Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 153–170
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

On the usefulness of STFT phase spectrum in human listening tests ☆

Kuldip K. Paliwal *, Leigh D. Alsteris

School of Microelectronic Engineering, Griffith University, Nathan campus, 4111 QLD, Australia

Received 4 December 2003; received in revised form 6 May 2004; accepted 15 August 2004

Abstract

The short-time Fourier transform (STFT) 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.

Keywords: Short-time Fourier transform; Phase spectrum; Magnitude spectrum; Speech perception; Overlap-add procedure; Automatic speech recognition

1. Introduction

In this paper, the usefulness of the phase spectrum [1] is explored in human speech perception. [2] The authors have a long-term goal of utilising phase spectra in an effort to improve automatic speech recognition (ASR) performance. It is common practice in ASR to discard the phase

☆ Audio files at http://maxwell.me.gu.edu.au/spl/research/phase/project.htm.

* Corresponding author. Tel.: +61 7 3875 6536; fax: +61 7 3875 5384.
E-mail addresses: k.paliwal@griffith.edu.au (K.K. Paliwal), l.alsteris@griffith.edu.au (L.D. Alsteris).

[1] Throughout this paper, the modifier 'short-time' (i.e., finite-time) is implied when mentioning the phase spectrum and magnitude spectrum.

[2] There is a large amount of literature available on the topic of the perception of phase in speech, dating back to Ohm's study (Ohm, 1843) in 1943. In our work, we only refer to a small percentage of this literature. In addition to the papers referenced throughout this text, the following selected papers may be of interest to the reader: Goldstein (1967), von Helmholtz (1912), Kim (2000), Patterson (1987), Plomp and Steeneken (1969), Pobloth and Kleijn (1999), Schroeder (1959).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.08.001

spectrum in favour of features that are derived purely from the magnitude spectrum [3] (Picone, 1993). 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 (20–40 ms) in a human perception experiment.

A few studies have been reported in the literature which discuss whether the phase spectrum provides any information which can contribute to intelligibility for human speech recognition (HSR). Schroeder (1975), and Oppenheim and Lim (1981) performed some informal perception experiments, concluding that the phase spectrum is important for intelligibility when the window duration of the short-time Fourier transform (STFT) is large ($T_w > 1$ s), while it seems to convey negligible intelligibility at small window durations (20–40 ms).

Liu et al. (1997) have recently investigated the intelligibility of phase spectra through a more formal human speech perception study. 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 six consonants. Their results (Fig. 1) show that intelligibility of magnitude-only stimuli decreases while the intelligibility of the phase-only stimuli increases as the window duration increases.

[3] There are other speech processing applications where spectral phase information is overlooked. For example, in speech enhancement it is common practice to modify the magnitude spectrum and keep the corrupt phase spectrum (Lim and Oppenheim, 1979; Wang and Lim, 1982).

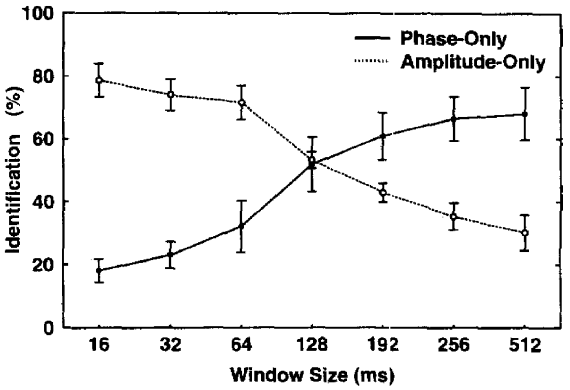

Fig. 1. Average identifcation performance and standard deviation as a function of window size for phase-only and magnitude-only stimuli, from the paper by Liu et al. (after Liu et al. (1997)).

For small window durations ($T_w < 128$ ms), magnitude-only stimuli are 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 (see Fig. 2). The modifications produce results which are different from Liu's results and more interesting from an ASR application's 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 magnitude-only and phase-only stimuli. In our experiments, we find that the intelligibility of phase-only stimuli is improved significantly and becomes 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 (1977), 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 this paper, to be on the safer side, we use a frame shift of $T_w/8$. Our study also differs from Liu's study with