On Time-Frequency Masking in Voiced Speech

Jan Skoglund, *Member, IEEE,* and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—This paper addresses the issue of masking of noise in voiced speech. First, we examine the audibility of cyclostationary narrow-band noise bursts 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 found that a change of phase of the noise in the high frequency region is more perceptible for a low-pitched sound than for a high-pitched sound. We then propose a pitch-dependent temporal weighting function which can be employed in quantization of pitch cycle waveforms. In a second experiment, we found that the audibility of high-frequency noise added to natural speech can be significantly reduced using this weighting function.**

***Index Terms*—Auditory masking, phase spectrum, speech coding, temporal weighting.**

I. Introduction

THE perceived quality of coded speech results from the process of tracking and 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 magnitude spectrum is often 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 traditional psycho-physical masking experiments, stimuli with a simple temporal or spectral structure such as noise, clicks or pure tones are often used. The experiments have also focused on phenomena belonging to one of three temporal masking classes: simultaneous masking, pre-masking or post-masking. It may be difficult to extend results from such studies into more complex physical signals like speech and music. For example, the masking properties of voiced speech are due to a combination of all the three classes.

Let us consider the vowel segment displayed in Fig. 1. Spectrally, this segment has a harmonic structure while temporally it has a structure with a periodic envelope. (The envelope in the figure is the analytical envelope of the signal.) In the processing of a sound in the auditory system, the function of the inner ear can be viewed as a bank of parallel bandpass filters. The bandwidths, so-called critical bandwidths, of these auditory filters increase with frequency. In Fig. 2, the transfer characteristics of an auditory filter bank described in [2] (which is an implementation of filter types from [3]) is depicted. A common notion is to assume that the envelopes of the auditory filter outputs contain the relevant features and cues for detection and discrimination of sounds [4]–[9]. The envelope detector is often modeled as a nonlinear device, e.g., a half-wave rectifier [4] or a square-law [10], followed by a low-pass filter or an integrator.

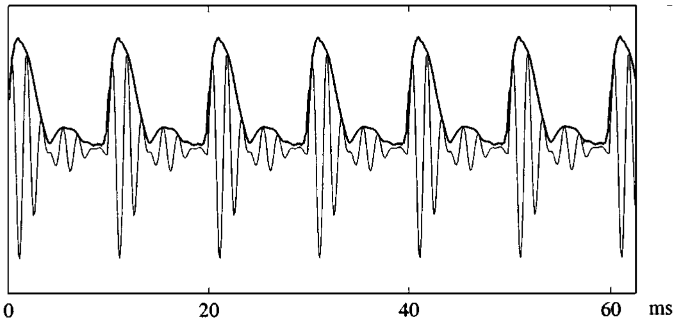

Fig. 1. Vowel (thin line) and its envelope (bold line).

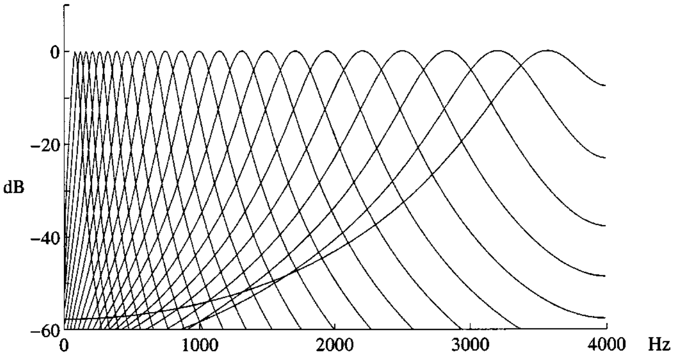

Fig. 2. Transfer characteristics of auditory filters.

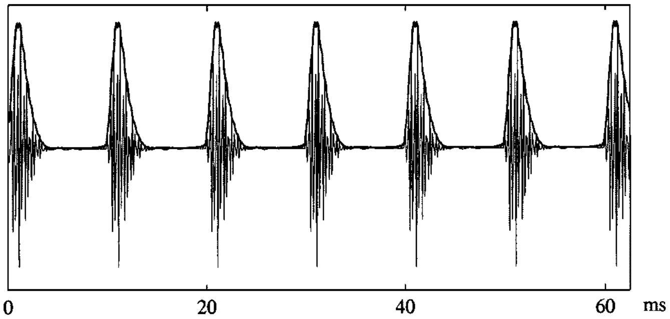

Fig. 3. Critical bandpass filtered vowel (thin line) and its envelope (bold line). The center frequency of the filter is 3200 Hz.

Fig. 3 depicts the output of a single high-frequency auditory channel (with a center frequency of 3200 Hz) in response to the

Manuscript received May 15, 1998; revised May 7, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Kroon.

J. Skoglund was with the Information Theory Group, Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Göteborg, Sweden. He is now with Global IP Sound, Inc., San Francisco, CA 94133 USA (e-mail: jan.skoglund@globalipsound.com).

W. B. Kleijn is with the Department of Speech, Music, and Hearing, Royal Institute of Technology, SE-100 44 Stockholm, Sweden.

Publisher Item Identifier S 1063-6676(00)05330-X.

vowel segment of Fig. 1. We see that the signal consists of short bursts with a distinct temporal structure. The figure illustrates the need for considering the time-frequency aspects of voiced speech in studying the masking properties. When decomposing the signal and considering single auditory channels, the information gained from the use of simple stimuli as described above may be sufficient to explain some of the masking effects observed in voiced speech.

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 result in unnatural speech, and that a more accurate phase representation increases speech quality [11]–[13]. For low-rate coding, it is important to understand the significance of the phase spectrum. Low-rate speech coders based on harmonic models often employ very crude descriptions of the phase spectrum and yet yield good quality [14] although they sometimes sound buzzy. 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 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 thorough study of the audibility of stationary wide-band and narrow-band noise is presented in [15] and [16]. Stationary noise was added to periodic impulse trains and the audibility thresholds were measured as a function of the fundamental frequency of the impulse trains. The results of these experiments indicated that in the low frequency region, the threshold for noise targets is mainly determined by the sharpness of spectral resolution. In the high frequency region, the critical bands are wide enough to contain several harmonics, thereby causing a temporally modulated waveform, so that the threshold for noise targets is detected by temporal analysis. The balance between these two effects is strongly dependent on the fundamental frequency of the impulse train.

To study the second effect further, we examined the audibility of stationary narrow-band noise in natural speech, and the audibility of the temporal distribution of cyclostationary[1] noise within the pitch cycle for synthetic vowels and natural speech. There are other factors that also contribute to the overall masking, e.g., the previously mentioned simultaneous masking. We have in this work concentrated on temporal masking phenomena. The results are relevant for the coding of voiced speech signals.

II. Experimental Setup

The signals in the experiments were produced on a Macintosh computer at an 8 kHz sampling frequency, upsampled to 32 kHz and low-pass filtered at 4 kHz prior to 16-bit analog reconstruction. The signals were amplified using a Sony FH-3 low-pass filter with a 4 kHz cut-off frequency and a NAD 3020B power amplifier. The experiments were performed using a pair of Beyerdynamic DT 990 headphones. The background noise was measured to have an SPL of less than 50 dB and the signal levels were around 80 dB. Three to six listeners with normal hearing were used for the experiments.

III. Time-Frequency Noise Experiments

In this section, we will investigate the audibility of critical band limited noise of different time-frequency regions when masked by natural and synthetic speech.

In the following, the speech signal will be denoted as the masking or masker signal and the added noise will be denoted as the target signal. Let the target-to-masker-ratio, TMR_{f_c}, denote the ratio between the power of the target and the masker in a critical band, CB_{f_c}, having a center frequency f_c. Hence

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

where $S_N(f)$ and $S_S(f)$ are the 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 [18] as

$$\mathrm{ERB} = 24.7(4.37 f_c + 1) \qquad (2)$$

where f_c is in kHz and ERB is in Hz. In the experiments, the TMR_{f_c} was computed as the energy ratio of critical band-pass filtered signal segments.

A. Preliminary Experiment

A pilot experiment was performed to investigate whether the isolated vowel results of [15] could be translated to an entire utterance with changing pitch and formant structure. The sentence "Joe brought a young girl" was spoken by one male and one female speaker. Narrow-band noise of critical bandwidth with varying center frequency was added to a tenth order linear prediction residual, having a sampling frequency of 8 kHz, and the speech was re-synthesized. The prediction order of ten was chosen as typical for predictive speech coders. To track the dynamic intensity of the speech, the noise was added at a constant segmental signal-to-noise ratio in the designated critical band. This means that the TMR_{f_c} was constant in each segment. The segment lengths were 20 ms. Noisy speech with a decreasing noise level was presented to four subjects who then indicated at what level the noise became inaudible. The noise level was decreased in 5-dB steps. The results are depicted in Fig. 4.

Although the results are affected by the time-varying formant structure of the utterance, some general effects of the time-frequency resolution can be observed. The sensitivity for the female speaker decreases as a function of frequency. After an initial decrease, the sensitivity for the male speaker increases. The

[1] Here we use the term "cyclostationary" of a random process if its statistical properties are invariant to a shift of the origin by integral multiples of the pitch period [17].

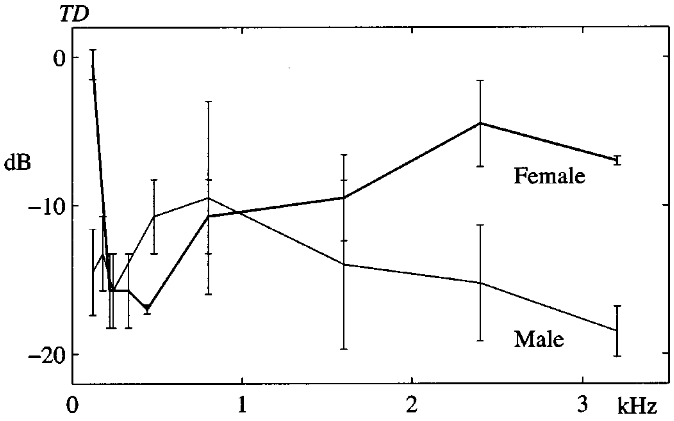

Fig. 4. Average audibility thresholds and standard deviations (vertical bars) for critical band limited noise of different center frequencies. A TD of 0 means equal target and masker energy in the critical band. Low values correspond to high sensitivity.

decreasing sensitivity of the female speaker with increasing frequency, and the initial decrease for the male speaker is consistent with the decreasing frequency resolution of the auditory filter bank with increasing frequency. The increasing sensitivity for the male speaker is consistent with increasing time-resolution with increasing frequency of the auditory filterbank.

As mentioned, the perceived quality of the female speech is more sensitive to additive noise at low frequencies. In these frequency regions, the spectral resolution of the auditory filterbank is sufficient to observe the noise between the harmonics. Naturally, this sensitivity will be apparent from our measurements only if our target-to-masker ratio measurement includes at least one harmonic. This is not always the case for females at low frequencies, explaining the anomalous point with 0 dB TMR$_{f_c}$ in Fig. 4.

B. Experiments

The results in Fig. 4 indicate that the audibility of quantization noise is pitch dependent. The higher audibility of stationary noise at high frequencies for male than for female speakers can be explained by the limited time resolution of the envelope detectors (temporal masking), the longer pitch cycle, and the nonuniform energy distribution within a pitch cycle. This, in turn, would mean that the temporal distribution of quantization noise within the pitch cycle has a higher perceptual importance for male than female speech. In the following experiment we will confirm this argument by examining the masking of cyclostationary noise by a synthetic vowel.

We use noise bursts of critical bandwidth, having a center frequency $f_c = 3200$ Hz (ERB = 370 Hz) and $f_c = 1600$ Hz (ERB = 200 Hz), as a basic building block. We noted in the previous experiment a higher temporal masking effect at higher frequencies and we limit the experiment to two bands in the region of interest. The masker is a synthetic vowel and the target consists of cyclostationary noise bursts generated by multiplying a stationary noise signal of critical bandwidth with a periodic window. The noise was added pitch-synchronously to an impulse train. Hence, in each pitch cycle of the impulse train, one noise burst is located. The windows consisted of concatenated Kaiser windows with a fixed support of 5 ms for each pitch cycle. For our work, we selected the procedure with fixed window support over the alternative where the window support is proportional to the pitch period. The latter method has a target bandwidth which varies with pitch. The power spectrum of the cyclostationary noise is slightly wider than the unwindowed stationary noise (an expression for the spectrum is given in the Appendix), but measurements show that the cyclostationary noise has approximately critical bandwidth. 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. Using the results given in the Appendix, the four different masker-plus-target signals have identical power spectra. Fig. 5 illustrates the impulse excitation added with cyclostationary noise in a specified phase position.

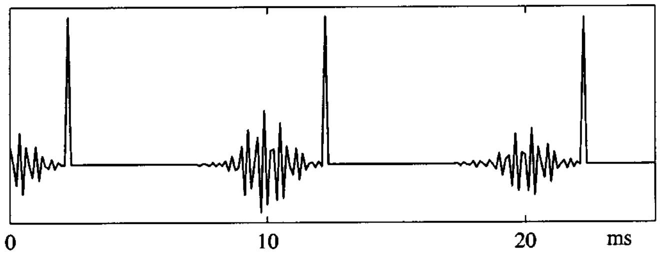

Fig. 5. Composite of cyclostationary noise and impulse train. The phase position of the noise is $\varphi = 3\pi/2$.

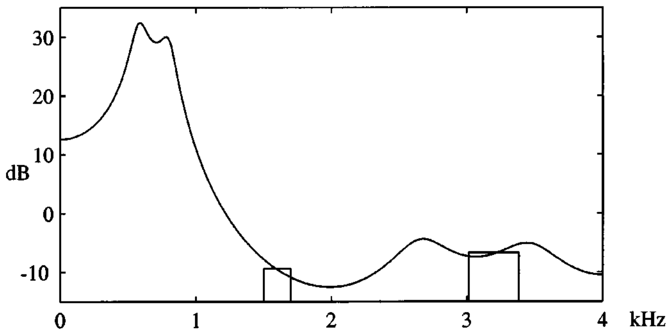

Fig. 6. Vowel spectrum used in the experiments. The power density spectrum of the target noise ($f_c = 3200$ Hz and $f_c = 1600$ Hz) is shown at a TMR$_{f_c}$ = 0 dB.

When the noise signal alone was presented to three subjects, they reported that the modulation effect is clearly audible at low pitch frequencies. However, at pitch frequencies higher than 170–180 Hz, the pitch-modulated noise is perceived as an almost stationary signal. This result is consistent with previous temporal masking experiments where the audibility of the modulation effect in amplitude-modulated wide band [4] and narrow band [19] noise was found to decrease with increasing modulation frequencies. This is another indication that for higher pitch frequencies the time location of the noise burst within a pitch cycle in a speech signal is of less importance for perception.

A synthetic vowel (c.f. Fig. 1) was then generated by exciting an all-pole filter, with a transfer function depicted in Fig. 6, with the distorted impulse train. The masking signal, i.e. the clean synthetic vowel, had an SPL of 84 dB. In the figure, the power density spectra of the two noise signals at a $\mathrm{TMR}_{f_c} = 0$ dB are also presented. The filter coefficients are $(a_0 = 1)$

$$\begin{aligned} a_1 &= -1.5672 \quad a_2 = 0.1299 \quad a_3 = 0.7840 \\ a_4 &= 0.5414 \quad a_5 = -0.7231 \quad a_6 = -0.3695 \\ a_7 &= 0.2431 \quad a_8 = 0.4628 \quad a_9 = -0.2915 \\ a_{10} &= 0.0253. \end{aligned}$$

A vowel generated by a zero-phase impulse train sounds unnatural, as previously mentioned. This is not a problem for the current masking experiment, since the subjects only indicated if noise was present or not and no judgements were made about the naturalness of the vowel. In a study by Hermes [20], similar types of noise bursts were added to synthetic vowels and it was found that adding noise increases naturalness.

The signals were generated using Matlab and the duration of both the masker and the target was 500 ms, including 20-ms half-period sinusoidal rise and fall windows. The TMR_{f_c} was calculated using the 500 ms signals. Audibility thresholds were measured using an adaptive two-interval, forced-choice procedure using a three-down, one-up decision rule that estimates the 79.4% correct decision point of a psychometric function [21]. If three correct answers were given, the noise level was decreased. For an incorrect answer, the noise level was increased. Each run consisted of 60 pairs. The noise level was initially well above the threshold. An initial step size was set to 4 dB and the downward step size was reduced to 2 dB after the first two reversals. The threshold estimate was based on the average of reversal points. If the total number of reversals was even, the first two reversals were discarded. Otherwise, the first three reversals were discarded. If the standard deviation of reversal points for a run was greater than 5 dB, that run was discarded. The final estimate of the threshold was based on an average of at least three runs. In this experiment, six listeners participated. The subjects received practice runs before the experiments and the listening sessions were segmented into 20-min sessions with intermediate pauses.

C. Results

If the results are presented as a function of the different time location within the period of the masker we get what is referred to in the literature as a masking period pattern [22]. Note that traditionally the target has been tone bursts [7], [22], [23] and not noise bursts as in this study. Fig. 7 depicts masking period patterns 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 postion $\varphi = 0$, 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 (in TMR_{f_c}), 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 Fig. 8.

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 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 pitch pulse excitation. A natural excitation pulse is often less peaky than an impulse, therefore we might expect a slightly lower difference for a natural vowel.

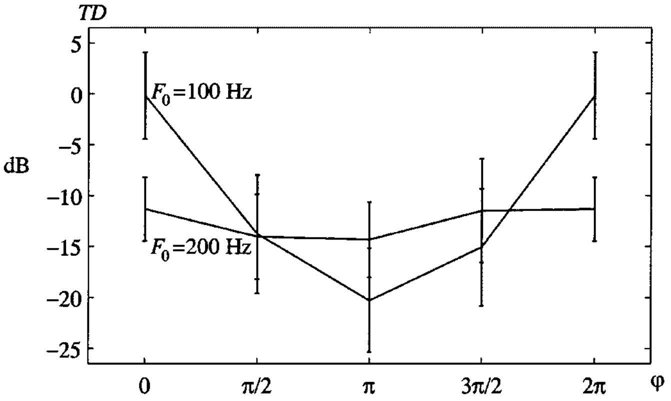

Fig. 7. Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions within a pitch cycle. Noise center frequency $f_c = 3200$ Hz.

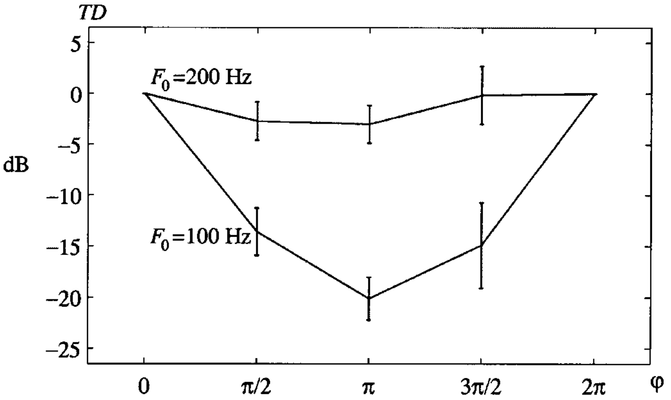

Fig. 8. Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions within a pitch cycle, normalized for $\varphi = 0$. Noise center frequency $f_c = 3200$ Hz.

Thresholds for two additional pitch values, $F_0 = 133$ Hz and $F_0 = 160$ Hz, were measured for one of the subjects. The results are presented together with the thresholds for $F_0 = 100$ Hz and $F_0 = 200$ Hz for that subject in Fig. 9. The figure clearly illustrates the pitch dependency of the temporal sensitivity.

In Figs. 10 and 11, masking period patterns for noise bursts centered at $f_c = 1600$ Hz are presented. We see that there still is a difference between the 100-Hz and 200-Hz masker, although the difference is smaller. This is in agreement with the expected increasing pitch dependency of temporal masking with increasing center frequency.

D. Discussion

Masking period patterns of broad-band maskers have previously been presented for noise maskers and tone bursts targets. Using a 3 ms long tone target at 3 kHz Zwicker [22] measured a decrease of masking by 15 dB in the center of the silent half-period of square-wave modulated broad band

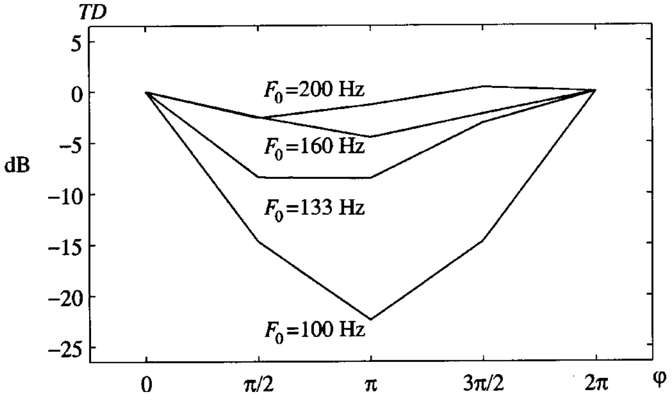

Fig. 9. Audibility thresholds of one listener for different noise burst positions within a pitch cycle, normalized for $\varphi = 0$. Noise center frequency $f_c = 3200$ Hz.

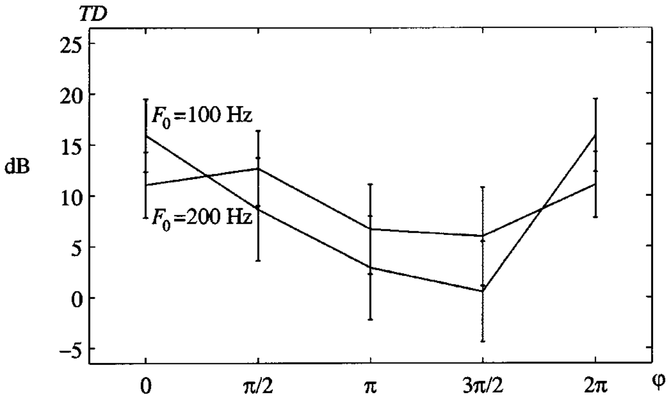

Fig. 10. Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions within a pitch cycle. Noise center frequency $f_c = 1600$ Hz.

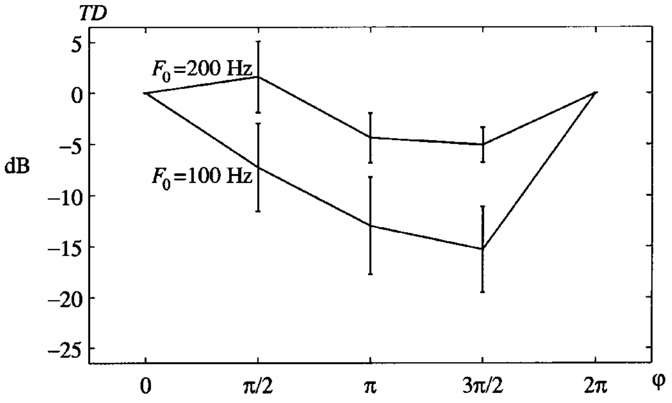

Fig. 11. Average audibility thresholds and the standard deviations (vertical bars) for different noise burst positions within a pitch cycle, normalized for $\varphi = 0$. Noise center frequency $f_c = 1600$ Hz.

noise with modulation frequency of 100 Hz. At 100 Hz, the masking-period pattern was quite symmetric and at lower modulation frequencies the pattern was more asymmetric due to the asymmetry between pre-masking and post-masking. In a similar experiment, Fastl [23] obtained a 15 dB deep valley with a 5 ms 2 kHz tone target 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, which means that the associated integration process yields a higher amount of masking. An investigation closely related to ours was performed by Duifhuis [24]. 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. The results in this work shows that Duifhuis' figures for coherent distortion, i.e., a harmonic tone, also are valid for an incoherent, i.e. uncorrelated noise, target signal. This is notable and quite unexpected, because the masking behavior of noise has in many other studies differed from the masking behavior of tones, e.g. [25].

One important feature often employed to explain masking effects is the temporal envelope of the waveform at the output of the auditory filters. There are a number of aspects of the envelope that have been proposed as cues for detection and discrimination, e.g., power [4], max-min ratio [6] and power spectrum [5], [9]. Although the phase characteristics have some impact [7], the time resolution of the auditory system is determined by two factors in the previously described model, the bandwidths of the auditory filters and the bandwidth (time-constant) of the integrator in the subsequent envelope detector. Moore *et al.* have suggested an envelope detector consisting of an integration of the instantaneous squared amplitude multiplied with a sliding window [10]. We have plotted the output of such an envelope detector (in dB scale) for the signals in the experiments in Figs. 12 and 13. It is clearly seen in these figures that the temporal resolution of the envelope detectors for the 200-Hz signal is not sufficient to distinguish between the different noise settings.

IV. Quantization Noise Experiments

The results of the previous experiments and the results of [15] and [16] suggest that, in 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 will investigate this further by simulating high-frequency quantization noise and examining its audibility in natural speech.

A. Experiments

In these experiments we simulate a simple subband coding scheme by extracting each pitch cycle vector $\mathbf{s} = [s(0), s(1), \cdots, s(N-1)]^T$, 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 discrete Fourier transform 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

$$\tilde{\mathbf{s}}_H = \mathbf{s}_H + \mathbf{y}_H \tag{3}$$

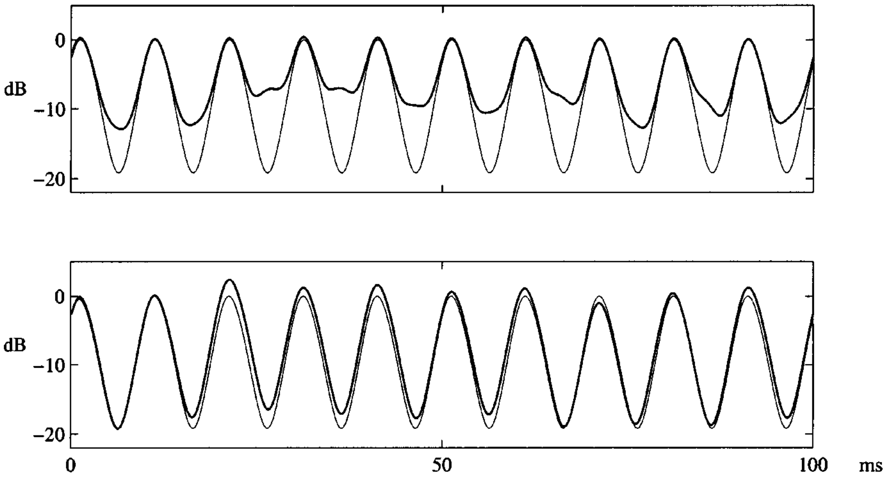

Fig. 12. Envelopes of outputs of auditory filter at $f_c = 3200$ Hz. 100-Hz vowel (thin line) and 100-Hz vowel + noise (bold line), TMR$_{f_c} = -12$ dB. Top: $\varphi = \pi$. Bottom: $\varphi = 0$.

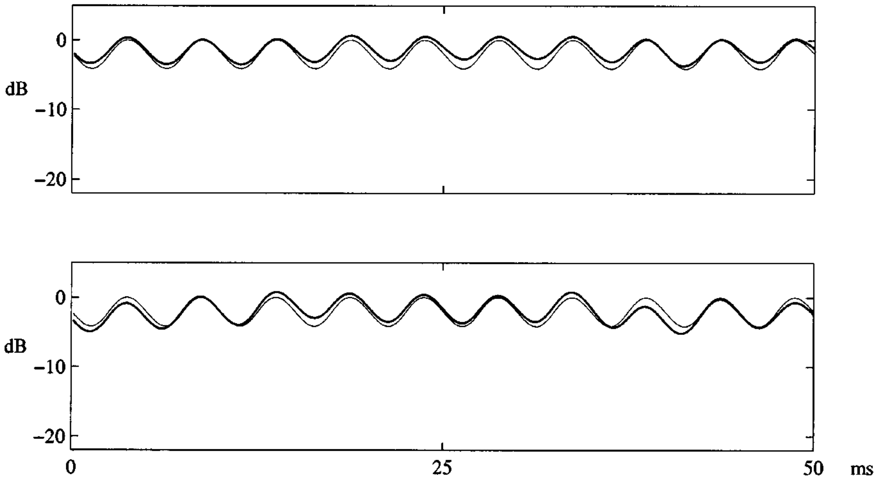

Fig. 13. Envelopes of outputs of auditory filter at $f_c = 3200$ Hz. 200-Hz vowel (thin line) and 200-Hz vowel + noise (bold line), TMR$_{f_c} = -12$ dB. Top: $\varphi = \pi$. Bottom: $\varphi = 0$.

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

$$D = \sum_{n=0}^{N-1} \left((\tilde{s}_H(n) - s_H(n))w(n)\right)^2 \tag{4}$$

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

$$w(n) = 1 - \rho(N) + \frac{\rho(N)}{2}\left(1 - \cos\left(2\pi \frac{n - \dfrac{N-1}{2}}{N-1}\right)\right) \tag{5}$$

with a pitch-dependent attenuation factor $\rho(N) = 10^{(aN^2 - bN/20)}$, where $a = 3 \cdot 10^{-3}$ and $b = 5 \cdot 10^{-2}$. This corresponds to a maximum attenuation of 15 dB and 3 dB for a pitch frequency of 100 Hz and 200 Hz, respectively. The high-frequency pulse, s_H, 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 Fig. 14.

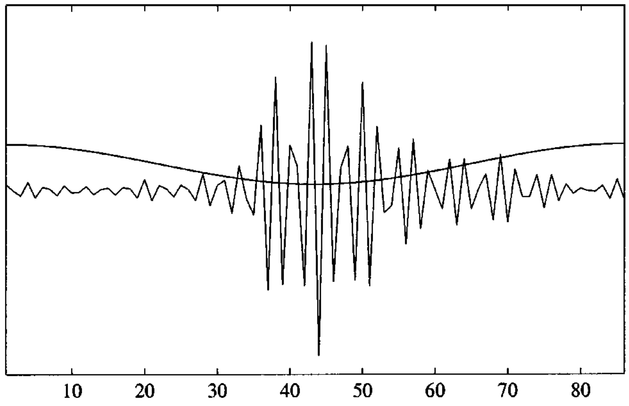

Fig. 14. A pitch pulse of $N = 86$ samples and the corresponding weighting function. The pulse is bandlimited from 3000 Hz to 4000 Hz.

To eliminate the effects of possible statistical peculiarities of a given codebook, a new codebook of M entries was generated for each pulse. The vectors in the codebook were normalized so that the final signal-to-noise ratio for the distorted pulse was constant and equal to 0 dB. We examined two ways of generating the noise. The first and most straightforward method

TABLE I

RESULTS OF PREFERENCE TEST FOR SUBJECTS S1–S4. NOISE HAVING DIFFERENT DEGREE OF SPEECH CORRELATION WAS SELECTED FROM CODEBOOKS OF SIZE M. THE NUMBERS CORRESPOND TO HOW MANY TIMES THE WEIGHTED CRITERION WAS PREFERRED TO THE UNWEIGHTED CRITERION. THE AVERAGE PREFERENCE IS GIVEN IN PERCENT

		Male speech (5 utterances)					Female speech (3 utterances)				
	M	S1	S2	S3	S4	Pref.	S1	S2	S3	S4	Pref.
No corr.	32	8	10	9	8	87%	2	4	6	4	67%
	64	8	10	7	8	82%	3	5	5	3	67%
Corr. $\gamma = -0.3$	32	10	10	10	9	97%	6	4	6	5	87%
	64	10	10	6	10	90%	6	6	3	6	79%
Corr. $\gamma = -0.8$	32	7	10	9	6	80%	6	5	5	4	83%
	64	8	10	7	7	80%	3	4	3	5	63%

was to generate a vector $\mathbf{x} = [x(0), x(1), \cdots, x(N-1)]^T$ consisting of independent and identically distributed (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 [26]. However, the quantization noise in speech coders is typically correlated with the speech. A coder accounting for such correlation was presented in [27]. In an optimal quantizer, the reconstruction vectors are the centroids of the coding regions [26]. For such a quantizer, the quantization error is correlated with the unquantized vector and its autocorrelation matrix is

$$\begin{aligned}\mathbf{C}_\mathbf{y} &= E[(\tilde{\mathbf{s}}_H - \mathbf{s}_H)(\tilde{\mathbf{s}}_H - \mathbf{s}_H)^T] = E[\mathbf{y}_H \mathbf{y}_H^T] \\ &= -E[\mathbf{s}_H \mathbf{y}_H^T] \end{aligned} \tag{6}$$

where the last equality is obtained from the fact that the reconstruction vectors are centroids. To generate noise vectors having the correlation according to (6) we let the vectors be of the form

$$\mathbf{y}_H = \mathbf{A}\mathbf{x} + \gamma \mathbf{s}_H \tag{7}$$

where $\mathbf{x}$, as before, consists of i.i.d. Gaussian components. Insertion of (7) in (6) yields the following relation:

$$-\gamma(\gamma+1)E[\mathbf{s}_H \mathbf{s}_H^T] = -\gamma(\gamma+1)\mathbf{C}_\mathbf{s} = \mathbf{A}\mathbf{A}^T. \tag{8}$$

We see that $\mathbf{A}$ is a real-valued matrix when $-1 < \gamma < 0$. In the experiments, the matrix $\mathbf{A}$ was chosen as one solution to (8) for different values of γ. The pulse correlation matrix $\mathbf{C}_\mathbf{s}$ 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 quasicoded version of eight utterances (three female and five male). In one version, the 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 $\gamma = -0.3$ and -0.8. The results are presented in Table I.

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.

B. Discussion

Although this is a simple simulation of a speech coding scheme, the results of Table I confirm that a temporal weighting 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 attenuation functions in different frequency bands.

V. CONCLUSIONS

The masking of noise in nearly periodic sounds such as voiced speech depends on the fundamental frequency of the sound [15] as well as many other factors. 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 [15], strongly suggest 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 temporal behavior of masking should lead to significant improvement in speech coder performance.

APPENDIX

In this Appendix we study the statistical relations between the pitch-synchronously modulated noise and the impulse excitation components of the synthetic vowels described in the article.

Let $W(t)$ denote a stationary zero-mean noise process having an autocorrelation function $R_W(\tau)$ and the corresponding power density spectrum $S_W(f)$. Let $v(t)$ denote a window with limited time support, $v(t) = 0, t \in \mathcal{R} - [-T/2, T/2]$, and let $u(t)$ denote a periodic repetition of $v(t)$

$$u(t) = \sum_{k=-\infty}^{\infty} v(t - kT) \tag{9}$$

with period T. By introducing a random phase position, θ, we form a stationary process

$$C(t) = u(t - \theta) = \sum_{k=-\infty}^{\infty} v(t - kT - \theta) \tag{10}$$

where θ is a uniform random variable on the interval $[-T/2, T/2]$. If we assume that $W(t)$ and $C(t)$ are independent we have that the windowed noise process $X(t) = C(t)W(t)$ is stationary with a mean $E[X(t)] = E[C(t)]E[W(t)] = 0$, and an autocorrelation function $R_X(\tau) = E[X(t)X(t + \tau)] = R_C(\tau)R_W(\tau)$. It is straight-forward to derive an expression for the power density spectrum of the windowed noise as

$$\begin{aligned} S_X(f) &= \int_{-\infty}^{\infty} S_C(\nu) S_W(f - \nu)\, d\nu \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} S'_C\left(\frac{k}{T}\right) S_W\left(f - \frac{k}{T}\right) \end{aligned} \tag{11}$$

where $S_W(f)$ is the power density spectrum of the stationary noise $W(t)$ and $S'_C(f)$ is the Fourier transform of one period of the autocorrelation function $R_C(\tau)$

$$R'_C(\tau) = \begin{cases} R_C(\tau) & \tau \in [-T/2, T/2] \\ 0 & \tau \notin [-T/2, T/2]. \end{cases} \tag{12}$$

As long as $S'_C(f)$ has a small bandwidth compared with $S_W(f)$, the spectral splatter caused by the windowing may be neglected.

Now consider the addition of $X(t)$ and a stationary train of impulses $D(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT - \varphi)$, where the signals have a specified phase relation, ψ

$$Y(t) = D(t) + X(t + \psi). \tag{13}$$

Since $W(t)$ is independent of the other signals the correlation between $D(t)$ and $X(t + \psi)$ is

$$E[D(t)X(t+\psi)] = E[D(t)C(t+\psi)]E[W(t+\psi)] = 0. \tag{14}$$

The phase relation between $D(t)$ and $X(t + \psi)$ will thereby only affect $E[D(t)C(t + \psi)]$. Hence, the autocorrelation function becomes $R_Y(\tau) = R_D(\tau) + R_X(\tau)$, where the cross terms cancel according to (14). The power spectrum of $Y(t)$ is consequently $S_Y(f) = S_D(f) + S_X(f)$.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions for improving our paper. They would also like to express their gratitude to the participants of the listening tests for their time and effort.

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. ASSP-27, pp. 247–254, Mar. 1979.

[2] M. Slaney, "Auditory toolbox: A MATLAB toolbox for auditory modeling work," Apple Computer, Tech. Rep. 45, 1993.

[3] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Amer.*, vol. 82, pp. 1560–1586, 1987.

[4] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *J. Acoust. Soc. Amer.*, vol. 66, no. 5, pp. 1364–1380, 1979.

[5] J.-P. Martens, "A new theory for multitone masking," *J. Acoust. Soc. Amer.*, vol. 72, no. 2, pp. 397–405, 1982.

[6] T. G. Forrest and D. M. Green, "Detection of partially filled gaps in noise and the temporal modulation transfer function," *J. Acoust. Soc. Amer.*, vol. 82, no. 2, pp. 1933–1943, 1987.

[7] A. Kohlrausch and A. Sander, "Phase effects in masking related to dispersion in the inner ear. II. Masking period patterns of short targets," *J. Acoust Soc. Amer.*, vol. 97, no. 3, pp. 1817–1829, 1995.

[8] E. A. Strickland and N. F. Viemeister, "Cues for discrimination of envelopes," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3638–3646, 1996.

[9] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narro-band carriers," *J. Acoust. Soc. Amer.*, vol. 97, no. 3, pp. 1817–1829, 1995.

[10] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *J. Acoust. Soc. Amer.*, vol. 83, no. 3, pp. 1102–1116, 1988.

[11] 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.

[12] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 1–10, Jan. 1997.

[13] X. Sun and B. M. G. Cheetham, "Speech excitation modeling for low bit speech coding," in *Proc. IEEE Workshop Speech Coding Telecommunications*, Pocono Manor, PA, 1997, pp. 9–10.

[14] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 121–173.

[15] C. Ma, "Psychophysical and Signal-Processing Aspects of Speech Representation," Ph.D. dissertation, Eindhoven Univ. Technol., Eindhoven, The Netherlands, 1992.

[16] 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 Commun.*, vol. 14, pp. 103–118, 1994.

[17] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[18] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[19] D. A. Eddins, "Amplitude modulation detection of narrow-band noise: Effects of absolute bandwidth and frequency region," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 470–479, 1993.

[20] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Commun.*, vol. 10, pp. 497–502, 1991.

[21] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Amer.*, vol. 49, pp. 467–477, 1971.

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[23] H. Fastl, "Temporal masking effects: I. Broad band noise masker," *Acoustica*, vol. 353, no. 5, pp. 287–302, 1976.

[24] H. Duifhuis, "Audibility of high harmonics in a periodic pulse II. Time effect," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1155–1162, 1971.

[25] R. Hellman, "Asymmetry of masking between noise and tone," *Percept. Psychophys.*, vol. 11, no. 3, pp. 241–246, 1972.

[26] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Dordrecht, The Netherlands: Kluwer, 1991.

[27] S. V. Andersen, S. H. Jensen, and E. Hansen, "Quantization noise modeling in low-delay speech coding," in *Proc. IEEE Workshop Speech Coding Telecommunications*, Pocono Manor, PA, pp. 65–66.

Jan Skoglund (S'93–M'98) was born in Göteborg, Sweden, in 1967. He received the M.S. degree in electrical engineering and the Lic.Eng. and Ph.D. degrees in information theory from Chalmers University of Technology, Göteborg, in 1992, 1996, and 1998, respectively. His Ph.D. thesis addressed different aspects of speech coding, such as spectrum quantization, pulse excitation modeling, and perceptual coding.

From 1992 to 1998, he was with the Department of Information Theory, Chalmers University of Technology. He worked on low bit rate speech coding as a Consultant at the Speech and Image Processing Service Research Lab, AT&T Labs-Research, Florham Park, NJ, from 1999 to 2000. Since 2000, he has been with Global IP Sound, Inc., San Francisco, CA, where he is working on speech processing tailored for packet-switched networks.

Bastiaan Kleijn (M'88–F'99) completed his undergraduate studies in The Netherlands in 1977 and received the M.S. degree in physics and the Ph.D. degree in soil science from the University of California, Riverside, in 1981. He received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1984 and the Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 1991. He held a postdoctoral position in physics at the University of California, Riverside, during 1981–1982, and was a Research Fellow in Physical Chemistry at the University of Melbourne, Melbourne, Australia, during 1982–1983.

He was a Member of Technical Staff at AT&T Bell Laboratories from 1984 to 1996. Since 1996, he has been with the Department of Speech, Music, and Hearing, Royal Institute of Technology, Stockholm, Sweden, where he is now Professor of speech processing. He held visiting professorships at Delft University of Technology and Vienna University of Technology in 1996 and 1998, respectively.

Dr. Kleijn was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1992 to 1997, a member of the IEEE Signal Processing Society Technical Committee on Neural Networks for Signal Processing from 1991 to 1994, and technical co-chair of ICASSP'99. He is currently a member of the IEEE Signal Processing Society Technical Committee on Speech Processing.