Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 48 (2006) 727–736
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Further intelligibility results from human listening tests using the short-time phase spectrum

Leigh D. Alsteris *, Kuldip K. Paliwal *

School of Microelectronic Engineering, Griffith University, Nathan Campus, Brisbane QLD 4111, Australia

Received 27 January 2005; received in revised form 16 August 2005; accepted 18 October 2005

Abstract

State-of-the-art automatic speech recognition systems (ASRs) use only the short-time magnitude spectrum for feature extraction; the short-time phase spectrum is generally ignored in these systems. Results from our recent human listening tests indicate that the short-time phase spectrum can significantly contribute to speech intelligibility over small window durations (i.e., 20–40 ms). This is an interesting result, indicating the possible usefulness of the short-time phase spectrum for ASR, which commonly employs small window durations of 20–40 ms for spectral analysis. In this paper, we continue our investigation of the short-time phase spectrum. We explore the use of partial short-time phase spectrum information, in the absence of all the short-time magnitude spectrum information, for intelligible signal reconstruction. We create two types of stimuli; one in which its frequency-derivative (i.e., group delay function, GDF) is preserved and another in which its time-derivative (i.e., instantaneous frequency distribution, IFD) is preserved. We do this to determine the contribution that each of these derivatives provides toward intelligibility. Reconstructing stimuli from knowledge of only the GDF or only the IFD results in poor intelligibility. However, when we create stimuli using knowledge of both the GDF and the IFD, reasonable intelligibility is obtained. In light of these results, we conclude that both the GDF and IFD components of the short-time phase spectrum are needed to reconstruct an intelligible signal. In addition, we also perform some experiments to quantify the intelligibility of stimuli reconstructed from the short-time phase and magnitude spectra of noisy speech. The intelligibility of stimuli constructed from either the short-time magnitude spectrum or the short-time phase spectrum degrades at a similar rate under increasing noise levels. The intelligibility of the original signals under noisy conditions also degrades with increased noise, but in all cases the intelligibility is superior to that provided by the stimuli constructed from the separate short-time components. Therefore, we argue that knowledge of both short-time magnitude and phase spectrum information results in superior human speech recognition performance.

Keywords: Short-time Fourier transform; Phase spectrum; Magnitude spectrum; Speech perception; Overlap-add procedure; Automatic speech recognition; Feature extraction; Group delay function; Instantaneous frequency distribution

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

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.10.005

1. Introduction

This paper is motivated by our desire to use the short-time phase spectrum[1] for automatic speech recognition (ASR). The idea that the phase spectrum may be useful for ASR is not commonly accepted. The argument against the use of the phase spectrum for ASR can be attributed to some well-known perception experiments conducted by several authors (Schroeder, 1975; Oppenheim and Lim, 1981; Liu et al., 1997), the results of which strongly suggest that the phase spectrum lacks intelligibility information at small window durations used in short-time Fourier analysis (20–40 ms). In addition, from a signal processing viewpoint, the phase spectrum is difficult to interpret due to phase wrapping and other problems (Murthy et al., 1989; Duncan et al., 1989; Yegnanarayana and Murthy, 1992). In its raw form, it is not possible to extract features for ASR from the phase spectrum.

Schroeder (1975) and Oppenheim and Lim (1981) concluded through their informal perception experiments that the phase spectrum is important for intelligibility when the window duration of the short-time Fourier transform is large (>1 s), while it seems to convey negligible intelligibility at small window durations (20–40 ms). Liu et al. (1997) investigated the intelligibility of the phase spectrum through a more formal human speech perception study. Their results, like the studies before it (Schroeder, 1975; Oppenheim and Lim, 1981), show that the phase spectrum contributes little intelligibility at small window durations (20–40 ms). However, we have recently suggested a number of modifications to Liu's experimental procedure,[2] producing results which are different from Liu's results and more interesting from an ASR viewpoint (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005). Our results indicate that, even for small window durations (20–40 ms), the phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum. This provides the motivation to investigate the use of the phase spectrum to derive an alternative, or supplemental, feature representation for ASR (which generally uses small window durations of 20–40 ms for spectral analysis).

Unlike the magnitude spectrum, resonances can not be observed in the phase spectrum. A physical connection between the phase spectrum and the structure of the vocal apparatus is not apparent. The phase spectrum values suffer from wrapping and other problems (Murthy et al., 1989; Duncan et al., 1989; Yegnanarayana and Murthy, 1992). It is therefore necessary that the phase spectrum be transformed into a more tangible representation, which does not suffer from the aforesaid problems. The short-time phase spectrum has two independent variables: frequency and time. Thus, while there may be many ways to represent the information present in the phase spectrum, two representations that first come to mind are those that can be obtained either by taking its frequency-derivative or its time-derivative. Signal processing issues related to the frequency-derivative of the phase spectrum, commonly known as the group delay function[3] (GDF), have been discussed extensively in the literature (Oppenheim and Schafer, 1975; Yegnanarayana et al., 1984; Yegnanarayana and Murthy, 1992; Bozkurt et al., 2004). Also, the GDF has been used in a number of speech processing applications; such as formant extraction (Murthy et al., 1989; Duncan et al., 1989), pitch extraction (Smits and Yegnanarayana, 1995; Satyanarayana and Yegnanarayana, 1999), spectrum estimation (Yegnanarayana and Murthy, 1992), minimum-phase signal reconstruction (Yegnanarayana et al., 1984), speech segmentation (Prasad et al., 2004), and speaker identification (Hegde et al., 2004a). Murthy and her colleagues have recently used GDF-based features for ASR (Murthy and Gadde, 2003; Hegde et al., 2004b,c; Alsteris and Paliwal, 2005). The time-derivative of the phase spectrum, most often referred to as the instantaneous frequency distribution (IFD), has been used in the past for pitch extraction (Abe et al., 1995; Charpentier, 1986; Nakatani et al., 2003) and formant extraction (Potamianos and Maragos, 1996; Friedman, 1985). Potamianos and Maragos (2001), Dimitriadis and Maragos (2003), Paliwal and Atal (2003) and Wang et al. (2003) have recently investigated IFD-based features for ASR.

As mentioned earlier, we have recently established that significant intelligibility is provided by

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

[2] For a detailed comparison of Liu's procedure and our procedure, refer to (Paliwal and Alsteris, 2005).

[3] Also referred to as group delay spectrum.

the phase spectrum at small window durations of 20–40 ms. This paper continues our investigation of the phase spectrum, serving to satisfy two objectives: (1) We explore the use of partial phase spectrum information, in the absence of all the magnitude spectrum information, for intelligible signal reconstruction. Using the same analysis–modification–synthesis procedure as we did in (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005), we will create two types of stimuli; one in which the phase spectrum frequency-derivative (i.e., GDF) is preserved and another in which the phase spectrum time-derivative (i.e., IFD) is preserved. We do this to determine the contribution that each component of the phase spectrum provides toward intelligibility. If we obtain significant intelligibility from either component, then it would be wise to investigate the component's potential as a basis for an ASR representation. Conversely, if we obtain poor intelligibility, perhaps we should consider other phase spectrum representations (other than the GDF and IFD). (2) Ultimately, the speech community is searching for ASR features that are robust to noise. Hence, in addition to the first objective, we also perform some experiments to quantify the intelligibility of stimuli reconstructed from the phase spectrum and the magnitude spectrum of noisy speech.[4]

The paper outline is as follows: In Section 2, we detail the analysis–modification–synthesis procedure (reported earlier in Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005) used to create the stimuli for the human perception experiments. In Section 3, we describe three experiments. The first experiment has been reported earlier (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005); however, it is presented here for the sake of completeness. In this experiment, we determine the amount of intelligibility provided by the phase spectrum and the magnitude spectrum at small window durations (32 ms). In the second experiment, we synthesize stimuli from knowledge of only partial phase spectrum information (i.e., GDF or IFD), in order to determine the contribution that each component of the phase spectrum provides toward intelligibility. In the third experiment, we determine the intelligibility of the phase spectrum and the magnitude spectrum under noisy conditions.

2. Analysis–modification–synthesis procedure (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005)

Although speech is a non-stationary signal, it is generally assumed to be quasi-stationary and, therefore, can be processed through a short-time Fourier analysis (Allen and Rabiner, 1977; Crochiere, 1980; Griffin and Lim, 1984; Portnoff, 1981; Quatieri, 2002). The short-time Fourier transform (STFT) of a speech signal, $x(t)$, is given by

$$X(\omega,t)=\int_{-\infty}^{\infty}x(\tau)w(t-\tau)\mathrm{e}^{-\mathrm{j}2\pi\omega t}\,\mathrm{d}\tau, \tag{1}$$

where $w(t)$ is a window function. In ASR, the Hamming window function (Nuttall, 1981) is typically used and its duration is normally 20–40 ms.

We can decompose $X(\omega,t)$ as follows:

$$X(\omega,t)=|X(\omega,t)|\mathrm{e}^{\mathrm{j}\phi(\omega,t)}, \tag{2}$$

where $|X(\omega,t)|$ is the short-time magnitude spectrum and $\phi(\omega,t)=\angle X(\omega,t)$ is the short-time phase spectrum. The signal $x(t)$ is completely characterized by its short-time magnitude and phase spectra.

The aim of the experiments in Section 3 is to determine the contribution that the phase spectrum (and partial phase spectrum) and the magnitude spectrum provide toward speech intelligibility. Accordingly, stimuli are created either from knowledge of the phase spectrum values or the magnitude spectrum values. In order to construct, for example, an utterance from the phase spectrum values, the signal is processed through the STFT analysis using Eq. (1) and the magnitude spectrum is made unity in the modified STFT, $\widehat{X}(\omega,t)$; that is,

$$\widehat{X}(\omega,t)=\mathrm{e}^{\mathrm{j}\phi(\omega,t)}. \tag{3}$$

This modified STFT is then used to synthesize the signal $\hat{x}(t)$ using the overlap-add method (Allen and Rabiner, 1977; Griffin and Lim, 1984). The synthesized signal, $\hat{x}(t)$, contains all the information about the short-time phase spectra contained in the original signal $x(t)$, 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, we retain each frame's magnitude spectrum and randomise each frame's

[4] Thus far (see Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005), all of our perception experiments have been performed with stimuli created from clean speech.

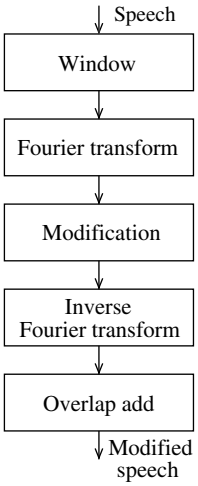

Fig. 1. Speech analysis–modification–synthesis system, used to create stimuli for our human perception experiments.

phase spectrum; that is, the modified STFT is computed as follows:

$$\widehat{X}(\omega,t) = |X(\omega,t)|e^{j\theta(\omega,t)}, \quad (4)$$

where $\theta(\omega,t)$ is a random variable uniformly distributed between 0 and 2π.

In the STFT-based analysis–modification–synthesis system of Fig. 1, there are a number of design issues that must be addressed. First, what type of window function, $w(t)$, should be used for computing the STFT (Eq. (1))? In Experiment 1, we investigate the use of both tapered windows (specifically, a Hamming window and a triangular window) and a rectangular window. Second, what should be the duration, T_w, of the window function? In our work, we investigate a small duration of 32 ms, since this is similar to the size used for ASR. Third, how often should we compute the STFT; i.e., how often should we sample the STFT across the time axis in order to avoid aliasing during reconstruction? 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 $T_w/4$ (Allen and Rabiner, 1977). To be on the safer side, we have used a sampling period of $T_w/8$. Although the rectangular and triangular windows can be used with a larger sampling period, we use the same sampling period (i.e., $T_w/8$) to maintain consistency.[5] The last design issue to consider is that of zero-padding. For a windowed frame of N samples (where N is a power of 2), the DFT 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. An inverse FFT of the modified STFT results in 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.

[5] We also refer to the STFT sampling period as the frame shift.

3. Human perception experiments (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005)

3.1. Experiment 1

In this experiment, we create phase-only stimuli and magnitude-only stimuli in accordance with the analysis–modification–synthesis procedure detailed in Section 2. We compare the intelligibility of both types of stimuli using rectangular, triangular, and Hamming analysis windows.[6]

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, totaling 64 recordings. The recordings are sampled at 16 kHz (16-bit). Each recording is processed, as described in Section 2, to retain either only the phase spectrum information or only the magnitude spectrum information.

As listeners, we use 12 native Australian English speakers with normal hearing, all within the age group of 20–35 years. The reconstructed signals and the original signals are played in random order to each listener. The task is to identify each utterance as one of the 16 consonants. The results are summarised in Table 1. Note that there are 768 tokens used for the evaluation of each stimuli type (i.e., four speakers, each speaking 16 consonants, all of which are presented to 12 listeners).

[6] Experiment 1 has been mentioned in a series of our papers (Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005) as it provides the foundation and motivation for continued experimentation. Therefore, once again, we provide a summary of the experimental setup and results for the convenience of the reader. For the most comprehensive description of this experiment, refer to (Paliwal and Alsteris, 2005).

Table 1
Result for Experiment 1 (Paliwal and Alsteris, 2005)

Type of stimuli	Analysis window type (%)		
	Hamming	Triangle	Rectangle
Original	89.9	89.9	89.9
Magnitude-only	84.2	84.0	78.1
Phase-only	59.8	53.8	79.9

Average consonant intelligibility of phase-only and magnitude-only stimuli (a frame duration of 32 ms is used in the STFT analysis). Results for three analysis window types are given.

It is observed that when a rectangular analysis window is used (at an analysis duration of 32 ms) in the analysis–modification–synthesis system, the phase spectrum conveys a large amount of intelligibility (79.9%). Note that in previous studies (Schroeder, 1975; Oppenheim and Lim, 1981; Liu et al., 1997), results were produced with a tapered window. The intelligibility conveyed by the magnitude spectrum is maximised when a Hamming window is used (intelligibility of 84.2%). The triangular window provides similar results (intelligibility of 84.0%).[7] Accordingly, in the following experiments, the phase-only stimuli are constructed with a rectangular analysis window and the magnitude-only stimuli are constructed with a Hamming analysis window. These results have been analysed in detail in (Paliwal and Alsteris, 2005). A discussion of why the rectangular window may provide better intelligibility than a tapered window for phase-only stimuli is provided in Appendix A.

3.2. Experiment 2

In this section, we explore the use of partial phase spectrum information for intelligible signal reconstruction. In addition to the phase-only stimuli, we create the following types of stimuli from the original 64 utterances, using the analysis–modification–synthesis framework described in Section 2 (using a rectangular analysis window of duration 32 ms):

(1) *IFD-only stimuli:* Take the phase spectrum of each short-time section and randomise it across frequency, such that $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$ is preserved. In other words, add the same random sequence (across frequency) to the phase spectrum values of each frame. The resulting phase spectra are used in place of the original phase spectra in the reconstruction algorithm (and magnitude spectra are set to unity).

(2) *GDF-only stimuli:* In a similar vein, take the phase spectrum of each short-time section and randomise it across time, such that $\mathrm{d}\phi(\omega,t)/\mathrm{d}\omega$ is preserved. That is, generate a random sequence whose length is equal to the number of frames in the utterance, then add this same sequence to the time-trajectory of the phase spectrum values for each DFT bin. Reconstruction is performed with the resulting phase spectra (and magnitude spectra are set to unity).

(3) *IFD + GDF stimuli:* We reconstruct a signal from the knowledge of both $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$ and $\mathrm{d}\phi(\omega,t)/\mathrm{d}\omega$. In order to do this, we must first reconstruct the phase spectra from these known quantities. Notice that the first-segment phase spectrum can only be reconstructed to within a time-shift of the original first-segment phase spectrum, since all we know about it is $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$. The remaining segments are reconstructed in relation to this segment. Consequently, we cannot recover the original phase spectra.[8] Reconstruction is performed with the altered phase spectra (and magnitude spectra are set to unity).

In this experiment we employ five listeners (a subset of four listeners[9] from the 12 used in Experiment 1 and one new listener). The reconstructed signals and the original signals are played in random order to each listener. The average consonant identification scores[10] are given in Table 2. Note that

[7] While many other observations about this data can be made, we only discuss what is relevant in the context of this paper. For a complete discussion of these results and significance figures, see (Paliwal and Alsteris, 2005).

[8] The raw phase spectrum values are only meaningful in the context of a fixed-time reference. All that we have lost in this reconstructed signal is the original fixed-time reference. Time referencing is now in relation to the phase spectrum values of the first frame (i.e., we still have a time reference, but it is different to that of the original phase spectra values).

[9] For reference, the average Experiment 1 intelligibility scores for the four listener subset were: original intelligibility was 91.8%, and phase-only (rectangular window) intelligibility was 85.5%.

[10] The intelligibility of the original signals and the phase-only stimuli are both higher than that reported in the Experiment 1. This can most likely be attributed to the following two reasons: (1) these results are based on a subset of listeners used in Experiment 1 (plus one additional listener), and (2) this experiment was conducted at a different time and location than Experiment 1. Regardless, the absolute intelligibility scores are not that important; it is the relative intelligibility that is more interesting.

Table 2
Results for Experiment 2

Type of stimuli	Intelligibility score (%)
Original	95.3
IFD-only	50.9
GDF-only	53.8
IFD + GDF-only	85.6
Phase-only	86.9

Average consonant intelligibility of stimuli constructed from partial phase spectrum information (rectangular analysis window of duration 32 ms used in the STFT analysis).

there are 320 tokens used for the evaluation of each stimuli type (i.e., four speakers, each speaking 16 consonants, all of which are presented to five listeners).

Reconstructing stimuli from knowledge of only the IFD or the GDF results in poor intelligibility: the intelligibility of the stimuli reconstructed from knowledge of only the IFD is 50.94% and the intelligibility of the stimuli reconstructed from knowledge of only the GDF is 53.75%. However, when we create stimuli using knowledge of both the IFD and the GDF, intelligibility on par with the stimuli reconstructed from the original phase spectra is achieved: the intelligibility of the stimuli reconstructed from knowledge of both the IFD and GDF is 85.63% and the intelligibility of the stimuli reconstructed from knowledge of the original phase spectra is 86.88%. The results imply that both IFD and GDF are required for good intelligibility from the phase spectrum. Furthermore, the intelligibility score of the original signals is by far the best (95.31%). That is, all of the phase spectrum and the magnitude spectrum information must be retained for superior intelligibility. This will be addressed further in Experiment 3.

3.3. Experiment 3

This experiment serves to quantify the intelligibility provided by the phase spectrum and the magnitude spectrum components of the STFT, under noisy conditions. In accordance with the results of Section 2, we use a rectangular analysis window to construct the phase-only stimuli and a Hamming analysis window to construct the magnitude-only stimuli. Once again, the duration of the analysis window is 32 ms (which complies with the standard frame sizes of 20–40 ms used in ASR). This time, however, the 64 original utterances are contaminated with white noise over several signal-to-noise ratios (SNRs) of -10 dB, 0 dB, 10 dB, 20 dB and ∞ dB (i.e., no noise added). We employ listeners (one new listener and a subset of two listeners[11] from the 12 used in Experiment 1). The reconstructed signals and the noisy original signals are played in random order to each listener. The average consonant identification scores are plotted in Fig. 2. Note that there are 192 tokens used for the evaluation of each stimuli type (i.e., four speakers, each speaking 16 consonants, all of which are presented to three listeners).

The results indicate that the intelligibility of both the phase-only stimuli and the magnitude-only stimuli degrade at a similar rate under decreasing SNR value. While the intelligibility provided by the original signals also degrades at a similar rate, the intelligibility is consistently better than that provided by the phase-only stimuli and the magnitude-only stimuli. It is particularly interesting to see that the intelligibility provided by the original signals is far better than that provided by the magnitude-only stimuli. This result seems to be at odds with the common practice in ASR; which is to discard the phase spectrum in favour of features that are derived only from the magnitude spectrum. Should ASR features also encapsulate information about the phase spectrum? According to these perception results, robustness in human speech recognition requires that both the magnitude spectrum and the phase spectrum be retained (where a frame duration of 32 ms is used in the STFT analysis). Thus, a feature set that represents information from both the magnitude spectrum and the phase spectrum may result in improved ASR performance.

4. Summary

The results from our previous study (Paliwal and Alsteris, 2005) demonstrate that significant intelligibility can be obtained from the short-time phase spectrum of speech at small analysis window durations of 20–40 ms. These results may have positive implications for ASR, which commonly employs analysis window durations of 20–40 ms for spectral analysis.

Experiment 1 was a reproduction of the pertinent details and results of an experiment previously

[11] For reference, the average Experiment 1 intelligibility scores for the two listener subset were: original intelligibility was 92.2%, and phase-only (rectangular window) intelligibility was 84.4%.

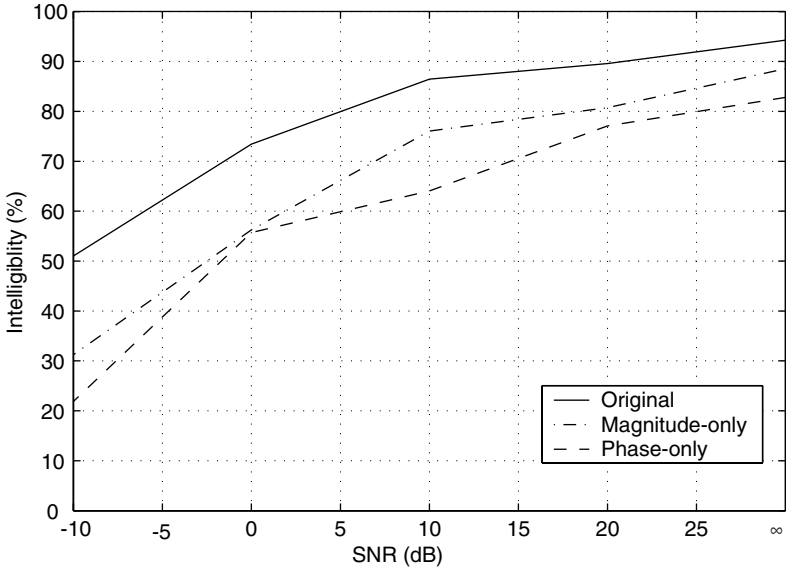

Fig. 2. Results for Experiment 3. Average consonant intelligibility of phase-only and magnitude-only stimuli constructed from white-noise contaminated speech over several SNRs (phase-only and magnitude-only stimuli are constructed with a rectangular and Hamming analysis window respectively, of duration 32 ms). Average intelligibility scores for the original (noisy) speech are also provided.

described in (Paliwal and Alsteris, 2005). In this experiment, we showed that when a rectangular analysis window is used (at an analysis duration of 32 ms) in the analysis–modification–synthesis system, the phase spectrum conveys a large amount of intelligibility.

In Experiment 2, we synthesized stimuli from knowledge of only the time-derivative of the phase spectrum information or only the frequency-derivative of the phase spectrum (i.e., IFD or GDF), in order to determine the contribution that each component provides toward intelligibility. Reconstructing stimuli from knowledge of only the IFD or only the GDF results in poor intelligibility. However, when we create stimuli using knowledge of both the IFD and the GDF, reasonable intelligibility is obtained. The results imply that both the IFD and GDF are required for good intelligibility from the phase spectrum. Note that, although the IFD and GDF are derived from the phase spectrum, neither of these quantities are fully representative of the information in the phase spectrum. The results of this experiment suggest that a possible avenue of future research could be to derive a feature representation from both the IFD and GDF information; such a representation would more faithfully describe the information in the phase spectrum and perhaps provide improved recognition performance over features based on the individual components.

In Experiment 3, we determined the intelligibility of the phase spectrum and the magnitude spectrum under noisy conditions. We observed that the intelligibility provided by both phase-only and magnitude-only stimuli degrades at a similar rate under decreasing SNR value. That is, one component is no more robust than the other. However, while the intelligibility provided by the original signals also degrades at a similar rate, it is consistently better than the intelligibility provided by either the phase-only or magnitude-only stimuli. Therefore, we argue that superior human speech recognition requires that both the magnitude spectrum and the phase spectrum be retained (where a frame duration of 32 ms is used in the STFT analysis). This result does not agree with the common practice in ASR; which is to discard the phase spectrum in favour of features that are derived only from the magnitude spectrum. A feature set that represents information from both the magnitude spectrum and the phase spectrum may result in improved ASR performance.[12]

[12] One of the reviewers of this paper mentioned that the phase spectrum may carry intelligibility information only up to 1 kHz. The reviewer has suggested that we perform additional experimentation in order to explore this phenomenon. We provide preliminary results in Appendix B.

Appendix A. A qualitative discussion on the effect of window shape for construction of phase-only and magnitude-only stimuli

The multiplication of a speech signal with a window function is equivalent to the convolution of the speech spectrum $S(\omega)$ with the spectrum $W(\omega)$ of the window function (we ignore the time dependency in this discussion). The window's magnitude spectrum[13] $|W(\omega)|$ 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 $|W(\omega)|$. 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. Reddy and Swamy (1985) have also recommended the use of a rectangular window function in the computation of the group delay spectrum, which is a frequency-derivative of the phase spectrum.

Appendix B. Intelligibility provided by bandlimited magnitude-only and phase-only stimuli

We provide the preliminary results of an experiment in which we attempt to determine the intelligibility of bandlimited magnitude-only and phase-only stimuli.

According to Delgutte (1996), "The phase-locking of auditory nerve fiber responses to sinusoidal stimuli falls off rapidly for frequencies above 1 kHz, until it becomes very hard to detect for frequencies above 4–5 kHz". One of the reviewers of this paper mentioned that the phase spectrum may carry intelligibility information only up to 1 kHz. The reviewer has suggested that we explore this phenomenon using our analysis–modification–synthesis framework. In order to do this, we first create a finite impulse response (FIR) low-pass filter. This is designed using Parks–McClellan optimal equiripple method (201 coefficients, passband edge frequency of 1 kHz, and stopband edge frequency of 1.15 kHz such that stopband attenuation is approximately 40 dB). For interest, we also create an FIR high-pass filter to explore the importance of the phase spectrum above 1 kHz (201 coefficients, stopband edge frequency of 1 kHz and passband edge frequency of 1.15 kHz).

We conceive of three ways to apply these filters:

(1) *Pre-modification filtering:* Filter each of the 64 original VCV speech signals and subsequently create the magnitude-only and phase-only stimuli from these.
(2) *Post-modification filtering:* Filter the magnitude-only and phase-only signals.
(3) *Pre- and post-modification filtering:* Filter the original signals and apply the filter again to the magnitude-only and phase-only stimuli. Note that magnitude-only and phase-only modifications on a bandlimited signal will introduce measurable magnitudes at frequencies outside of the band due to STFT modifications (since STFT modification is a non-linear process).

This results in 17 types of stimuli for this experiment (this also includes the non-bandlimited original, magnitude-only, and phase-only signals). The magnitude-only and phase-only modifications are made in the same way as described in Experiments 2 and 3 (i.e., 32 ms window, 1/8 overlap, Hamming window for magnitude-only stimuli and rectangular window for phase-only stimuli). The stimuli types and their associated intelligibility scores are presented in Table B.1. At this point in time, we have only collected the results of one listener (therefore, there are only 64 tokens used for the evaluation of

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

Table B.1
Average consonant intelligibility of stimuli constructed from bandlimited speech

Type of stimuli			Intelligibility score (%)
Category	Filter type	Filter position	
Original	None	n/a	95.3
	Low-pass	n/a	84.4
	High-pass	n/a	96.9
Phase-only	None	n/a	93.8
	Low-pass	Pre	64.1
		Post	82.8
		Both	64.1
	High-pass	Pre	76.6
		Post	75.0
		Both	73.4
Magnitude-only	None	n/a	89.1
	Low-pass	Pre	46.9
		Post	70.3
		Both	48.4
	High-pass	Pre	95.3
		Post	82.8
		Both	89.1

each stimuli type). The following indicative observations can be made[14]:

- For low-pass filtering, on average, the intelligibility of phase-only stimuli seems to be better than that of magnitude-only stimuli (average intelligibility of low-pass phase-only stimuli is 70.3% and average intelligibility of low-pass magnitude-only stimuli is 54.7%).
- For high-pass filtering, on average, the intelligibility of magnitude-only stimuli seems to be better than that of phase-only stimuli (average intelligibility of high-pass phase-only stimuli is 75.0% and average intelligibility of high-pass magnitude-only stimuli is 89.1%).
- For both magnitude-only and phase-only stimuli, low-pass filtering seems to be best when applied as post-modification filtering.

These preliminary results seem to indicate that below 1 kHz the phase spectrum is providing more toward intelligibility than the magnitude spectrum. Further testing will enable us to verify these findings conclusively.

[14] The significance of these statements will be verified by collecting more data and subsequently presented in a future paper.

References

Abe, T., Kobayashi, T., Imai, S., 1995. Harmonics tracking and pitch extraction based on instantaneous frequency. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, pp. 756–759.

Allen, J.B., Rabiner, L.R., 1977. A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE 65 (11), 1558–1564.

Alsteris, L.D., Paliwal, K.K., 2004. Importance of window shape for phase-only reconstruction of speech. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, May, pp. I-573–I-576.

Alsteris, L.D., Paliwal, K.K., 2005. Evaluation of the modified group delay feature for isolated word recognition. In: Proc. Internat. Symposium on Signal Processing and its Applications, August, pp. 715–718.

Bozkurt, B., Doval, B., D'Alessandro, C., Dutoit, T., 2004. Appropriate windowing for group delay analysis and roots of *z*-transform of speech signals. In: EUSPICO.

Charpentier, F.J., 1986. Pitch detection using the short-term phase spectrum. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, April, pp. 113–116.

Crochiere, R.E., 1980. A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust., Speech and Signal Processing ASSP-28 (1), 99–102.

Delgutte, B., 1996. Physiological models for basic auditory percepts. In: Hawkins, H.L., McMullen, T.A., Popper, A.N., Fay, A.R. (Eds.), Auditory Computation. Springer-Verlag, New York.

Dimitriadis, D., Maragos, P., 2003. Robust energy demodulation based on continuous models with application to speech recognition. In: Proc. Eurospeech, September, pp. 2853–2856.

Duncan, G., Yegnanarayana, B., Murthy, H.A., 1989. A nonparametric method of formant estimation using group delay spectra. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, May, pp. 572–575.

Friedman, David H., 1985. Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, March, pp. 1121–1124.

Griffin, D.W., Lim, J.S., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust., Speech Signal Process. ASSP-32 (2), 236–243.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2004a. Application of the modified group delay function to speaker identification and discrimination. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, May, pp. I-517–I-520.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2004b. Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Proc. Internat. Conf. on Speech, Language Processing, October.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2004c. The modified group delay feature: a new spectral representation of speech. In: Proc. Internat. Conf. on Speech, Language Processing, October.

Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervocalic stop consonants. Speech Comm. 22 (4), 403–417.

Murthy, H.A., Gadde, V., 2003. The modified group delay function and its application to phoneme recognition. In: Proc.

IEEE Internat. Conf. on Acoust., Speech, Signal Processing, April, pp. I-68–I-71.

Murthy, H.A., Madhu Murthy, K.V., Yegnanarayana, B., 1989. Formant extraction from Fourier transform phase. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, May, pp. 484–487.

Nakatani, T., Irino, T., Zolfaghari, P., 2003. Dominance spectrum based v/uv classification and F_0 estimation. In: Proc. on Eurospeech, September, pp. 2313–2316.

Nuttall, A.H., 1981. Some windows with very good sidelobe behaviour. IEEE Trans. Acoust., Speech Signal Process. ASSP-29 (1), 84–91.

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69 (May), 529–541.

Oppenheim, A.V., Schafer, R.W., 1975. Digital Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Paliwal, K.K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proc. on Eurospeech, September, pp. 2117–2120.

Paliwal, K.K., Alsteris, L.D., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Comm. 45 (2), 153–170.

Paliwal, K.K., Atal, B.S., 2003. Frequency-related representation of speech. In: Proc. on Eurospeech, September, pp. 65–68.

Portnoff, M.R., 1981. Short-time Fourier analysis of sampled speech. IEEE Trans. Acoust., Speech Signal Process. ASSP-29 (3), 364–373.

Potamianos, A., Maragos, P., 1996. Speech formant frequency and bandwidth tracking using multiband energy demodulation. J. Acoust. Soc. Amer. 99, 3795–3806.

Potamianos, A., Maragos, P., 2001. Time-frequency distributions for automatic speech recognition. IEEE Trans. Speech Audio Process. 9 (March), 196–200.

Prasad, V.K., Nagarajan, T., Murthy, H.A., 2004. Automatic segmentation of continuous speech using minimum phase group delay functions. Speech Comm. 42, 429–446.

Quatieri, T.F., 2002. Discrete-time Speech Signal Processing. Prentice-Hall, Upper Saddle River, NJ.

Reddy, N.S., Swamy, M.N.S., 1985. Derivative of phase spectrum of truncated autoregressive signals. IEEE Trans. Circ. Systems CAS-32 (6).

Satyanarayana, P., Yegnanarayana, B., 1999. Robustness of group-delay based method for extraction of significant instants of excitation from speech signals. IEEE Trans. Speech Audio Process. 7 (6), 609–619.

Schroeder, M.R., 1975. Models of hearing. Proc. IEEE 63, 1332–1350.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. 3 (5), 325–333.

Wang, Y., Hansen, J., Allu, G.K., Kumaresan, R., 2003. Average instantaneous frequency and average log envelopes for ASR with the aurora 2 database. In: Proc. on Eurospeech, September, pp. 25–28.

Yegnanarayana, B., Murthy, H.A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Process. 40 (9), 2281–2289.

Yegnanarayana, B., Saikia, D.K., Krishnan, T.R., 1984. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans. Acoust., Speech Signal Process. ASSP-32 (3), 610–623.