Intelligibility of Speech from Phase Spectrum

Leigh D. Alsteris and Kuldip K. Paliwal

***Abstract*— The short-time Fourier transform of a speech signal has two components: the magnitude spectrum and the phase spectrum. In this paper, the relative importance of short-time magnitude and phase spectra for speech perception is investigated. Human perception experiments are conducted to measure intelligibility of speech stimuli synthesized either from magnitude spectra or phase spectra. It is traditionally believed that the magnitude spectrum plays a dominant role for small window durations (20-40 ms); while the phase spectrum is more important for large window durations ($>$ 1 s). It is shown in this paper 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.**

I. 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], [6], [7], [8]. 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 window[1]. The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

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

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:

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

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, we describe an experiment to evaluate the importance of short-time phase spectra and short-time magnitude spectra in human speech perception.

II. Human Perception Experiment

We compare the intelligibility of magnitude-only and phase-only stimuli using 2 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.

The authors are with the Signal Processing Laboratory, School of Microelectronic Engineering, Faculty of Engineering and Information Technology, Griffi th University, Brisbane, QLD 4111 Australia (E-mail: l.alsteris@griffi th.edu.au; k.paliwal@griffi th.edu.au)

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

TABLE I
CONSONANTS USED IN PERCEPTION TESTING.

a-Consonant-a	As in
aba	bat
ada	deep
afa	five
aga	go
aka	kick
ama	mum
ana	noon
apa	pea
asa	so
ata	tea
ava	vice
aza	zebra
adha	then
asha	show
atha	thing
azha	measure

A. Recordings

We record 16 commonly occurring consonants in Australian English in aCa context (Table I) 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 4 speakers: 2 males and 2 females, providing a total of 64 utterances. The recordings are made in a silent room with a SONY ECM-MS907 microphone (90 degree position). The signals are sampled at 16 kHz with 16-bit precision. The duration of each recorded signal is approximately 3 seconds[2].

B. Stimuli

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,

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

This modified STFT is then used to synthesize the signal $\hat{s}(t)$ using the overlap-add method [1]. 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 procedure as the *phase-only* utterances. Similarly, for generating *magnitude-only* utterances,

[2]This time is inclusive of leading and trailing silence periods.

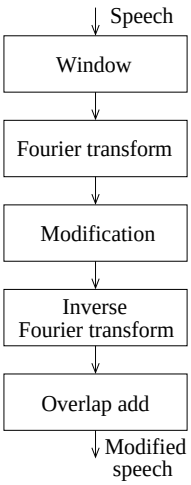

Fig. 1. Speech analysis-modification-synthesis system.

we retain each frame's magnitude spectrum and randomise each frame's phase spectrum; that is, the modified STFT is computed as follows:

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

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 4 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 2 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. Zero-padding is used in the construction of all stimuli in this study, unless otherwise stated.

There are 8 types of stimuli for this experiment. The description of each type is provided in Table II.

TABLE II
STIMULI USED FOR PERCEPTION TESTING (WITH FRAME SHIFT OF $T_w/8$).

Type of stimuli	Retained Spectrum	Window Type	Window Duration (ms)
A	Magnitude	Hamming	32
B	Magnitude	Rectangular	32
C	Phase	Hamming	32
D	Phase	Rectangular	32
E	Magnitude	Hamming	1024
F	Magnitude	Rectangular	1024
G	Phase	Hamming	1024
H	Phase	Rectangular	1024

C. Subjects

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.

D. Procedure

The perception tests for this experiment are conducted over 2 sessions. In the first session, stimuli types A, B, C, and D are presented ($T_w = 32$ ms). In the second session, stimuli types E, F, G, and H are presented ($T_w = 1024$ ms).

The subjects are tested in isolation in a silent room. The reconstructed signals and the original signals (a total of 320 for each session) are played in random order via SONY MDR-V5000DF 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. In both sessions, the subjects are first familiarised with the task through a short practice test. Session 1 results are provided in Table III and session 2 results are provided in Table IV. Results are averaged over the 12 subjects. The intelligibility of the original recordings is averaged over both sessions.

Responses are collected through software. The software displays the 16 aCa possibilities as well as an extra option for a null response. Participants are instructed to only choose the null response when they have no clue as to what the consonant may be. Responses are input via the keyboard in the form of numbers (1-17). Each audio file is presented once. No feedback is provided.

E. Results and discussion

The following observations can be made from Tables III and IV:

TABLE III

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 IV

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

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], [9].
2) The difference in intelligibility between magnitude-only stimuli constructed with a Hamming window and magnitude-only stimuli constructed with a rectangular window at a large window duration of 1024 ms is insignificant ($F[1,11] = 0.63$, $p < 0.01$). The same can also be said for phase-only signals constructed with either window type at the large window duration ($F[1,11] = 1.18$, $p < 0.01$).
3) 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.
4) For a small window duration of 32 ms, the Hamming window provides better intelligibility than the rectangular window for magnitude-only stimuli ($F[1,11] = 29.38$, $p < 0.01$); while the rectangular window is better than the Hamming window for the construction of phase-only stimuli ($F[1,11] = 176.30$, $p < 0.01$).
5) 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 spectrum[3] $|W(f)|$ 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. It is because of this that the rectangular window results in better intelligibility than the Hamming window for phase-only utterances.

III. **Conclusion**

In this paper, the relative importance of short-time magnitude and phase spectra on speech perception is investigated. The results of our experiment show that the phase spectrum not only contributes to speech intelligibility at large analysis window durations ($T_w = 1024$ ms), but also at small analysis window durations ($T_w = 32$ ms), if the analysis-modification-synthesis parameters are properly selected. 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.

IV. **Acknowledgment**

This work was partly supported by ARC (Discovery) grant (No. DP0209283). The authors also wish to thank the volunteers who took part in the subjective listening tests reported in this paper.

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.

[3]The 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 $S(f)$.

[2] D.W. Griffi n and J.S. Lim, "Signal estimation from modifi ed 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] M.R. Portnoff "Short-time Fourier analysis of sampled speech" IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364-373, 1981.
[7] T.F. Quatieri, *Discrete-time speech signal processing*, Prentice Hall, Upper Saddle River, NJ, 2002.
[8] L.R. Rabiner and R.W. Schafer, *Discrete-time speech signal processing, principles and practice*, Prentice Hall, Englewood Cliffs, NJ, 1978.
[9] M.R. Schroeder, "Models of hearing", Proc. IEEE, Vol. 63, pp. 1332-1350, 1975.