Speech Intelligibility Estimation Based on Phase Correlation Index

S.GOTOH(1),M.KAZAMA(2),and M.TOHYAMA(3)
(1)ahjinsei@yahoo.co.jp, Kogakuin University, Tokyo, Japan
(2)mich@ann.hi-ho.ne.jp, Waseda University, Tokyo, Japan
(3)m_tohyama@waseda.jp, Waseda University and University of York, UK

Abstract
Intelligible speech information is contained in the phase for long-term spectrum, in the amplitude for mid-term spectrum, or in the phase for short-term spectrum. The phase or magnitude spectrum dominance on intelligible speech representation can be evaluated by narrow-band envelope recovery. This article proposes a method for estimation of speech intelligibility in a noisy environment based on frequency analysis of the narrow-band envelopes, which can be represented by phase correlation index(PCI). PCI could be evaluated from the signals which are used for listening tests of speech intelligibility. Intelligibility tests using a set of noisy three-nonsense syllables indicated that PCI might be a good candidate for estimator of speech intelligibility.

Keywords
Speech analysis and synthesis, Speech intelligibility, SI

1 Introduction

Phase information has not been used much in speech processing. Oppenheim and Lim [1], however, found that when a speech signal is of sufficient length, speech intelligibility is lost in Fourier-transform amplitude-only reconstruction but not in phase-only reconstruction. Liu et al[2] showed that even for shorter window size the effect of the phase on the perception could be observable by investigating the effect of the phase on intervocalic stop consonant perception for VCV speech signals.

On the contrary Drullman[3] indicated that the narrow-band speech temporal envelope is an essential element for producing intelligible speech. The authors also have been intensively investigating the effect of phase on intelligible speech reconstruction. Recently Kazama et al[4] described that the dominance of magnitude or phase can be estimated by recovery of narrow-band envelopes of speech. In this article we will introduce a new measure for speech intelligibility based on the frequency analysis of the narrow-band envelope, which can be done using phase correlation statistics. Speech intelligibility scores obtained by listening tests in a noisy environment are estimated by the new measure defined as the phase correlation index.

2 Intelligibility for Reconstructed Speech from Magnitude or Phase-only Spectrum

In the first experiment, we presented an example of the significance of the amplitude or phase spectrum to intelligible speech, using amplitude and phase combined (hybrid) speech signals[2][4]. The primary signals were speech and noise. Thus, the two hybrid signals were of the nature "speech amplitude, random phase (Type A)" or "random amplitude, speech phase (Type B)". The effect of window length was estimated based on the results of intelligibility tests, as well as on those of narrow-band envelope-correlation studies.

The listeners were three females and seven males, aged 22-53. They were all native Japanese with normal hearing. As original speech signals we used two sets of 27 phrases, spoken by two female speakers. All of the speech materials were in Japanese and digitized at a sampling rate of 16kHz using a 16-bit A/D. Each 1.5-s-long speech phrase had additional silent parts at the start and end, so the total length was 4s. We produced a white-noise signal that was 4s long by using MATLAB software.

We analyzed all the original pairs using short-term Fourier transform (STFT) (Fig. 1). We used a rectangular-window function to cut the speech data into frames. Each frame started with the data points in the previous frame's last half. Hybrid signal type A (B) was synthesized by inverse STFT using the amplitude spectrum of speech (noise) and the phase spectrum of noise (speech). A triangular window was applied to each synthesized frame waveform to avoid discontinuities between successive frames.

We used nine frame lengths (8, 16, 32, 64, 128, 256, 512, 1024, and 2048 ms). Each of the nine possible pairs resulted in nine (frame lengths) pairs of type A or type B hybrid signals for use in the listening tests. To determine the narrow-band envelope cross-correlations, we split the original primary speech signals and the hybrid signals into 17 1/4-octave bands following Drullman's approach [3] by using an FIR filter bank (fourth-order Butterworth filter) at 250 to 4000 Hz, which is almost equal to the telephone band. The envelopes were defined by a Hilbert transform in each 1/4-octave frequency band.

Each subject listened to a different subset composed of 54 hybrid signals using each of the 54 original phrases no more than once. Each kind of hybrid sig-

nal (type A or B and nine frame lengths) was included three times in each subset. The 54 hybrid signals were presented in random order through headphones. Each subject was asked to write down the phrases as they listened. The data were pooled with regard to type A or B and frame length. With nine frame lengths, each data point was based on 30 presentations. A sentence was considered intelligible only if the complete sentence was written down correctly. We evaluated the preservation of the narrow-band temporal envelopes by using the cross-correlation between the envelopes of the original speech and the hybrid signals. The correlation was calculated for every 1/4-octave band and averaged over the frequency bands.

As shown in Fig. 2(a), in terms of intelligibility, the type A hybrid signals (i.e., using the speech amplitude and the random phase spectrum) showed the strongest effect of frame length: from perfectly intelligible for short frames to totally unintelligible for the longer frames. The type B signals showed the opposite behavior, though less extreme. The correlation data as a function of frame length for the type A and B hybrid signals (Fig. 2(b)) show a similar pattern, although not identical.

First, the crossover point falls at a slightly longer frame length. This may indicate that the speech envelope includes slow modulations, which are included in the correlation values but do not contribute to speech intelligibility. Second, and more surprisingly, the correlation values for the type A and B signals show an almost perfect complementary behavior. The two values appear to add up to one, but we cannot provide a theoretical framework as to why this would be the case.

All in all, the qualitative correspondence between Figs. 2(a) and (b) confirms that the preservation of the narrow-band temporal envelopes constitutes an important factor for the preservation of speech intelligibility.

Interestingly the narrow-band envelopes are recovered from the phase spectral records, when the frame length becomes shorter than 8 ms and even if it reaches a single sample of length[5]. It suggests that the phase information is crucial for both cases of short and long frame lengths.

3 Envelope Frequency Analysis Using Phase Correlation

Envelope frequency can be estimated by phase spectrum analysis. Figure 3 is an example of a modulated noise. Figure 3(a) shows a time waveform $x(n)$, and 3(b) gives the phase spectrum, and Fig.3(c) presents the phase correlation sequence $PCS(k)$, which is defined as

$$PCS \equiv \sqrt{S^2 + C^2} \quad (1)$$

$$S \equiv \frac{1}{N}\sum_{l=0}^{N-1} \sin \Delta\theta(k,l) \quad (2)$$

$$C \equiv \frac{1}{N}\sum_{l=0}^{N-1} \cos \Delta\theta(k,l) \quad (3)$$

$$\Delta\theta(k,l) \equiv \theta(l+k) - \theta(l) \quad (4)$$

$$\theta(k) \equiv \text{phase spectrum of } x(n). \quad (5)$$

We can see the frequency spectrum of the envelope by $PCS(k)$. We will use $PCS(k)$ for speech intelligibility estimation, because envelope recovery is significant for intelligible speech reconstruction.

4 Speech Intelligibility Tests in Noisy Environment

Figure 4 illustrates the loudspeaker locations in an anechoic room for speech intelligibility tests. Uncorrelated random noise were continuously generated from six loudspeakers. Every target signal is composed of three nonsense syllables in Japanese. A series of the targets were addressed by one of the six loudspeakers which was randomly selected for every target. The signal to noise ratio of signals reproduced from the loudspeaker which radiates speech syllables and random noise was set to be 0 or 30 dB. A set of 150 targets, composed of 120 targets for 0 dB(30 samples for 30 dB), were used for the experiments. No signal-manipulation such as frequency-band limitation or equalization was performed for the input signals fed into loudspeakers. And no other environmental noise was added.

All the signals were recorded at the center position as shown in Fig.4 using a dummy head manufactured by Brüel and Köer, and digitized at a sampling rate of 44.1kHz using a 16-bit A/D. Listeners were four males including one of the authors, aged 22-57, and native Japanese with normal hearing, but they were naive in the intelligibility tests except one of the authors. They listened to the recorded 150 targets through audio-headphones(AKG) by preferable listening levels under dichotic condition and were asked to write down the syllables as they listened. Intelligible scores were recorded if the second syllable was written down correctly.

Figure 5 is an example of narrow-band waveform for the noisy targets. Panel(a) and (c) in the upper row are noise records without speech, while panel (b) shows a record of three syllables including noise. The lower panels are PCS corresponding the upper ones, respectively. Here panels (a) and (c) denoted by PCS_{fN} and PCS_{bN} are PCS for the noise intervals, while panel (b) indicated by PCS_{SN} is for the interval including the target syllables. We assume that speech syllables might be intelligible, as the difference becomes large in PCS among the intervals described above. Therefore we evaluate the speech intelligibility score according to the phase correlation index PCI, which is defined as the difference of PCS

$$PCI(i) \equiv \frac{1}{M}\sum_{m=1}^{M} W(m) PCI(i,m) \quad (6)$$

$$PCI(i,m) \equiv 10 \log \frac{D(i,m)}{D(m)} \quad (7)$$

where we set

$$D(i,m) \equiv \sum_{k=0}^{N-1} B^2(i,m) + F^2(i,m) \quad (8)$$

$$B(i,m) \equiv PCS_{SN}(i,m) - PCS_{bN}(i,m) \quad (9)$$

$$F(i,m) \equiv PCS_{SN}(i,m) - PCS_{fN}(i,m) \quad (10)$$

$$D(m) \equiv \frac{1}{Q}\sum_{i=1}^{Q} D(i,m) \quad (11)$$

where Q is the number of samples(equal to 150) and M is the number of frequency bands(equal to 17), and $W(m)$ is a weighting function for getting the index. The frequency band from 250 to around 4700 Hz is split into 17 1/4 oct.bands.

Figure 6 demonstrates the results of speech intelligibility scores. The horizontal axis shows the PCI for each sample and the lower part of the panel illustrates the distribution of samples corresponding each PCI. The weighting function for getting an effective $PCI(i)$ is still under study. Here we simply chose the best one of the $PCI(i,m)$ in the frequency bands of 250, 500, 1,000 and 2000(Hz). And all the intelligibility scores are obtained under the dichotic listening conditions. Therefore we used the better one, which takes higher $PCS(i,m)$ between both ears following the best ear hypothesis[6].

We can see there is a good correlation between the intelligibility score and PCI. It means PCI might be a good candidate for the intelligibility estimator. However, the distributions of samples corresponding PCI is not uniform in this example. Thus we have to make another experiment by directly controlling PCI so that we might reconfirm estimating accuracy for speech intelligibility by PCI.

Figure 7 gives the relationship between the intelligibility and MTF-STI[7]. Here MTF-STI was estimated based on the signal to noise ratio of samples. We can confirm again STI could be also a good indicator for the intelligibility. It suggests that the intelligibility scores obtained in the experiment was also reasonable. STI requires the signal to noise ratio or measurements using modulated noise signals. However PCI requires only noisy speech samples which are addressed in situ conditions for the listening tests.

5 Summary

Intelligible speech can be represented by narrow-band envelopes. We have described that if the narrow-band envelopes are recovered, then intelligible speech is synthesized irrespective of using the magnitude or phase information. The envelope-frequencies can be estimated by the phase correlation, if spectral resolution for the phase is sufficient for the envelope-frequency analysis. Consequently we proposed a new measure for speech intelligibility based on the phase correlation index, PCI.

The index was derived by a weighting average for the phase correlation sequence in each frequency band. Although the weighting factor is still under investigation, we could see a good relationship between the index and intelligible scores. In this article we only took the best PCI for representative frequency bands. We also confirmed that the intelligible scores obtained by listening tests were normal by checking MTF-STI. MTF-STI is also a measure related to the envelopes, however, the relationship between the MTF-STI and the phase information has not been clearly understood, and it requires measurements using modulation noise signals or signal to noise ratio.

The phase correlation index PCI which the authors proposed in this article has a great potential in estimating speech intelligibility, and it can be obtained from noisy speech samples which are used for listening tests in situ conditions without other additional measurements. The authors would like to thank Prof. T.Houtgast for intensive discussions about the phase dominance on intelligible speech representation, and appreciate Prof. Y. Yamasaki and Prof. K. Shinohara for their continuous encouragement during this research.

References

[1] A.V.Oppenheim and J.S. Lim, J. S., The importance of phase in signals, Proc. IEEE, 69, pp.529-541, 1981.

[2] L. Liu, J.He, and G. Palm, Effects of phase on the perception of intervocalic stop consonants, Speech Communication 22, pp.403-417, 1997.

[3] R. Drullman, Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Am. 97(1), pp. 585-592, 1995.

[4] M.Kazama, M.Tohyamam, and T. Houtgast, Speech Reconstruction by Using only Its Magnitude Spectrum or Only Its Phase, 17th Int. Cong. Acoust. 7p51,2001.

[5] M.Kazama, M. Tohyama, and T. Houtgast, under preperation

[6] M.Tohyama, Modern Techniques for improving Speech Intelligibility In Noisy Environments, Noise and Man '93, Proc. 6th Int. Cong. Noise as a Public Health Problem, 3, pp.238-246, 1993

[7] T.Houtgast, H.J.M.Steeneken, and R. Plomp, Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I General Room Acoustics. Acoustica, 46, pp. 60-72,1980.

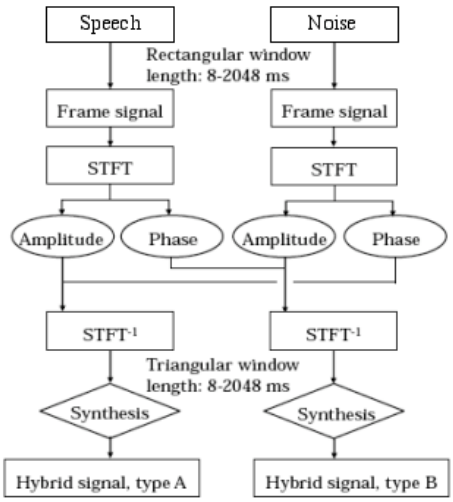
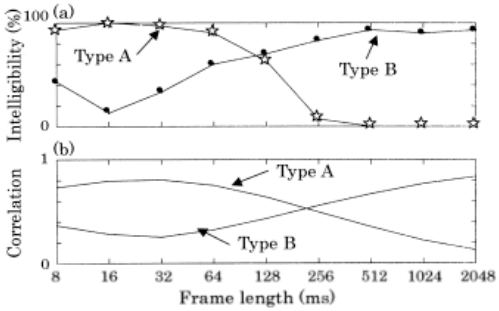


Fig.1 Method for deribing two types of hybrid signals from two primary signals using a cross-wise combination of the amplitude and phase spectra in the STFT overlap-add procedure



Type A: synthesized using speech amplitude and noise phase

Type B: synthesized using speech phase and noise amplitude

Fig. 2 Listening test and envelope correlation analysis for synthesized signals derived from speech and noise:
(a)Intelligibility scores for speech segments
(b)averaged narrow-band envelope correlation between synthesized and original speech

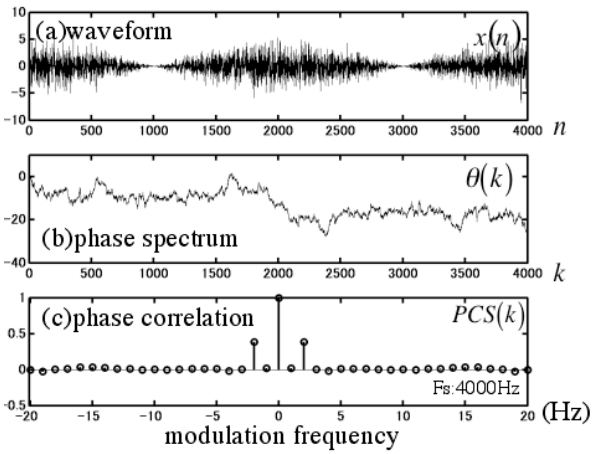


Fig.3 Phase spectrum correlation for a modulated noise signal

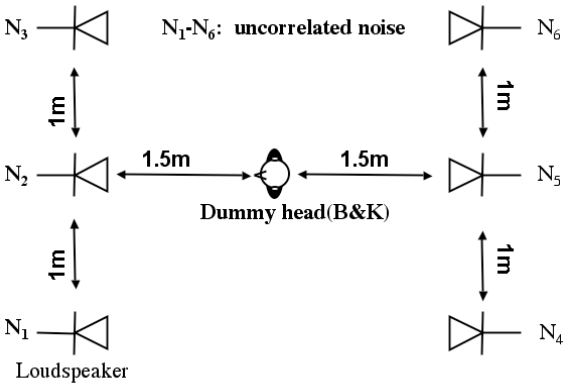


Fig.4 Noise environment for speech intelligibility tests

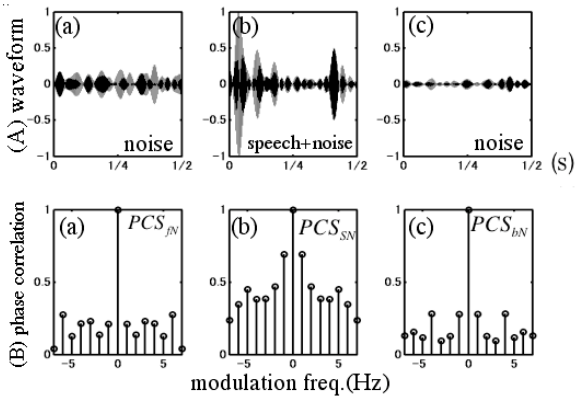


Fig.5 Samples of narrow-band noisy speech f_c: 250(Hz) bandwidth: 1/4oct.band (a)(c) noise without speech
(b) speech+noise interval speech: a series of three nonsense syllables

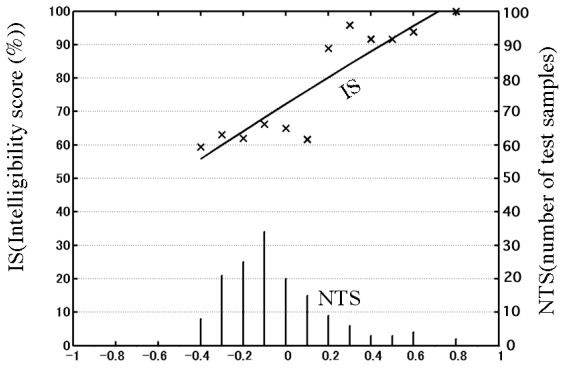


Fig.6 Intelligibility score and *PCI*

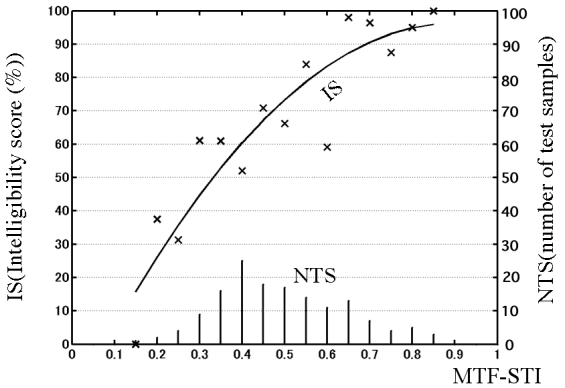


Fig.7 Intelligibility score and MTF-STI