

IMPORTANCE OF WINDOW SHAPE FOR PHASE-ONLY RECONSTRUCTION OF SPEECH

Leigh D. Alsteris and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, Brisbane, Australia
e-mail: K.Paliwal@griffith.edu.au, L.Alsteris@griffith.edu.au

ABSTRACT

The authors recently conducted a human perception experiment [6] to measure the intelligibility of speech stimuli synthesised either from short-time magnitude spectra or short-time phase spectra. The results of the experiment indicate that even for small window durations (of relevance for automatic speech recognition applications), the phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum if the analysis-modification-synthesis parameters are properly selected. This intelligibility is significantly more than that reported by Liu et al. [3], who carried out a similar experiment with the same analysis-modification-synthesis framework. The significant improvement in intelligibility over Liu's results may be attributed to the differences in the parameter settings adopted. In this paper, we review our previous experiment and conduct an additional experiment to determine the contribution that each parameter setting provides towards the intelligibility of stimuli reconstructed from short-time phase spectra. The parameter selection that contributes most to the intelligibility of the phase-only stimuli is that of a rectangular analysis window, as opposed to a Hamming window (which is generally used in speech analysis).

1. INTRODUCTION

Although speech is a non-stationary signal, it can be assumed to be quasi-stationary and, therefore, can be processed through a short-time Fourier analysis [1, 2, 8, 9, 10]. Note that the modifier 'short-time' implies a finite-time window over which the properties of speech may be assumed stationary; it does not refer to the actual duration of the window¹. The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

$$S(f, t) = \int_{-\infty}^{\infty} s(\tau)w(t - \tau)e^{-j2\pi f \tau} d\tau, \quad (1)$$

where $w(t)$ is a window function of duration T_w . In speech processing, the Hamming window function is typically used and its width T_w is normally 20-40 ms.

We can decompose $S(\nu, t)$ as follows:

$$S(f, t) = |S(f, t)|e^{j\psi(f, t)}, \quad (2)$$

This work was partly supported by ARC (Discovery) grant (No. DP0209283).

¹We use the qualitative terms 'small' and 'large' to make reference to the duration.

where $|S(f, t)|$ is the short-time magnitude spectrum and $\psi(f, t) = \angle S(f, t)$ is the short-time phase spectrum. The signal $s(t)$ is completely characterized by its short-time magnitude and phase spectra.

In this paper, the usefulness of the phase spectrum² is explored in human speech perception. The authors have a longer-term goal of utilising phase spectra in an effort to improve automatic speech recognition (ASR) performance. Although the phase spectrum carries half of the information about the speech signal (as seen from Eq. (2)), ASR systems generally discard the phase spectrum in favour of cepstral features, which are derived purely from the magnitude spectrum [7]. In the ASR framework, speech is processed frame-wise using a temporal window of duration 20-40 ms. If the phase spectrum is to be of any use for ASR applications, it should provide some information about speech intelligibility using small window durations in a human perception experiment.

Liu et al. [3] have conducted such an experiment. They recorded six stop-consonants from 10 speakers in vowel-consonant-vowel context. Using these recordings, they created *magnitude-only* and *phase-only* stimuli. Magnitude-only stimuli were created by analysing the original recordings with a STFT, replacing each frame's phase spectra with random phase values, then reconstructing the speech signal using the overlap-add method. In the case of phase-only stimuli, the original phase of each frame was retained, while the magnitude of each frame was set to unity for all frequency components. The stimuli were created for various window lengths from 16 ms to 512 ms. These were played to subjects, whose task was to identify each as one of the 6 consonants. Their results show that intelligibility of magnitude-only stimuli decreases while the intelligibility of the phase-only stimuli increases as the window duration increases. For small window durations ($T_w < 128$ ms), magnitude-only stimuli is significantly more intelligible than phase-only stimuli (while the opposite is true for larger window lengths). This implies that for small window durations (which are of relevance for ASR applications), the magnitude spectrum contributes much more towards intelligibility than the phase spectrum.

The authors of this paper initially set out to reproduce Liu's results; in doing so, made a number of modifications in Liu's analysis-modification-synthesis procedure. The modifications produce results which are different from Liu's results and more interesting from an ASR applications viewpoint. The first suggested modification is that of the analysis window type. Liu and his collaborators employed a Hamming window for construction of both the

²From here in, the modifier 'short-time' is implied when mentioning the phase spectrum and magnitude spectrum.

magnitude-only and phase-only stimuli. In our experiments, we find that the intelligibility of phase-only stimuli is improved significantly and comparable to that of magnitude-only stimuli when a rectangular window is used. The second suggested modification is the choice of analysis frame shift; Liu et al. used a frame shift of $T_w/2$. As shown by Allen and Rabiner [1], in order to avoid aliasing errors during reconstruction, the STFT sampling period (or frame shift) must be at most $T_w/4$ for a Hamming window. In our work, to be on the safer side, we use a frame shift of $T_w/8$. Our study also differs from Liu's study with respect to the number of consonants used (16 for this study compared to 6 for Liu et al.). The design parameters are discussed in further detail later in this paper. Our results indicate that even for small window durations, the phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum if the analysis-modification-synthesis parameters are properly selected.

In the rest of this paper, we present two experiments (the first of which has been reported earlier [6], however, is presented here for completion). In the first experiment, we compare intelligibility of phase-only stimuli and magnitude-only stimuli constructed at both small and large window durations. In the second experiment, we synthesize different types of phase-only stimuli, all of which are constructed with a small analysis window duration. We test for the intelligibility with a number of combinations of the settings for the remaining design parameters (ie., window type, frame shift and zero-padding) in order to ascertain their respective contribution to intelligibility.

2. HUMAN PERCEPTION EXPERIMENTS

2.1. Experiment 1

In this experiment we compare the intelligibility of magnitude-only and phase-only stimuli using two window types: 1) a rectangular window, and 2) a Hamming window. This comparison is done at a small window duration of 32 ms as well as a large window duration of 1024 ms. We employ a frameshift of $T_w/8$ and zero padding (to reduce aliasing effects).

We record 16 commonly occurring consonants in Australian English in aCa context spoken in a carrier sentence "Hear aCa now". For example, for the consonant /d/, the recorded utterance is "Hear ada now". These 16 consonants in the carrier sentence are recorded for four speakers: two males and two females, providing a total of 64 utterances. The recordings are made in a silent room, sampled at 16 kHz with 16-bit precision.

Each of the recordings are processed through a STFT-based speech analysis-modification-synthesis system (Fig. 1) to retain either only phase information or only magnitude information. In order to construct, for example, an utterance with only phase information, the signal is processed through the STFT analysis using Eq. (1) and the magnitude spectrum is made unity in the modified STFT $\hat{S}(f, t)$; that is,

$$\hat{S}(f, t) = e^{j\psi(f, t)}. \quad (3)$$

This modified STFT is then used to synthesize the signal $\hat{s}(t)$ using the overlap-add method. The synthesized signal $\hat{s}(t)$ contains all of the information about the short-time phase spectra contained in the original signal $s(t)$, but will have no information about the short-time magnitude spectra. We refer to this procedure as the STFT *phase-only synthesis* and the utterances synthesized by this

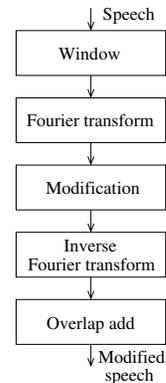


Fig. 1. Speech analysis-modification-synthesis system.

procedure as the *phase-only* utterances. Similarly, for generating *magnitude-only* utterances, we retain each frame's magnitude spectrum and randomise each frame's phase spectrum; that is, the modified STFT is computed as follows:

$$\hat{S}(f, t) = |S(f, t)|e^{j\phi}, \quad (4)$$

where ϕ is a random variable uniformly distributed between 0 and 2π .

In the STFT-based speech analysis-modification-synthesis system of Fig. 1, there are four design issues that must be addressed.

1. **Analysis window type.** This refers to the type of window function $w(t)$ used for computing the STFT (Eq. (1)). A tapered window function (such as Hanning, Hamming or triangular) has been used in earlier studies [3]. Considering these studies have found the phase spectrum to be unimportant at small window durations, a rectangular (non-tapered) window function is investigated in this study in addition to a Hamming window function.
2. **Analysis window duration.** In this experiment, we investigate the importance of phase spectra for two window durations: 1) $T_w = 32$ ms and 2) $T_w = 1024$ ms.
3. **STFT sampling period (frame shift).** In order to avoid aliasing during reconstruction, the STFT must be adequately sampled across the time axis. The STFT sampling period is decided by the window function $w(t)$ used in the analysis. For example, for a Hamming window, the sampling period should be at most $T_w/4$ [1]. To be on the safer side, we have used a sampling period of $T_w/8$. Although the rectangular window can be used with a larger sampling period, we use the same sampling period (i.e., $T_w/8$) to maintain consistency. In this paper, we also refer to the STFT sampling period as the frame shift.
4. **Zero-padding.** For a windowed frame of length N , the Fourier transform is computed using the fast Fourier transform (FFT) algorithm with a FFT size of $2N$ points. This is equivalent to appending N zeros to the end of the N -length frame prior to performing the FFT. The resulting STFT is modified, then inverse Fourier transformed to get a reconstructed signal of length $2N$. Only the first N points are retained, while the last N points are discarded. This is done in order to minimise aliasing effects.

Table 1. Consonant intelligibility (or, identification accuracy) of magnitude-only and phase-only stimuli for a small window duration of 32 ms (with $T_w/8$ frame shift).

Type of stimuli	Intelligibility (in %) for	
	Hamming window	Rect. window
Original	89.9	89.9
Magn. only	84.2	78.1
Phase only	59.8	79.9

Table 2. Consonant intelligibility (or, identification accuracy) of magnitude-only and phase-only stimuli for a large window duration of 1024 ms (with $T_w/8$ frame shift).

Type of stimuli	Intelligibility (in %) for	
	Hamming window	Rect. window
Original	89.9	89.9
Magn. only	14.1	13.3
Phase only	88.0	89.3

As listeners, we use 12 native Australian English speakers with normal hearing, all within the age group of 20-35 years. The group of listeners and the group used for the recordings are mutually exclusive.

The subjects are tested in isolation in a silent room. The re-constructed signals and the original signals are played in random order via earphones at a comfortable listening level. The task is to identify each utterance as one of the 16 consonants. This way, we attain consonant identification (or, intelligibility) accuracy for each subject for different conditions.

The perception tests for this experiment are conducted over two sessions. In the first session, we use a window duration of 32 ms. Results averaged over the 12 subjects are listed in Table 1. In the second session, a window duration of 1024 ms is used and the averaged results are provided in Table 2. The intelligibility of the original recordings is averaged over both sessions.

The following observations can be made from Tables 1 and 2:

1. For the large window duration of 1024 ms, the phase spectrum provides significantly more information than the magnitude spectrum for both the Hamming window function ($F[1, 11] = 2880.57, p < 0.01$) and the rectangular window function ($F[1, 11] = 1582.38, p < 0.01$). This observation is consistent with the results reported earlier in the literature [3, 4, 12].
2. For the small window duration of 32 ms, intelligibility of magnitude-only stimuli is significantly better than the phase-only stimuli when the Hamming window function is used ($F[1, 11] = 17.4, p < 0.01$), but these are comparable when the rectangular window function is used ($F[1, 11] = 2.91, p < 0.01$). Thus, if a rectangular window function is used in the STFT analysis-modification-synthesis system, the phase spectrum carries as much information about the speech signal as the magnitude spectrum, even for small window durations, which are typically used in speech processing applications.
3. For a small window duration of 32 ms, the Hamming window provides better intelligibility than the rectangular window for magnitude-only stimuli ($F[1, 11] = 29.38, p <$

Table 3. Comparison of consonant intelligibility (or, identification accuracy) for the phase-only stimuli used in experiment 3.

Type of Stimuli	Parameter Settings			Phase-only Intelligibility
	Window	Shift	Padding	
A	Ham	$T_w/2$	No	45.3%
B	Rect	$T_w/2$	No	76.6%
C	Rect	$T_w/8$	No	82.8%
D	Rect	$T_w/8$	Yes	85.9%

0.01); while the rectangular window is better than the Hamming window for the construction of phase-only stimuli ($F[1, 11] = 176.30, p < 0.01$).

4. For a small window duration of 32 ms, the best intelligibility results from magnitude-only stimuli (obtained by using a Hamming window) are significantly better than the best results from phase-only stimuli (obtained using a rectangular window) ($F[1, 11] = 17.14, p < 0.01$).

These results can be explained as follows. The multiplication of a speech signal with a window function is equivalent to the convolution of the speech spectrum $S(f)$ with the spectrum $W(f)$ of the window function. The window's magnitude spectrum³ $|W(f)|$ has a big main lobe and a number of side lobes. This causes two problems: 1) frequency resolution problem and 2) spectral leakage problem. The frequency resolution problem is caused by the main lobe of $|W(f)|$. When the main lobe is wider, a larger frequency interval of the speech spectrum gets smoothed and the frequency resolution problem becomes worse. The spectral leakage problem is caused by the sidelobes; the amount of spectral leakage increases with the magnitude of the side lobes. For magnitude-only utterances, we want to preserve the true magnitude spectrum of the speech signal. For the estimation of the magnitude spectrum, frequency resolution as well as spectral leakage are serious problems. Since the Hamming window has a wider main lobe and smaller side lobes in comparison to the rectangular window, the Hamming window provides a better trade-off between frequency resolution and spectral leakage than the rectangular window and, hence, it results in higher intelligibility for the magnitude-only utterances. For the estimation of the phase spectrum, it seems that the side lobes do not cause a serious problem; the smoothing effect caused by the main lobe appears to be more serious [11]. It is because of this that the rectangular window results in better intelligibility than the Hamming window for phase-only utterances.

2.2. Experiment 2

The intelligibility results for phase-only stimuli in experiment 1 are better than previously reported by Liu et al. [3]. This improvement is made by altering a number of parameter settings in the analysis-modification-synthesis framework. In this experiment, we determine the contribution that each analysis-modification-synthesis parameter setting provides towards the intelligibility of signals re-constructed from short-time phase spectra.

The previous experiment demonstrated that it is possible to attain a good intelligibility score for phase-only stimuli with a small

³The window's phase spectrum $\angle W(f)$ is a linear function of frequency and, hence, does not cause a problem in estimating the speech spectrum $S(f)$.

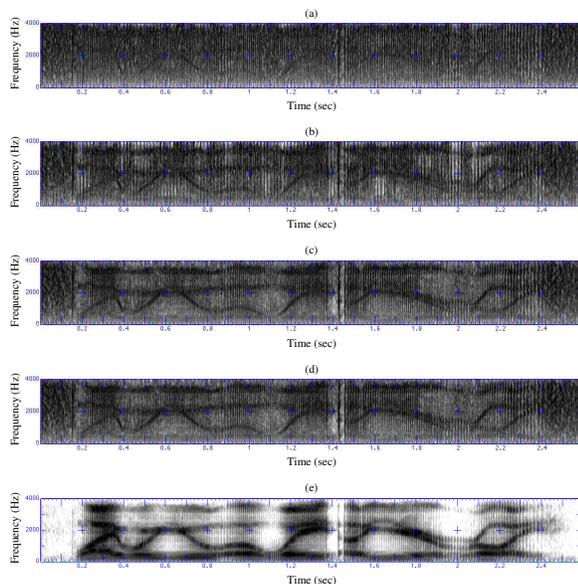


Fig. 2. The spectrograms of phase-only stimuli at an analysis window duration of 32 ms: (a) stimulus type A, (b) stimulus type B, (c) stimulus type C, (d) stimulus type D, and (e) spectrogram of the original speech sentence “Why were you away a year Roy?”. Stimulus construction parameters are given in Table 3.

analysis window duration of 32 ms. Therefore, in this experiment, the window duration is set constant at 32 ms. A number of combinations of settings for the other parameters are tested in order to ascertain their respective contribution to the intelligibility of phase-only stimuli. Table 3 details the parameters used to construct each type of stimuli and the names we will use to refer to them in this experiment.

All audio files are presented to each subject in a single session. The details of the experimental setup are the same as those used in experiment 1.

The intelligibility scores are provided in Table 3. From these scores, we can conclude that the major contribution to overall intelligibility comes from the use of the rectangular window (stimulus type B). The reason for this improvement is as discussed in experiment 1. It is not surprising to see further improvement from the decrease in frame shift (stimuli type C) and the use of zero-padding (stimuli type D), due to their roles in reducing aliasing effects.

Fig. 2 presents the spectrogram of a sentence of speech with its reconstructed phase-only stimuli A, B, C and D. The increasing clarity of the formant tracks in these spectrograms, from A through D, is indicative of the corresponding trend in the intelligibility of these stimuli.

3. CONCLUSION

The authors recently conducted a human perception experiment [6] to measure the intelligibility of speech stimuli synthesised either from short-time magnitude spectra or short-time phase spectra. In this paper, we review the results of that experiment (experiment 1) which demonstrate that even for small window durations, phase spectra can contribute to speech intelligibility as much as magni-

tude spectra if the analysis-modification-synthesis parameters are properly selected. An additional experiment is performed (experiment 2), the results of which indicate that the parameter selection that contributes most to the intelligibility of the phase-only stimuli is that of a rectangular analysis window, as opposed to a Hamming window (which is generally used in speech analysis).

Since the speech processing in ASR applications is done frame-wise over small analysis window durations (20-40 ms), it is logical to investigate the use of phase spectrum to extract features for these applications. Some preliminary results have already been reported earlier [5], which show the usefulness of phase spectrum for ASR. More detailed results will be reported in the future.

4. ACKNOWLEDGEMENT

The authors wish to thank the volunteers who took part in the subjective listening tests.

5. REFERENCES

- [1] J.B. Allen and L.R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis” Proc. IEEE, Vol. 65, No. 11, pp. 1558-1564, 1977
- [2] D.W. Griffin and J.S. Lim, “Signal estimation from modified short-time Fourier transform”, IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, pp. 236-243, 1984
- [3] L. Liu, J. He and G. Palm, “Effects of phase on the perception of intervocalic stop consonants”, Speech Communication, Vol. 22, pp. 403-417, 1997.
- [4] A.V. Oppenheim and J.S. Lim, “The importance of phase in signals” Proc. IEEE, Vol. 69, pp. 529-541, 1981.
- [5] K.K. Paliwal, “Usefulness of phase in speech processing”, Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan, pp. 1-6, Feb. 2003
- [6] K.K. Paliwal and L. Alsteris, “Usefulness of phase spectrum in human speech perception”, Proc. Eurospeech, Geneva, Switzerland, pp. 2117-2120, Sept. 2003
- [7] J.W. Picone, “Signal Modeling techniques in speech recognition”, Proc. IEEE, Vol. 81, No. 9, pp. 1215-1247, 1993.
- [8] M.R. Portnoff “Short-time Fourier analysis of sampled speech” IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364-373, 1981.
- [9] T.F. Quatieri, *Discrete-time speech signal processing*, Prentice Hall, Upper Saddle River, NJ, 2002.
- [10] L.R. Rabiner and R.W. Schafer, *Discrete-time speech signal processing, principles and practice*, Prentice Hall, Englewood Cliffs, NJ, 1978.
- [11] N.S. Reddy and M.N.S. Swamy, “Derivative of phase spectrum of truncated autoregressive signals”, IEEE Trans. Circuits and Systems, Vol. CAS-32, pp. 616-618, 1985.
- [12] M.R. Schroeder, “Models of hearing”, Proc. IEEE, Vol. 63, pp. 1332-1350, 1975.