On the significance of phase in the short term Fourier spectrum for speech intelligibility

Michiko Kazama, Satoru Gotoh, and Mikio Tohyama
Waseda University, 161 Nishi-waseda, Shinjuku-ku, Tokyo 169–8050, Japan

Tammo Houtgast
VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands

(Received 18 November 2008; revised 15 October 2009; accepted 29 December 2009)

This paper investigates the significance of the magnitude or the phase in the short term Fourier spectrum for speech intelligibility as a function of the time-window length. For a wide range of window lengths (1/16–2048 ms), two hybrid signals were obtained by a cross-wise combination of the magnitude and phase spectra of speech and white noise. Speech intelligibility data showed the significance of the phase spectrum for longer windows (>256 ms) and for very short windows (<4 ms), and that of the magnitude spectrum for medium-range window lengths. The hybrid signals used in the intelligibility test were analyzed in terms of the preservation of the original narrow-band speech envelopes. Correlations between the narrow-band envelopes of the original speech and the hybrid signals show a similar pattern as a function of window length. This result illustrates the importance of the preservation of narrow-band envelopes for speech intelligibility. The observed significance of the phase spectrum in recovering the narrow-band envelopes for the long term windows and for the very short term windows is discussed.
© *2010 Acoustical Society of America.* [DOI: 10.1121/1.3294554]

PACS number(s): 43.66.Nm, 43.72.Ja [MW] Pages: 1432–1439

I. INTRODUCTION

The research question of this paper is the significance of the magnitude and the phase in the short term Fourier spectrum for speech intelligibility. The magnitude spectrum is considered important in almost all types of applications of speech processing, while phase has received less attention. Speech signals are commonly analyzed using short-time Fourier transforms (STFTs), and their characteristics are conventionally represented by the magnitude spectrum for speech analysis and/or synthesis.[1] The difference between phonemes is reflected in the structure of the magnitude spectra, and power (or magnitude) spectrum subtraction is commonly used for noise reduction.[2,3]

From a point of view of audio engineering type of applications, it is well known that the phase spectrum of sound is important for rendering (or removing) room-reverberation effects from a reverberation-free (or reverberant) signal. It seems that human listeners are able to appropriately detect phase changes in longer signal segments than those commonly used for speech analysis.[4–8] Schroeder and Strube[5] and Traumueller and Schouten[9] reported that vowels could be synthesized using phase information, and Oppenheim and Lim[10] found that when a speech signal was of sufficient length, speech intelligibility was lost in Fourier-transform magnitude-only reconstruction but not in phase-only reconstruction. However, it is still unclear how effectively the phase information is used for the synthesis of intelligible speech, and formal listening tests were not performed.

On the other hand, Liu *et al.*[11] intensively investigated the effect of the phase on intervocalic stop consonant perception for VCV speech signals. It was shown that the perception of intervocalic stop consonants varies from magnitude dominance to phase dominance as the Fourier-analysis window size increases across the cross-over point between 192 and 256 ms. An effect of phase on perception was also observed for shorter time windows, in the range of 10 to 30 ms. The present study builds on and expands the Liu *et al.* study with respect to two main issues: (1) it investigates the significance of magnitude versus phase in the short term Fourier spectrum for *sentence intelligibility,* rather than for the perception of intervocalic consonants, and (2) it covers a wider range of time-window lengths, with a window covering the complete sentence as the upper limit, down to the lower limit of a single-sample window length.

It is generally believed that speech intelligibility is related to narrow-band envelopes.[12] Drullman[13] found that intelligible speech signals can be synthesized by modulating 24 1/4-octave noise bands covering 100–6400 Hz range, using the temporal speech envelopes obtained in the corresponding 1/4-octave bands. This was the motivation for investigating the role of the phase spectrum in relation to the preservation of narrow-band envelopes. For instance, it will be shown that in case where the window length used in the Fourier transform is substantially larger than the period of the envelope modulation of interest, it is the phase spectrum that carries the information about the temporal envelope, not the magnitude spectrum. It will be shown that the same applies in the case of very short window lengths.

The experimental approach adopted in this paper is similar to the one used by Liu *et al.*,[11] but applied to a spoken sentence and random noise. From these signals, two new signals were created by a cross-wise combination of the magnitude and phase spectra of the speech and noise signals. These two hybrid signals are made for a wide range of win-

 0001-4966/2010/127(3)/1432/8/$25.00 © 2010 Acoustical Society of America

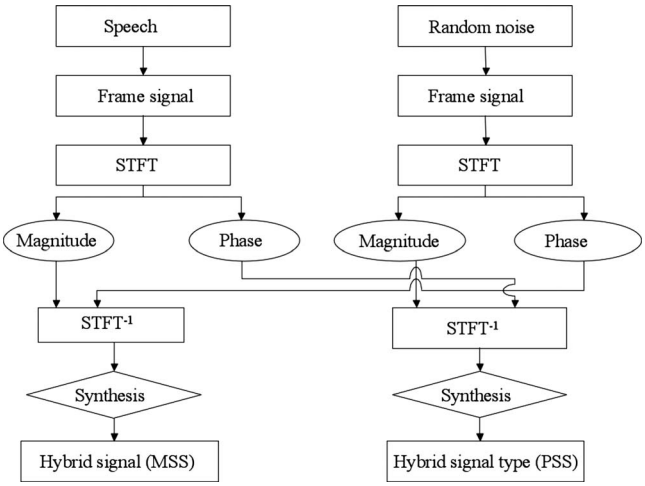

FIG. 1. Method for deriving two types of hybrid signals from speech and random noise using a cross-wise combination of the amplitude and phase spectra in the STFT overlap-add procedure.

dow lengths used in the STFT overlap-add procedure. Sentence intelligibility tests and envelope-correlation studies are performed to investigate the characteristics of the two hybrid signals as a function of window length.

II. LISTENING EXPERIMENT

Synthesized hybrid (magnitude- or phase-only) speech signals were obtained by using female-spoken speech and random-noise samples, as shown in Fig. 1. Sentence intelligibility for the two hybrid signals, as a function of the window length used in the STFT analysis and reconstruction, was estimated using listening tests.

A. Method

1. Test materials and signal processing

The original speech signals consisted of 96 sentences spoken by two female speakers. All of the speech materials were in Japanese and digitized at a sampling rate of 16 kHz, which seems well suited to the sentence intelligibility tests carried out in this study, using a 16-bit A/D converter. Each 1.5-s-long speech phrase had additional silent parts at the start and end, so the total length was 4 s. The speech tokens were simple everyday sentences, with a length of typically six to ten words, e.g., (translated), "This letter cannot be clearly seen from far away." A white-noise signal was produced using MATLAB software.

The speech and random-noise pairs were analyzed using STFT (Fig. 1) where a rectangular-window function was applied to cut the signals into frames. A 50% overlapped windowing was applied (except for the two- or the single-point frames). Two hybrid signals were synthesized by inverse STFT using the magnitude spectrum of the speech (or the noise) and the phase spectrum of the noise (or the speech). The first type will be referred to as magnitude-spectrum speech (MSS) and the second type as phase-spectrum speech (PSS). A triangular window, with a frame length equal to the rectangular window used for the analysis, was applied to each synthesized frame to avoid discontinuities between successive frames. Sixteen frame lengths (1/16–2048 ms) were used, including the limit case of only a single-sample time frame of 1/16 ms. The total set of materials consisted of 192 processed sentences (96×2): six sentences for each of the 16 frame lengths and two types of hybrid signals.

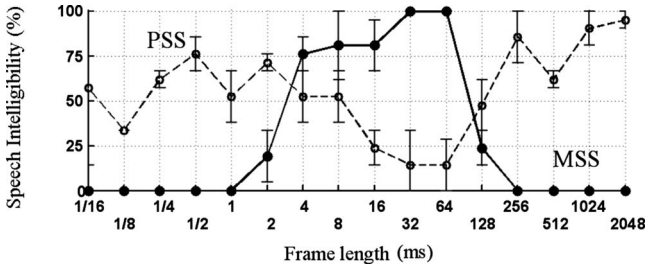

FIG. 2. Sentence intelligibility for PSS and MSS, as a function of the frame length used in the STFT procedure.

2. Subjects and procedure

The listeners were seven men, aged 24–49. They were all native speakers of Japanese. The total set of processed sentences were presented in random order through headphones (AKG-K240) under diotic listening conditions at an individually preferred level. Each subject was asked to write down the sentences as they listened. A sentence was considered intelligible only if the complete sentence was written down correctly.

B. Results

Figure 2 shows the sentence intelligibility scores (with the standard deviation) for each signal type and frame length. The horizontal axis shows the frame length. Each data point is based on six presentations to seven listeners. A score of 100% indicates that all subjects could correctly write down each sentence.

The MSS hybrid signal shows the strongest effect of frame length, ranging from perfectly intelligible for medium-range frames (4–64 ms) to totally unintelligible for long frames (256–2048 ms) and very short frames (1/16–1 ms). The PSS signals show the opposite behavior, though less extreme. For the shorter time frames, the results above suggest that frequency resolution finer than 250 Hz (frame length longer than 4 ms) is needed to get intelligible speech from the spectral magnitude. For the longer time frames, the results suggest that the temporal resolution required to obtain intelligible speech from the magnitude spectrum should be better than about 128 ms, corresponding to a modulation frequency of 8 Hz.

It is interesting to note that, where the magnitude spectrum fails in reproducing intelligible speech, the phase spectrum (partly) takes over this role. For the longer time frames, this corresponds with the earlier observations of Oppenheim and Lim[10] and Liu *et al.*[11] This is consistent with the idea that the temporal properties or signal dynamics represented by the envelopes can be expressed as the very local characteristics of the phase spectrum, such as the group delay. That

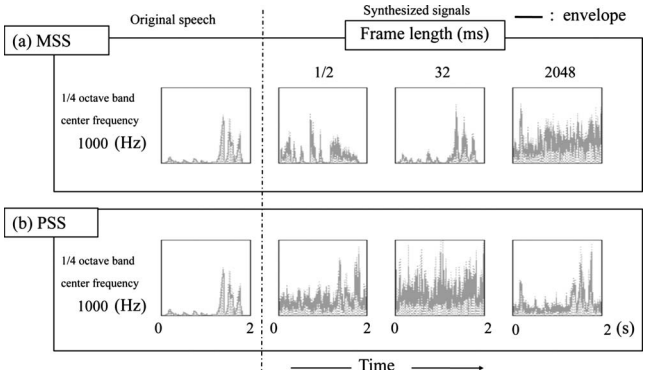

FIG. 3. Samples of squared sub-band waveforms with envelopes for the original speech, and for the MSS and PSS synthesized signals for three frame lengths used in the STFT. The sub-band considered is the 1/4-octave band with 1 kHz center frequency.

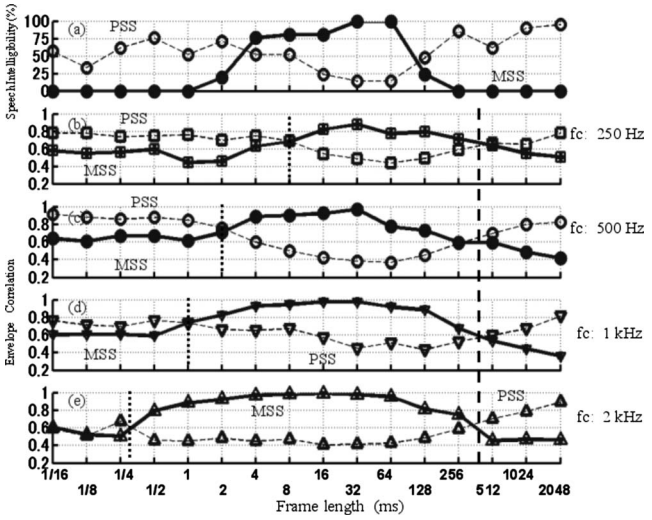

FIG. 4. Sentence intelligibility (a) and examples of envelope-correlation analysis [(b)–(e)] for MSS and PSS. Envelope correlation analysis was made in 1/4-octave bands following Eq. (1) in the text.

is, phase spectra with a fine spectral-resolution (as resulting from long time frames) will allow a partial reconstruction of the narrow-band temporal envelope.

The observed significance of the phase spectrum for the very short time frames is more surprising. This will be discussed later.

III. PRESERVATION OF NARROW-BAND ENVELOPES

As mentioned before, it is generally believed that preservation of intelligibility is related to preservation of narrow-band envelopes.[12–14] In this section, it will be investigated to what extent the narrow-band envelopes are preserved for the two types of hybrid signals. Sub-band signals of 1/4-octave bands between 250 and 4000 Hz were derived by applying a finite impulse response (FIR) filter bank (fourth-order Butterworth filter).[13] The envelope in each frequency band was defined by a Hilbert transform.

A. Observation of synthesized signals and their envelopes

Examples of the envelopes of the original and the hybrid speech signals of MSS and PSS are shown in Fig. 3 for one of the sentences in the stimulus set. This example takes one frequency band (1 kHz 1/4-octave band) and three choices of the time window (1/2, 32, and 2048 ms), motivated by the intelligibility data in Fig. 2. The narrow-band envelope of the MSS illustrated in Fig. 3(a) resembles the original envelope, only for the frame length of 32 ms. The envelope samples of PSS [Fig. 3(b)] show the opposite behavior: resemblance with the original envelope is seen only for the very short and long time frames. The observed qualitative agreement between envelope preservation and the intelligibility data motivated a more detailed study of the two types of hybrid signals in terms of narrow-band envelope correlations as a function of window length.

B. Narrow-band envelope-correlation analysis

The narrow-band envelope-correlation analysis is performed between the original and synthesized speech materials. The nature of the narrow-band temporal envelopes of the signals was evaluated by determining the correlation coefficients between the original and hybrid-signal envelopes. The correlation was calculated for every 1/4-octave band. The correlation coefficient of the ith frequency band for a sentence l is defined as

$$\rho_i(l) = \frac{\overline{\hat{e}_{oi}(l,n)\hat{e}_{si}(l,n)}}{\sqrt{\hat{E}_{oi}(l)\hat{E}_{si}(l)}}, \tag{1}$$

where

$$\hat{e}_{oi}(l,n) = e_{oi}(l,n) - \overline{e_{oi}(l,n)}, \tag{2}$$

$$\hat{e}_{si}(l,n) = e_{si}(l,n) - \overline{e_{si}(l,n)}, \tag{3}$$

$$\hat{E}_{oi}(l) = \overline{\hat{e}_{oi}(l,n)^2}, \tag{4}$$

$$\hat{E}_{si}(l) = \overline{\hat{e}_{oi}(l,n)^2}, \tag{5}$$

and $e_{oi}(l,n)$ and $e_{si}(l,n)$ denote the squared envelopes of the original and synthesized speech signals in the ith band for the sentence l, and the over-line shows taking an appropriate time average. For each frame length and for each of the two hybrid-signal types, the average was taken of the correlation coefficients for the six sentences used for that condition.

Figures 4(b)–4(e) present examples of the correlation coefficients between the narrow-band envelopes of the hybrid signals and the original speech for each of four 1/4-octave bands. Figure 4(a) is just a replication of the intelligibility test results in Fig. 2. The pattern of the correlation coefficients, as a function of the time-window length, is somewhat frequency-band dependent, but the complementary nature of the correlation values for MSS and PSS is observed for each frequency band. The intelligibility data and the narrow-band envelope correlations show the same trend with respect to the effect of frame length. This correspondence between the intelligibility data and the narrow-band envelope data confirm that the preservation of the narrow-band temporal envelopes is closely related to speech intelligibility.

The correlation data for MSS and PSS show two cross-over points. The cross-over point at a frame length of about 256 ms is almost independent of the frequency band considered, as can be seen by the vertical broken line through the figures. Since the observed decrease in the correlation for MSS toward long frame lengths reflects the loss of time resolution required for representing the temporal envelope, this crossover point is supposed to be related to the dominant frequency of the envelope modulations. The corresponding cross-over point in the intelligibility data is considerably lower, suggesting that the speech envelope includes slow modulations, which are included in the correlation values, but contribute little to speech intelligibility. This point is addressed in Sec. IV.

The other cross-over point is frequency dependent as can be seen by the vertical dotted lines in each of the figures. The cross-over points happen to correspond roughly with the duration of the period of the center frequency. We cannot provide a firm theoretical basis for this relation. In general terms, however, the frequency dependency of the cross-over point can be understood as a reflection of the limited frequency resolution associated with a short frame length. Given the increase in bandwidth for increasing center frequency f_c of the 1/4-octave bands considered in Fig. 4, a certain loss of frequency resolution (typically equal to the inverse of the frame length in the STFT) will have less effect for higher $f_c's$. Thus, in order to recover 1/4-octave band envelopes from the magnitude spectrum, the frame length used in the STFT should provide an adequate degree of frequency resolution, related to the width of the frequency band considered. Hence, shorter frames are allowed toward higher $f_c's$.

The results so far can be summarized as follows.

(1) *The MSS data are quite understandable*. For longer time frames (>256 ms), the temporal resolution is insufficient to follow the relevant envelope modulations, and for shorter time frames (<4 ms), the frequency resolution becomes insufficient (this appears to depend on the center frequency of a band).
(2) *The PSS data are more surprising*. The envelope is (partly) recovered for windows longer than 256 ms, and also for the very short time frames (which may not be intuitively obvious for many readers).

These observations on the phase dominance for longer and for very short time frames will be studied further by analyzing narrow-band envelope recovery from the phase spectrum only.

C. Recovery of narrow-band envelopes from the phase spectrum

1. Significance of phase spectrum for long window lengths

The importance of the phase spectrum for modulated signals is well illustrated by the difference between an amplitude- and a quasi-frequency-modulated (AM and QFM) sinusoid. It is well known that the phases of the two side-band components determine the temporal envelope: essentially flat in the QFM case and modulated in the AM case. Figures 5(a) and 5(b) show a stationary random noise and a noise modulated by a co-sinusoidal function, respectively. The corresponding magnitudes and phase spectra are shown in Figs. 5(c)–5(f). The envelope-modulation frequency in this example is given by $2(1/N)$, where N denotes the signal length. Here, STFT analysis was applied to the whole signal length.

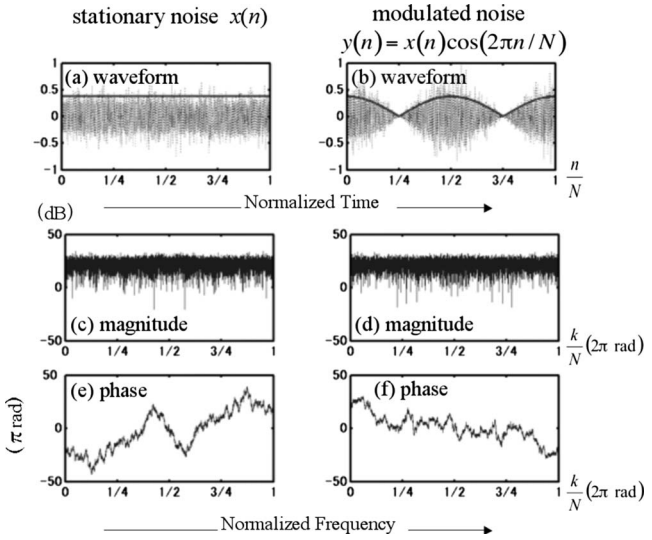

FIG. 5. Examples of stationary random noise (a) and modulated noise (b) with the magnitude [(c) and (d)] and phase [(e) and (f)] spectral characteristics.

Although there are no clear indications of the envelope frequency in the magnitude and phase spectra, the frequency can be observed by applying an auto-correlation analysis to the phase spectrum. This is illustrated in Fig. 6, where the modulation frequency is converted to a real quantity. When the phase difference between components (k_o) and $(k+k_o)$ is given by

$$\Delta\theta(k,k_o)=\theta(k+k_o)-\theta(k_o), \tag{6}$$

then the phase correlation function phc(k) can be obtained by

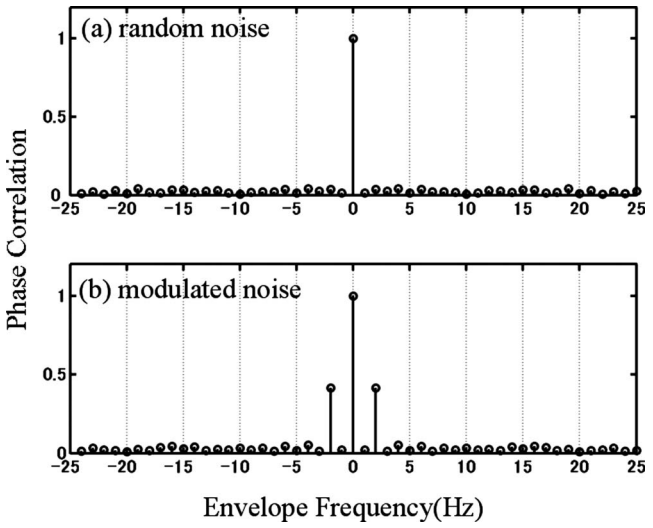

FIG. 6. Phase spectrum auto-correlation analysis for the signals shown in Fig. 5 according to Eqs. (6)–(9) in the text.

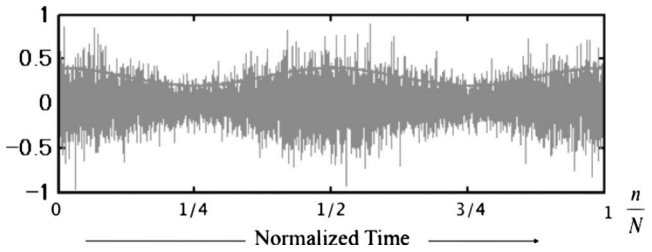

FIG. 7. Reconstruction of the modulated noise of Fig. 5(b), using the corresponding phase spectrum and a random magnitude spectrum.

$$\mathrm{phc}_c(k) = \frac{1}{K}\sum_{k_o=0}^{k_o=K-1} \cos\Delta\theta(k,k_o), \tag{7}$$

$$\mathrm{phc}_s(k) = \frac{1}{K}\sum_{k_o=0}^{k_o=K-1} \sin\Delta\theta(k,k_o), \tag{8}$$

$$\mathrm{phc}(k) = \sqrt{\mathrm{phc}_c(k)^2 + \mathrm{phc}_s(k)^2} \tag{9}$$

in a discrete form. Here, K denotes the number of frequency components of interest. In the figure, the horizontal axis shows the frequency shift, which can be interpreted as the envelope frequency.

Note that Fig. 6(a) shows that the fluctuations seen in the phase spectrum for stationary noise are random. Only the modulated-noise case [Fig. 6(b)] shows that the modulation frequency can be estimated from phase information alone. Figure 7 is an example of a hybrid signal made by substituting random magnitude for the original magnitude spectrum of the modulated noise shown in Fig. 5(b). This illustrates that the original envelope is partly preserved on the basis of the phase spectrum only. The spacing of the frequency components in the phase spectrum resulting from the STFT should be small enough to reflect the envelope frequency in the phase spectrum auto-correlation function. Since this frequency spacing is related to the inverse of the frame length used in the STFT, this implies that for envelope recovery from the phase spectrum, the frame length should be longer than the period of the envelope modulation of interest.

2. Significance of phase spectrum for very short window lengths

Phase dominance for very short frame lengths can be interpreted as the narrow-band envelope recovery from the zero-crossings of a waveform. As Figs. 4(b)–4(e) indicate, this requires that the frame length is shorter than the period of the center frequency of interest. For the present study, the limit case of a very short analysis window-lengths is a length of 1/16 ms (i.e., the sampling rate), corresponding to a single-point STFT. The result of a single-point STFT is for each sample, its magnitude, and the phase is just the sign of the sample, ±. Thus, the phase information of a single-point STFT keeps the zero crossings of the original signal, if the sample frequency is adequate. This is the same as applying infinite peak clipping to a signal, which also preserves the zero-crossing information while losing all amplitude information. It will be shown that the narrow-band envelopes can be partly recovered from the zero crossings of the original signal.

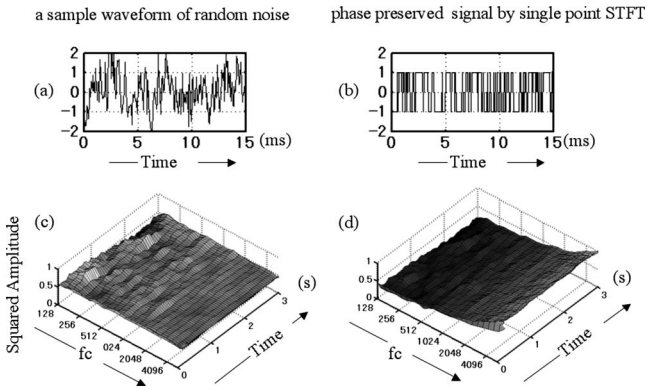

FIG. 8. Representation of stationary noise. Left part: original. Right part: after single-point STFT, with the phase (+ or −) of each sample preserved, and the magnitude set at unity.

Figure 8(a) shows a snap shot of a waveform from a stationary random noise. Figure 8(b) represents the random noise re-synthesized by preserving the phase only for a single point of STFT (just ±) with the magnitude of unity. This resulted in an infinitely peak-clipped version of the original signal. The three dimensional plots in Figs. 8(c) and 8(d) show the temporal change in the short-time sub-band energy of these two signals, indicating that the original spectro-temporal characteristics, especially in the low frequency range, are preserved to some extent in the signal synthesized from the phase information of the single-point STFT. The increase in the energy for the high frequency bands in Fig. 8(d) can be interpreted as a processing noise due to the hard clipping of the waveform shown in Fig. 8(b).

The example above refers to a stationary signal. It illustrates that the shape of the spectral magnitude information is hidden in the pattern of the zero-crossings, particularly for the low frequency bands. Figure 9 is another example for a random noise, but now including a modulated sub-band. The original zero-crossings are preserved in Fig. 9(b) by the single-point STFT. Despite losing the original magnitude information, the temporal envelope of the sub-band can be recovered to some extent from the zero-crossing information, as shown in Fig. 9(d). In other words, the narrow-band en-

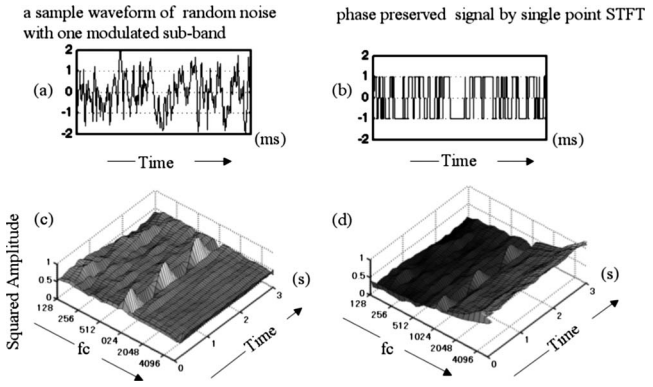

FIG. 9. A sample waveform of random noise with a modulated sub-band. Left part: original. Right part: similar to Fig. 8.

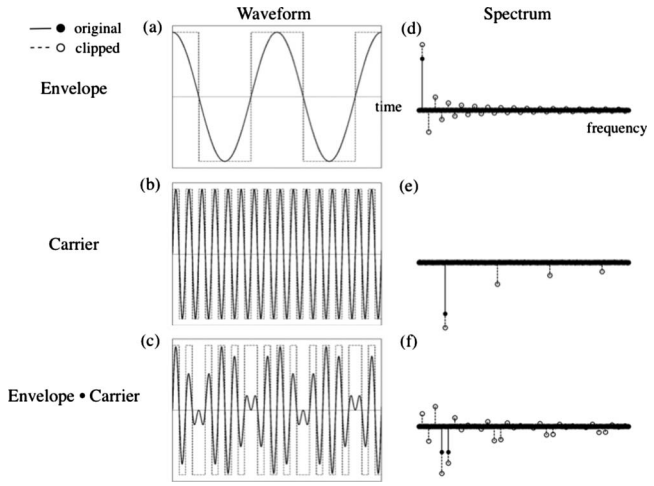

FIG. 10. Spectrum of the infinitely clipped version of a modulated sinusoidal signal.

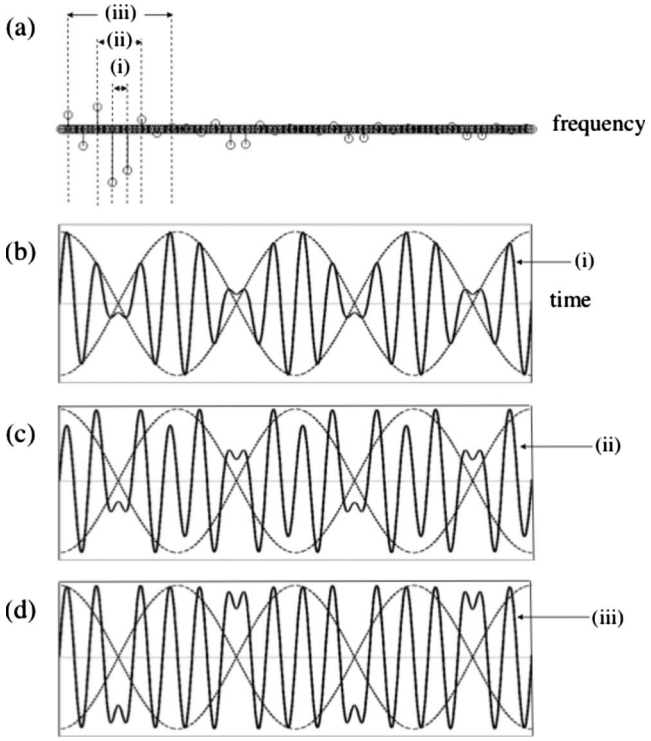

FIG. 11. Sinusoidal envelope recovery from the clipped wave as shown in Fig. 10, after applying sub-band filtering with increasing bandwidth, indicated by (i), (ii), and (iii).

velope can be partly recovered from the fine-structure (zero crossings) of the modulated noise samples after sub-band analysis.

This example may explain the significance of the phase for the very short time frames, as observed in Fig. 4. However, Fig. 4 also indicates that for the very short time frames the envelopes for the higher frequency bands are not well recovered from the phase-only information. This may be related to the fact that, given the typical shape of the power spectrum of speech, the higher bands represent only a small fraction of the total power, and consequently the modulation properties of these higher bands may only be marginally represented in the zero-crossing statistics of the over-all signal. Another possibility is that envelope reconstruction from phase is disrupted by the processing noise that yields higher energy at higher frequencies, as shown in Figs. 8(c), 8(d), 9(c), and 9(d).

Another example of envelope recovery from zero-crossing information is provided in Figs. 10 and 11. It concerns a modulated sinusoidal waveform, as shown in Figs. 10(a)–10(c). Spectral records for the envelope [Fig. 10(d)], its carrier [Fig. 10(e)], and the modulated signal [Fig. 10(f)] are represented by the line-spectral characteristics. Here, the solid lines and solid circles show the original ones, while the dotted lines and open circles indicate the infinitely clipped case. The spectral structure of the modulated signal can be expressed as the convolution of the spectral sequences for the envelope and the carrier. The convolution is performed periodically, because this numerical sample is composed of a discrete sequence. If only the zero-crossing property is preserved with the magnitude of unity (discarding the envelope of the modulated sinusoid), the convolved spectral-structure is expanded, including its higher harmonics. Although those higher harmonics are not contained in the original modulated signal, the modulation property, such as the temporal envelope, can be recovered by applying appropriate filtering, as shown in Fig. 11.

Figure 11(a) is close-up of Fig. 10(f). If we take a bandwidth denoted by (i) in the figure, representing sub-band analysis, then we get the waveform shown in Fig. 11(b). Here, the broken line represents the original envelope shown in Fig. 10(a). However, if we increase the bandwidth according to the examples denoted by (ii) or (iii) in Figs. 11(c) and 11(d), the original envelope is no longer recovered. This illustrates that the original envelope can be recovered from zero-crossing information when applying sub-band filtering, provided that the bandwidth is adapted to the modulation frequency of interest. Since in our analysis, higher frequencies are associated with broader absolute bandwidth, this may be a reason why envelope recovery from phase for very short windows is poorer at high frequencies (Fig. 4), and why the processing noise increases toward high frequencies [Figs. 8(d) and 9(d)].

In principle, characteristics of speech waveforms can be understood as a mixture of a random noise-like feature or a periodic structure. Thus, the two simplified examples presented above represent two extreme cases of signals with speech-like characteristics. For both some form of sub-band-envelope recovery from the zero-crossing information only was demonstrated.

IV. DISCUSSION

The results of the listening experiment, as presented in Fig. 2, provide the key data of the present study. We will first consider the strong effect shown in the MSS data, in particular, the decrease to 0% intelligibility for the long and the short frame lengths when using the magnitude spectra only.

In interpreting the loss of intelligibility of MSS speech for time windows of over about 250 ms, it was assumed that this reflects the loss of temporal resolution required for following the essential speech-envelope modulations. This

would suggest that the envelope modulations above 4 Hz are indispensable for maintaining intelligible speech. A related study on the effect of low-pass filtering narrow-band envelopes[15] indicated that only for a 2-Hz (or lower) low-pass cut-off frequency is sentence intelligibility severely reduced. The difference may be understood by realizing that, in this previous study, the envelope filtering was the only "distortion" applied to the speech, while in the present study many additional "distortions" are introduced in the MSS condition, as a result of disregarding the phase spectrum.

The loss of intelligibility of MSS speech for time windows of 1 ms or less was interpreted as reflecting the very limited frequency resolution associated with such brief time windows. This would imply that a frequency resolution of worse than about 1000 Hz makes speech unintelligible. In related studies, e.g., on the effect of spectral smearing on speech intelligibility,[16] or on the minimum number of bands required to produce intelligible speech,[17] the operations are performed on a logarithmic frequency scale, complicating the comparison with the present study (window-associated loss of spectral-resolution is constant over frequency). Still, the 1000-Hz limit suggested by the present data does not disagree with the findings of these other studies.

For a frame length in the range of 4–64 ms, the magnitude spectrum carries the essential information for speech intelligibility. Thus, for the commonly observed approach in speech processing, i.e., a spectral analysis with a window length of a few tens of ms, the use of the power spectrum does maintain the essential cues for speech intelligibility. Following the traditional view of the peripheral auditory system as a set of band pass filters, the auditory temporal window in the mid-frequency range amounts to typically 10 ms.[18] Figure 4 indicates that for this window length, both the speech intelligibility and the narrow-band envelope preservation are dominated by the magnitude spectrum, while the phase only plays a minor role. This implies that the envelopes of the auditory filter outputs carry the intelligibility-relevant information.

The main goal of the present study was to investigate the relative importance of magnitude versus phase in the short term Fourier-spectrum approach in speech analysis and synthesis, given that most studies concentrate on the magnitude or power spectrum only. The experimental signal manipulations used in this study (resulting in the hybrid signals) will have consequences for the preservation of the envelope and the fine structure in narrowband (auditory) filters. In this respect, there is a relation between this study and research on the relative importance for speech intelligibility of envelope cues and fine-structure (phase) cues at the auditory filter outputs. The consensus on this issue (see, among others, Refs. 14, 13, and 19) is that preservation of the envelopes at the filter outputs is the main factor for speech intelligibility, while the phase or the fine structure is of secondary importance. This firm distinction is somewhat complicated by studies, indicating that envelope and phase information at the filter outputs are not independent,[20] and that temporal-envelope cues can be recovered from the speech fine structure.[21] A detailed study on the effect of additive noise on speech intelligibility, quantifying the relative importance of disrupting the narrow-band envelopes or the fine structure, confirmed the importance of the envelopes.[22] However, besides the main role of envelope cues, that study also showed some reduction in intelligibility after disrupting the fine structure cues. This is in line with various other studies, indicating that temporal fine-structure cues do play a role in speech intelligibility, especially in case of complex (i.e., non-stationary) maskers.[23–25]

The relevance of narrow-band envelopes for speech intelligibility motivated the use of this concept (i.e., the degree of preservation of narrow-band envelopes) for interpreting the present intelligibility data on the relative importance of the magnitude or phase spectrum. However, it should be realized that, besides the loss of envelope correlation for the pre- and post-processed speech, other types of distortions are introduced as well. For instance, it has been shown that a loss of cross-spectral modulation phase coherence may reduce speech intelligibility.[26] It is very probable that the observed loss of correlation between the pre- and post-processed narrow-band envelopes is associated with a loss of cross-spectral modulation phase coherence. Also, as mentioned before, the processing will affect the narrow-band fine structure. Consequently, part of the observed relation in the present study between loss of intelligibility and loss of band-envelope correlation may well be caused by the associated effects of loss of cross-spectral modulation phase coherence or loss of fine structure. The present study does not allow to specify this any further.

V. CONCLUSION

Speech was subjected to Fourier analysis and synthesis, using the overlap-add procedure with window lengths ranging from 1/16 to 2048 ms. Experiments were performed on the intelligibility of the speech for two conditions applied for the synthesis: (a) MSS-mode, using the speech-magnitude spectra with randomized phase spectra, and (b) PSS-mode, using the speech-phase spectra with randomized magnitude spectra. Besides the intelligibility measurements, the signals were subjected to an analysis of the correlation between the narrow-band envelopes, before and after the MSS or the PSS synthesis mode. The main findings were as follows.

(1) Using the MSS synthesis mode (magnitude spectra only), intelligible speech was obtained only for frame lengths of 4–64 ms.
(2) When using the PSS synthesis mode (phase spectra only), reasonably intelligible speech was obtained for frames longer than 128 ms or shorter than 4 ms.
(3) Thus, the two curves of intelligibility, as a function of frame length for the MSS and PSS synthesis mode, show a complementary character. This means that, for speech intelligibility, for the medium-range windows (4–64 ms), the magnitude spectrum dominates, and for the two extreme regions (<4 and >64 ms), the phase spectrum dominates.
(4) Intelligibility scores and correlation coefficients between the synthesized and original envelopes in 1/4-octave bands showed the same trend with respect to the effect of frame length, although not identical. This qualitative cor-

respondence confirms that the preservation of narrow-band temporal envelopes constitutes an important factor for the preservation of speech intelligibility.

The interpretation of these findings may be summarized as follows.

(a) *Considering the MSS synthesis mode (magnitude spectra only)*. For long time frames, the speech becomes unintelligible due to a loss of time resolution, as the frame length becomes longer than the period of the dominant speech envelope modulations (about 256 ms, a 4-Hz modulation frequency). For frame lengths shorter than 4 ms, the speech becomes unintelligible because the corresponding frequency resolution is insufficient.
(b) *Considering the PSS synthesis mode (phase spectra only)*. For long time frames, it is shown that the phase spectrum contains envelope information, as reflected in phase-spectral auto-correlation analysis. For the very short frames, it is shown that the phase-only synthesized speech essentially keeps the zero-crossing interval information, from which the shape of the original power spectrum can be obtained. It is shown that the temporal envelope of a sub-band can be partly recovered from the zero crossings of the total signal.

The main result of this study is that, besides the dominance of the magnitude spectrum for the middle range of window lengths, there appear to be two regions of phase dominance with respect to intelligibility and preservation of narrow-band envelopes. This phase dominance applies to very short and long time windows.

ACKNOWLEDGMENT

This research was partly supported by the Telecommunications Advancement Research Fellowship, Japan.

[1]M. R. Schroeder, "Computer speech," *Springer Series in Information Sciences* (Springer-Verlag, Berlin Heidelberg, 1999), pp. 63–73.
[2]S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **27**, 113–120 (1979).
[3]P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," Signal Process. **8**, 387–400 (1985).
[4]M. R. Schroeder, "Computer speech," *Springer Series in Information Sciences* (Springer-Verlag, Berlin Heidelberg, 1999), pp. 113–127.
[5]M. Schroeder and H. Strube, "Flat-spectrum speech," J. Acoust. Soc. Am. **79**, 1580–1583 (1986).
[6]M. R. Schroeder, "Models of hearing," Proc. IEEE **63**, 1332–1350 (1975).
[7]H. Pobloth and W. Kleijn, "Squared error as a measure of phase distortion," in Proceedings of the EUROSPEECH (ISCA) (2001), pp. 1973–1976.
[8]R. Plomp and H. Steeneken, "Effect of phase on the timbre of complex tones," J. Acoust. Soc. Am. **46**, 409–421 (1969).
[9]A. Traumueller and M. Schouten, *The Psychophysics of Speech Perception* (Kluwer, Dordrecht, 1987).
[10]A. Oppenheim and J. Lim, "The importance of phase in signals," Proc. IEEE **69**, 529–541 (1981).
[11]L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," Speech Commun. **22**, 403–417 (1997).
[12]T. Houtgast, H. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72 (1980).
[13]R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592 (1995).
[14]R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science **270**, 303–304 (1995).
[15]R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064 (1994).
[16]M. ter Keurs, J. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception II," J. Acoust. Soc. Am. **93**, 1547–1552 (1993).
[17]L. Friesen, R. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163 (2001).
[18]C. Plack and B. Moore, "Temporal window shape as a function of frequency and level," J. Acoust. Soc. Am. **87**, 2178–2187 (1990).
[19]Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90 (2002).
[20]O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," J. Acoust. Soc. Am. **110**, 1628–1640 (2001).
[21]G. Gilbert and C. Lorenzi, "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444 (2006).
[22]F. Dubbelboer and T. Houtgast, "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871 (2007).
[23]D. Gnansia, V. Péan, B. Meyer, and C. Lorenzi, "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033 (2009).
[24]C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869 (2006).
[25]S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575 (2008).
[26]S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in Proceedings of the EUROSPEECH (ISCA) (2001), pp. 473–476.