaan@speech.kth.se

ABSTRACT

This paper addresses the issue of masking of noise in voiced speech. First, we examine the audibility of cyclostationary narrow-band noise added to voiced speech generated by synthetic excitation. Varying the temporal location of noise within a pitch cycle corresponds to varying its phase spectrum. Using this fact, we find that a phase change of the noise in the high frequency region is more perceptible for a low-pitched sound than for a high-pitched sound. We propose a pitch-dependent temporal weighting function and we show experimentally that it is beneficial to the quantization of pitch-cycle waveforms.

1. INTRODUCTION

The perceived quality of coded speech results from preserving important dynamic features such as spectral envelope, pitch frequency, and waveform shape. By exploiting the masking properties of the human auditory system, we can reduce the audibility of quantization noise. In linear predictive speech coders, error weighting based on the (short-time) magnitude spectrum is employed to adapt the spectral envelope of the quantization noise [1]. More detailed information about masking, in both the phase and the magnitude spectral domain, will likely lead to improved performance of speech coders.

In source-filter based coding schemes, the excitation signal for voiced speech usually consists of pulses having almost flat power spectra. It has long been known that the phase spectrum of the excitation pulses affects speech quality, but how important this phase is for speech coding is not very well understood, although it is known that zero-phase impulses results in unnatural speech, and that a more accurate phase representation increases speech quality [2, 3, 4]. For low-rate coding, it is important to understand the significance of the phase spectrum. Thus, it is natural to study how much phase information of the pitch-cycle waveform is required for attaining high quality reconstructed speech.

The phase spectrum is closely related to the temporal distribution of energy in the pitch cycle. Hence, one way of investigating the perceptually important regions in the magnitude and phase spectrum is to add noise distributed in different time-frequency regions. A study of the audibility of stationary wide-band and narrow-band noise is presented in [5], which indicated that limited spectral resolution for low frequencies and limited temporal resolution for high frequencies strongly affect masking. The relative importance of these two effects is determined by the fundamental frequency of the impulse train.

In this work, we continue the study of the effects of temporal masking in near periodic signals, paying particular attention to their potential benefit for speech coding. Thus, we examine the audibility of the temporal distribution of cyclostationary noise within the pitch cycle for synthetic and natural speech signals and propose and test a simple temporal weighting function for quantization in speech coders.

2. EXPERIMENTAL SETUP

The signal processing was performed on signals with a sampling rate of 8 kHz. High-quality D/A conversion and headphones were used in the listening experiments. The background noise was measured to have an SPL of less than 50 dB and the signal levels were set to be approximately 80 dB. Three to six normal-hearing listeners were used for the experiments.

In our experiments, the speech signal was the *masking* or *masker* signal and the added noise was the *target* signal. We denote the target-to-masker-ratio, TMR_{f_c}, as the ratio between the power of the target and the masker in a critical band CB_{f_c}, where f_c is the center frequency. Hence

$$\mathit{TMR}_{f_c} = 10 \log_{10} \frac{\int_{\mathit{CB}_{f_c}} S_N(f)\,df}{\int_{\mathit{CB}_{f_c}} S_S(f)\,df}, \tag{1}$$

where $S_N(f)$ and $S_S(f)$ are the short-time power spectral densities of the noise target and the speech masker, respectively. The lower limit of TMR_{f_c} that can be detected in listening will be referred to as the *audibility threshold*, *TD*, at frequency f_c. The critical bandwidths, expressed as equivalent rectangular bandwidths (*ERB*), were calculated following [6]. The TMR_{f_c} was computed as the energy ratio of critical band pass filtered signal segments over the entire interval of each experiment.

We measured audibility thresholds using an adaptive two-interval, forced-choice procedure using a 3-down, 1-up decision rule method which estimates the 79.4% correct decision point of a psychometric function [7]. Each run consisted of 60 pairs and the final threshold estimate was based on an average of at least three runs for six listeners, except where mentioned.

3. NOISE PERCEPTION IN PERIODIC OR NEARLY PERIODIC SIGNALS

The results in [5] suggest that the temporal distribution of quantization noise within the pitch cycle has a higher perceptual importance for male than female speech. Our first experiment on cyclostationary noise in a synthetic vowel confirms this. The masker is a synthetic vowel and the target is cyclostationary noise, constructed by windowing stationary critical band wide noise using Kaiser windows with a fixed support of 10 ms for each pitch cycle. Four phase positions, φ, of the burst relative to the impulse were examined for two pitch values, $F_0 = 100$ Hz and $F_0 = 200$ Hz, of the impulse trains and two center frequencies, $f_c = 3200$ Hz ($ERB = 370$ Hz) and $f_c = 1600$ Hz ($ERB = 200$ Hz), of the noise. Figure 1 illustrates the impulse excitation added with cyclostationary noise in a specified phase position.

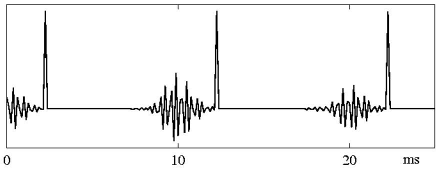

Figure 1: Composite of cyclostationary noise and impulse train. The phase position of the noise is $\varphi = 3\pi/2$.

The distorted synthetic vowel was generated by exciting an all-pole filter, having the transfer function depicted in Figure 2, with the distorted impulse train. In the figure, the spectral location of the two noise signals is also indicated. The duration of both masker and target was 500 ms, including 20-ms half-period sinusoidal rise and fall windows.

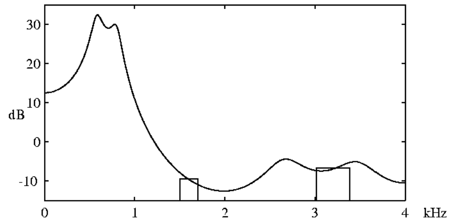

Figure 2: Vowel spectrum used in the experiments. The power density spectrum of the target noise ($f_c = 3200$ Hz and $f_c = 1600$ Hz) is shown at a $TMR_{f_c} = 0$ dB.

Figure 3 depicts the resulting *masking period patterns* (audibility thresholds as a function of cycle phase) for noise bursts centered at $f_c = 3200$ Hz. The average thresholds are connected with straight lines and the vertical bars show the standard deviations. Phase position $\varphi = 0$, corresponding to the pitch pulse being in the center of the noise burst, was perceived as the least sensitive phase position for all subjects and both pitch frequencies. Since the subjects vary in their absolute threshold levels, an interesting diagram can be obtained by normalizing the threshold for each subject at $\varphi = 0$ and plot the threshold difference for the other phase settings. This normalized diagram is depicted in Figure 4. For the 100 Hz vowel, there is a difference of around 20 dB between the most and least detectable phase positions. For the 200 Hz vowel, the difference is only around 3 dB. Hence, the difference in phase sensitivity between the high-pitched and the low-pitched vowel is 17-18 dB, confirming increasing phase sensitivity with decreasing pitch. Note that for the 100 Hz vowel the cyclostationary noise is most audible between the impulses. In a speech signal this means that the noise is masked most strongly around the location of the pitch pulse within the pitch cycle.

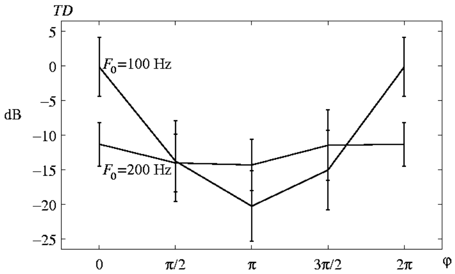

Figure 3: Average audibility thresholds and standard deviations (vertical bars) for different noise burst positions. Noise center frequency $f_c = 3200$ Hz.

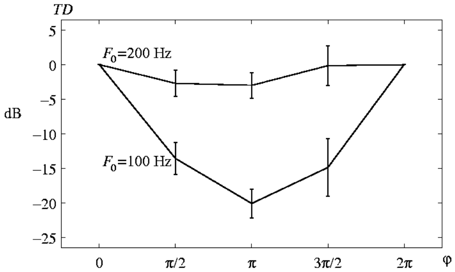

Figure 4: Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions, normalized for $\varphi = 0$. Noise center frequency $f_c = 3200$ Hz.

Thresholds for two additional pitch values, $F_0 = 133$ Hz and $F_0 = 160$ Hz, were measured for one of the subjects. All results for this subject are shown in Figure 5. The noise center frequency was $f_c = 3200$ Hz. The figure clearly illustrates the pitch dependency of the temporal sensitivity.

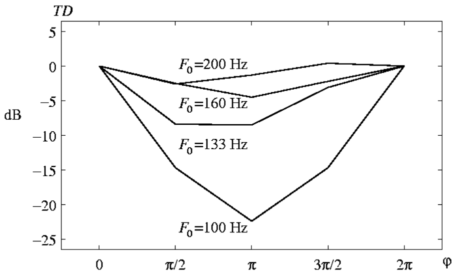

Figure 5: Audibility thresholds of one listener for different noise burst positions within a pitch cycle, normalized for $\varphi = 0$. Noise center frequency $f_c = 3200$ Hz.

When the center frequency of the noise bursts is decreased, the sensitivity to pitch decreases as well. In Figure 6, masking period patterns we obtained for noise bursts centered at $f_c = 1600$ Hz are shown.

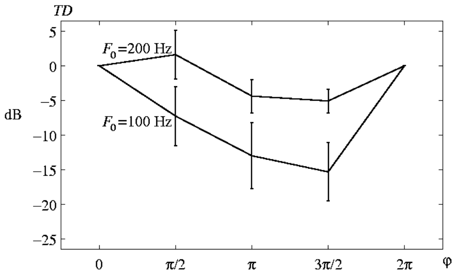

Figure 6: Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions, normalized for $\varphi = 0$. Noise center frequency $f_c = 1600$ Hz.

Our results are consistent with related results in the literature. Using a 3 ms long tone target at 3 kHz, Zwicker [8] measured a decrease of masking by 15 dB in the center of the silent half-period of square-wave modulated broad band noise with modulation frequency of 100 Hz. In a similar experiment, Fastl [9] obtained a 15 dB deep valley with a 5 ms 2 kHz tone and a modulation frequency of 67 Hz. This is consistent with our results, since the masking noise bursts in Fastl's experiment had longer duration than the vowel pulses in our experiment, yielding a higher amount of masking. An investigation closely related to ours was performed by Duifhuis [10]. He measured masking period patterns of a complex masker, consisting of a fundamental and a number of harmonics, and the target was bursts of a harmonic not present in the masker. For the corresponding frequencies he obtained masking period patterns similar to ours, with valleys 10-20 dB deep. Our results thus show that Duifhuis' figures for a coherent target, i.e. a harmonic tone, are also valid for an incoherent, i.e. uncorrelated noise, target signal.

4. QUANTIZATION WITH TEMPORAL WEIGHTING

The results of the previous experiments suggest that, in the quantization of a pitch-cycle waveform, low accuracy of the waveform matching of high frequencies can be tolerated around the peak of the pulse and high accuracy is needed in the valleys between the peaks. In the next experiment, we investigate the relevance of this effect to quantization by simulating high-frequency quantization noise with different temporal weightings and examining its audibility in natural speech.

In our experiments we simulate a simple subband coding scheme by extracting each pitch cycle $\mathbf{s} = [s(0)\, s(1) \ldots s(N-1)]^T$, of speech in a natural speech utterance, where N is the pitch period, and decomposing it into a low frequency part $\mathbf{s}_L$ and a high frequency part $\mathbf{s}_H$ using the DFT so that $\mathbf{s} = \mathbf{s}_L + \mathbf{s}_H$. The pulses were aligned so that the peak of the pulse was centered in the vector. A noise vector $\mathbf{y}_H$ was then added, $\tilde{\mathbf{s}}_H = \mathbf{s}_H + \mathbf{y}_H$, and the speech was reconstructed. The noise vector was selected from a random codebook of size M vectors so as to minimize a weighted squared distortion criterion

$$D = \sum_{n=0}^{N-1} \Big(\big(\tilde{s}_H(n) - s_H(n)\big) w(n) \Big)^2, \qquad (2)$$

where we selected the temporal weighting function $w(n)$ to be a damped and shifted von Hann window

$$w(n) = 1 - \rho(N) + \frac{\rho(N)}{2} \left(1 - \cos\Big(2\pi \frac{n - \frac{N-1}{2}}{N-1} \Big) \right) \qquad (3)$$

with a pitch-dependent damping factor $\rho(N) = 10^{-\frac{aN^2-bN}{20}}$, where $a = 3 \cdot 10^{-3}$ and $b = 5 \cdot 10^{-2}$. This corresponds to a maximum damping of 15 dB and 3 dB for a pitch frequency of 100 Hz and 200 Hz, respectively. The high frequency pulse contained frequencies in the 3000 Hz to 4000 Hz band. An example of a pulse extracted from an utterance spoken by a male speaker and the corresponding weighting function is depicted in Figure 7. To eliminate the effects of possible statistical peculiarities, a new codebook of M entries was generated for each pulse. The vectors in the codebooks were normalized so that the final signal-to-noise ratio for the distorted pulse was constant and equal to 0 dB.

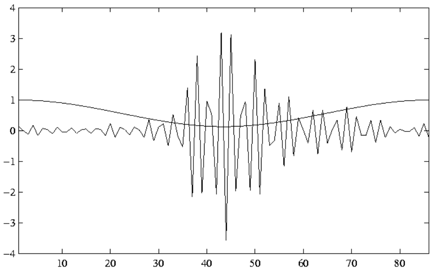

Figure 7: A pitch pulse, band-limited between 3 kHz and 4 kHz, of $N = 86$ samples and the corresponding weighting function.

We examined two ways of generating the noise. The first and most straightforward method was to generate a vector $\mathbf{x} = [x(0)\, x(1) \ldots x(N-1)]^T$ consisting of i.i.d. Gaussian components $x(n)$ and then decomposing it in the same manner as was done with the speech vector, thereby obtaining a noise vector $\mathbf{y}_H = \mathbf{x}_H = \mathbf{x} - \mathbf{x}_L$. Thus, the noise vectors are uncorrelated with the speech. For high rate quantizers this is often a feasible first-order approximation [11]. However, the quantization noise in speech coders is typically correlated with the speech. In a nearest-neighbor optimal quantizer the reconstruction vectors are the centroids of the coding regions [11]. For such a quantizer, the quantization error is correlated with the unquantized vector and its autocorrelation matrix is

$$\mathbf{C}_y = E[\mathbf{y}_H \mathbf{y}_H^T] = -E[\mathbf{s}_H \mathbf{y}_H^T]. \qquad (4)$$

To generate noise vectors having a correlation as in (4) we let the vectors be of the form

$$\mathbf{y}_H = \mathbf{A}x + \gamma \mathbf{s}_H \qquad (5)$$

where $\mathbf{x}$, as before, consists of i.i.d. Gaussian components and where γ determines the correlation level. Insertion of (5) in (4) yields the following relation

$$-\gamma(\gamma + 1)\mathbf{C}_s = \mathbf{A}\mathbf{A}^T. \qquad (6)$$

We see that $\mathbf{A}$ is a real-valued matrix when $-1 < \gamma < 0$. The pulse correlation matrix $\mathbf{C}_s$ was estimated using pulses from several speech files. Since pulses have different dimensions N, zero-padding was applied to obtain a normalized dimension $N_0 > N$.

For each experiment, we created a quasi-coded version of eight utterances (three female and five male). In one version, the

Correlation	M	S1	S2	S3	S4	Pref.
No corr.	32	8	10	9	8	87%
	64	8	10	7	8	82%
$\gamma = -0.3$	32	10	10	10	9	97%
	64	10	10	6	10	90%
$\gamma = -0.8$	32	7	10	9	6	80%
	64	8	10	7	7	80%

Table 1: Results of preference test for male utterances for subjects S1-S4. The numbers correspond to how many times the weighted criterion was preferred to the unweighted criterion. The average preference is given in percent.

Correlation	M	S1	S2	S3	S4	Pref.
No corr.	32	2	4	6	4	67%
	64	3	5	5	3	67%
$\gamma = -0.3$	32	6	4	6	5	87%
	64	6	6	3	6	79%
$\gamma = -0.8$	32	6	5	5	4	83%
	64	3	4	3	5	63%

Table 2: Results of preference test for female utterances.

codebooks were searched with an unweighted squared-error criterion and in the other version they were searched with our new, weighted squared-error criterion. The utterances were presented in random order to the listeners who had to indicate which utterance they preferred in a forced-choice pairwise comparison. We used three types of noise: uncorrelated noise and speech correlated noise with γ = -0.3 and -0.8. The results are presented in Table 1 and Table 2.

From the tables, we see a clear preference for the weighted criterion for the utterances spoken by male speakers while the preference is less strong for the female speakers. These results are consistent with our previous experiments. The preferences do not depend strongly on the amount of speech correlation of the simulated quantization noise.

Although we used simulated quantization noise, the results of Table 1 and 2 show that a temporal weighting in the quantizer can improve the speech quality for higher frequency bands for low-pitched speech. The proposed weighting criterion can be made more sophisticated by using different damping functions in different frequency bands.

5. CONCLUSIONS

The masking of noise in nearly periodic sounds such as voiced speech depends on the fundamental frequency of the sound [5]. For high-pitched sounds, the auditory system sensitivity to low-frequency noise is strongest in the valleys between the harmonics in the spectral domain. For low-pitched sounds, the sensitivity to high-frequency noise is strongest in the valleys between the pulse peaks in the time domain. Varying the temporal distribution of noise during a pitch cycle corresponds to a change in its phase spectrum. Although phase changes could be detected in a high pitched vowel, the effect of a phase change is significantly more audible in a low-pitched vowel. Our results, and those of [5], suggest strongly that phase changes are more audible for male than for female speakers.

In speech coding, this suggests that for female speakers it is important to maintain the harmonic structure of the (short-term) magnitude spectrum at low frequencies but that low accuracy suffices for the phase spectrum of the pitch cycle. For male speakers, more bits should be allocated to the phase spectrum of the pitch cycle, but a degradation in the harmonic structure is not audible. The results are consistent with the relative performance commonly found for CELP and sinusoidal coders. In CELP, many bits are essentially spent on the description of the phase of the pitch-cycle waveform, which means that male speakers sound relatively good. However, the reconstruction accuracy of the harmonic structure of the short-term magnitude spectrum is relatively low in CELP (the local peak-to-valley ratio is reduced significantly). This is a result of inadequate performance by the long-term predictor. In sinusoidal coders, on the other hand, the reconstruction of the harmonic character of the speech is generally very good, but the pitch-cycle phase is usually modeled with low accuracy. Thus, female voices sound better than male voices in sinusoidal coders. Our results indicate that exploitation of the pitch-dependent behavior of temporal masking should lead to significant improvement in speech coder performance.

6. REFERENCES

1. B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.
2. A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–590, 1971.
3. W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 1–10, January 1997.
4. X. Sun and B. Cheetham, "Speech excitation modelling for low bit speech coding," in *IEEE Speech Coding Workshop*, (Pocono Manor, PA), pp. 9–10, 1997.
5. C. Ma and D. O'Shaughnessy, "The masking of narrowband noise by broadband harmonic complex sounds and implications for the processing of speech sounds," *Speech Communication*, vol. 14, pp. 103–118, 1994.
6. B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.
7. H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 467–477, 1971.
8. E. Zwicker and H. Fastl, *Psychoacoustics.* Springer, Berlin, 1990.
9. H. Fastl, "Temporal masking effects: II. critical band noise masker," *Acustica*, vol. 36, no. 5, pp. 317–330, 1976/77.
10. H. Duifhuis, "Audibility of high harmonics in a periodic pulse II, time effect," *J. Acoust. Soc. Am.*, vol. 49, no. 4, pp. 1155–1162, 1971.
11. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Dordrecht, Holland: Kluwer Academic Publishers, 1991.