ASR on Speech Reconstructed from Short-time Fourier Phase Spectra

Leigh D. Alsteris and Kuldip K. Paliwal

School of Microelectronic Engineering
Grif£th University, Brisbane, Australia
L.Alsteris@griffith.edu.au, K.Paliwal@griffith.edu.au

Abstract

In our earlier papers [1, 2], we have measured human intelligibility of speech stimuli reconstructed either from the short-time magnitude spectra (magnitude-only stimuli) or the short-time phase spectra (phase-only stimuli) of a speech stimulus. We demonstrated that, even for small analysis window durations of 20-40 ms (of relevance to automatic speech recognition), the short-time phase spectrum can contribute to speech intelligibility as much as the short-time magnitude spectrum. In this paper, we perform automatic speech recognition on magnitude-only and phase-only stimuli. When employing an MFCC-based front-end, the recognition achieved for these phase-only stimuli is much worse than magnitude-only stimuli at small analysis window durations, which is not consistent with their corresponding human intelligibility results. This implies that the MFCC feature set is not capturing all of the discriminating information present in the speech signal.

1. Introduction

In automatic speech recognition (ASR), speech is processed frame-wise using a temporal window of duration 20-40 ms. At such small durations, it is generally believed that the phase spectrum[1] does not contribute much to speech intelligibility and, as a result, state-of-the-art systems discard the phase spectrum in favour of features that are derived purely from the magnitude spectrum.

In a recent study we have conducted a number of human perception experiments, the results of which indicate that the phase spectrum can contribute signi£cantly to speech intelligibility over small window durations of 20-40 ms. These results may have direct implications for ASR. In this paper we summarise the human perception experiments, conduct ASR experiments on some popular databases, and compare the two sets of results.

The paper outline is as follows: In Section 2, we detail the analysis-modi£cation-synthesis (AMS) method used to create the phase-only (PO) and magnitude-only (MO) stimuli. The perception experiments are explained in Section 3. In the £rst experiment, we compare intelligibility of PO and MO stimuli constructed at both small and large window durations. In the second experiment, we synthesise different types of PO stimuli (at a small analysis window duration), testing for the intelligibility with a number of combinations of design parameter settings. ASR results for PO and MO stimuli are presented in Section 4.

2. Analysis-modi£cation-synthesis procedure

Although speech is non-stationary, it can be assumed to be quasi-stationary and, therefore, can be processed through a short-time Fourier analysis[2] [3, 4, 5, 6, 7]. 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 $S(f,t)$ 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.

The aim of the experiments in the following sections is to determine the contribution that phase and magnitude provide towards speech intelligibility. Accordingly, stimuli are created either from phase or magnitude. 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 modi£ed STFT $\hat{S}(f,t)$; that is,

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

This modi£ed STFT is then used to synthesize the signal $\hat{s}(t)$ using the overlap-add method [3, 4]. The synthesized signal $\hat{s}(t)$ contains all 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, we retain each frame's magnitude spectrum and randomise each frame's phase spectrum; that is, the modi£ed STFT is computed as follows:

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

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

In the STFT-based AMS 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))? A tapered window function (eg. Hamming) has been used in earlier studies [8]. Considering these studies have found the phase spectrum to be unimportant at small window durations, a rectangular window function is investigated in this study in addition to a Hamming window function. Second, what should be the duration T_w

[1] From here in, the modi£er 'short-time' is implied when mentioning the phase spectrum and magnitude spectrum.

[2] The modi£er 'short-time' implies a £nite-time window over which the properties of speech may be assumed stationary; it does not refer to the actual duration of the window. We use the qualitative terms 'small' and 'large' to make reference to the duration.

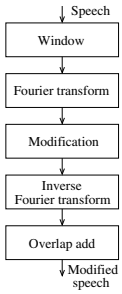

Figure 1: *Speech analysis-modi£cation-synthesis system.*

of the window function? In our experiments we consider a small duration of 32 ms and a large duration of 1024 ms. Third, how often should we compute STFT; i.e., how often should we sample the STFT across 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 $T_w/4$ [3]. 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. We also refer to the STFT sampling period as the frame shift. The last design issue to consider is that of zero-padding. For a windowed frame of length N (where N is a power of 2), 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 modi£ed, then inverse Fourier transformed to get a reconstructed signal of length $2N$. Only the £rst N points are retained, while the last N points are discarded. This is done in order to minimise aliasing effects.

3. Human perception experiments

3.1. Experiment 1

We compare the intelligibility of MO and PO 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 frame shift 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 sampled at 16 kHz (16-bit). Each recording is processed by the AMS system (Fig. 1) to retain either only phase information or only magnitude information.

As listeners, we use 12 native Australian English speakers with normal hearing, all within the age group of 20-35 years. Each person are tested in isolation in a silent room. The reconstructed 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 identi£cation (or, intelligibility) accuracy for each subject for different conditions.

The following observations can be made from the results in Table 1 (see [2] for signi£cance £gures):

1. For the large window duration of 1024 ms, phase spectrum provides signi£cantly more information than magnitude spectrum for both the Hamming and rectangular window functions. This is consistent with results reported in [8, 9, 10].
2. Intelligibility of MO32 stimuli is signi£cantly better than the PO stimuli when the Hamming window function is used, but these are comparable when the rectangular window function is used. Thus, if a rectangular window function is used in the AMS 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. The Hamming window provides better intelligibility than the rectangular window for MO32 stimuli, while the rectangular window is better than the Hamming window for construction of PO32 stimuli.
4. The best intelligibility results from MO32 stimuli (obtained by using a Hamming window) are signi£cantly better than the best results from PO32 stimuli (obtained using a rectangular window).

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 window function spectrum $W(f)$. The window's magnitude spectrum $|W(f)|$ has a big main lobe and a number of side lobes, causing 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 MO 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 MO 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 PO utterances.

3.2. Experiment 2

Intelligibility results for PO32 stimuli are better than previously reported by Liu et al. [8]. This improvement is made by altering a number of parameter settings in the AMS framework. In this experiment, we wish to determine the contribution to intelligibility by

Table 1: *Consonant intelligibility of MO and PO stimuli constructed using small and large window durations (32 ms and 1024 ms). Results for two types of analysis window (rectangular and Hamming) are given. A frame shift of $T_w/8$ is used.*

Type of stimuli	Intelligibility (in %) for			
	32 ms		1024 ms	
	Ham.	Rect.	Ham.	Rect.
Original	89.9	89.9	89.9	89.9
Magn. only	84.2	78.1	14.1	13.3
Phase only	59.8	79.9	88.0	89.3

Table 2: *Consonant intelligibility for PO32 stimuli constructed with various parameter settings.*

Type of stimuli	Parameter Settings			PO Intelligbililty
	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%

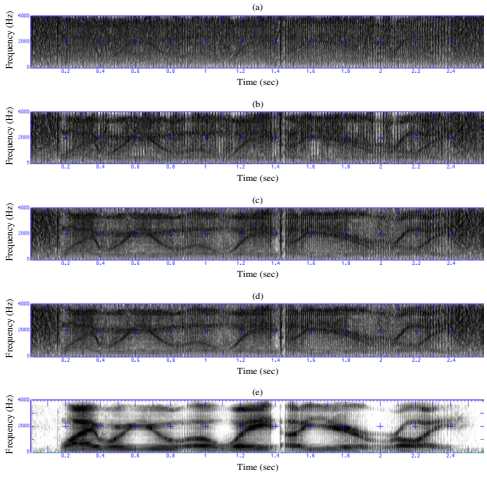

Figure 2: *The spectrograms of PO stimuli at an analysis window duration of 32 ms: (a) stimulus type A, (b) stimulus type B, (c) stimulus type C, (d) stimulus type D, and (e) spectrogram of the original speech sentence "Why were you away a year Roy?".*

each of these parameter settings. In order to do this, a number of combinations of settings for the AMS parameters are tested. Table 2 details the parameters used to construct each type of stimuli and the names we use to refer to them in this experiment, as well as intelligbility scores. Details of the experimental setup are the same as those used previously.

From the scores, we conclude that the major contribution to overall intelligibility comes from the use of the rectangular window (stimulus 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 (stimulus C) and the use of zero-padding (stimulus D), due to their roles in reducing aliasing effects. Fig. 2 presents the spectrogram of a sentence of speech with its reconstructed PO stimuli A, B, C and D. The increasing clarity of formant tracks in these spectrograms, from A through D, is indicative of the corresponding trend in intelligibility of these stimuli.

4. ASR experiments

In this section, we perform ASR experiments with PO and MO stimuli created from speech in the ISOLET and Aurora II databases. The PO and MO stimuli are constructed using a rectangular and Hamming analysis window respectively, with a frame shift of $T_w/8$. This is done at both small and large analysis durations (32 ms and 1024 ms).

We use an MFCC-based front-end with the following settings:

- Frame duration: 20 ms
- Frame shift: 10 ms
- Window type: Hamming
- Pre-emphasis coef£cient: 0.97
- Frequency range: 0-4 kHz
- Number of £lter-bank energies: 24
- Number of cepstral ceof£cents: 12 (exlcuding c0).

Using the Cambridge HMM Toolkit [12], we do both training and testing with the original speech, PO speech and MO speech. Systems are trained with SNR=∞ and tested over a range of SNRs[3].

4.1. Isolated word task (ISOLET)

The ISOLET database is an isolated-word, speaker-independent task, sampled at 8 kHz. The vocabulary is 26 English alphabet letters. Two repetitions of each letter are recorded for each speaker. Speakers are divided into 2 sets: 90 for training, 30 for testing. Each word is modeled by a HMM with 5 emitting states and 5 Gaussian mixtures per state. Although the vocabulary is relatively small, this is a dif£cult task as all words are short and highly confusable. Word accuracy scores, over a range of test SNRs, are given in Fig. 3.

4.2. Connected digit task (Aurora II)

Aurora II caters for speaker-independent experiments using several noise types and SNRs. Speech consists of digit sequences derived from the TI digit database down-sampled to 8 kHz and £ltered with a G.712 characteristic. Each digit (0-9) is modelled using a HMM with 16 emitting states and 3 Gaussian mixtures per state. We train with the clean training set (8440 utterances). The test set (28028 utterances) is divided evenly among 7 SNRs (∞ 20,15,10,5,0,-5 dB) and 4 noise types (subway, babble, car, exhibition). Word accuracy scores (which take into account insertions and deletions) are given in £gures 4 and 5.

4.3. Results

MO32, MO1024, and PO1024 recognition scores agree with the trends observed in the human perception experiments (thus, we will not discuss these further). However, PO32 results for both databases are worse than expected. While these signals sound quite intelligible, the ASR system provides poor word accuracy. This poor performance is attributed to a small dynamic range of the PO32 magnitude spectra. This translates to less discriminating power for the

[3]Note that for the case of PO and MO stimuli, the noise must be added to original test-set speech before modi£cation.

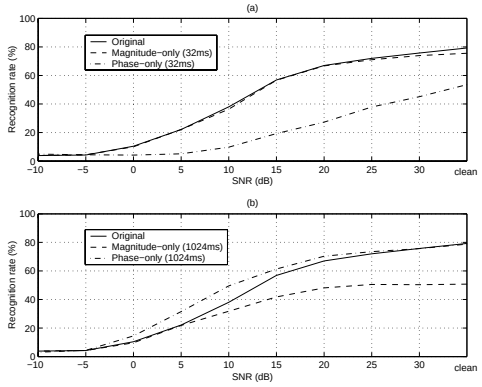

Figure 3: *Word accuracy versus SNR for ISOLET. White noise results provided. PO and MO stimiuli are constructed with analysis window durations of (a) 32 ms and (b) 1024 ms.*

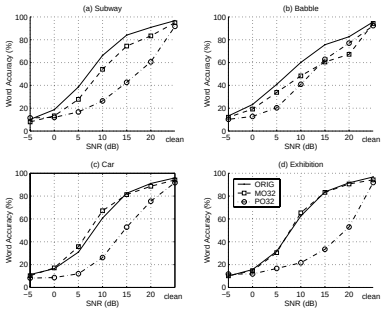

Figure 4: *Word accuracy versus SNR for Aurora II. Four noise types are investigated. PO and MO stimuli are constructed with analysis window duration of 32 ms.*

MFCC features. To demonstrate this, we take an MFCC vector from original speech as well as the vectors at the corresponding locations in the MO32 and PO32 reconstructions. A zero is appended to the beginning of each of these vectors to account for the absence of c0. An inverse DCT is then calculated for each vector (Figs. 6(a,c,e)). The inverse DCT allows us to view the reduced spectral information with which the models are trained. This procedure is repeated for the addition of white noise at an SNR of 20dB, at 20 different seed values for the random noise generator (Figs. 6(b,d,f)).

The small amount of discriminating power in the PO32 spectra (Fig. 6(e)) is suf£cient to obtain good word accuracy in clean conditions for Aurora II, but insuf£cient in clean conditions for the highly confusable Isolet task. The poorer performance in lower SNRs is best explained by viewing Figs. 6(b,d,f), which illustrate the relative variability in the magnitude spectra of the original speech and its MO32 and PO32 reconstructions. Note that the PO32 magnitude spectrum exhibits much more variability than the MO32 magnitude spectrum.

5. Conclusions

In this paper, we have reviewed the results of our recently conducted perception experiments, which demonstrate that the phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum at small analysis window durations. We have conducted ASR experiments on MO and PO stimuli constructed from speech stimuli in the Isolet and Aurora II databases. The ASR performance for PO stimuli is much worse than MO stimuli at small analysis window durations. This result is not consistent with the high PO intelligibility measured from human perception experiments. Since both MO and PO stimuli are almost equally intelligible to humans, and ASR recognises one much better than the other, it could be possible that the MFCC feature set is inadequate. That is, there seems to be some discriminating information in the phase part of the speech signal that is not being captured by the MFCC representation. These results demonstrate that there is a need for further research in the £eld of feature extraction.

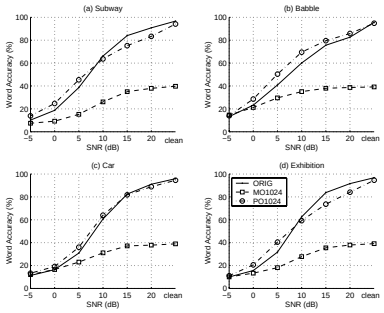

Figure 5: *Word accuracy versus SNR for Aurora II. Four noise types are investigated. PO and MO stimuli are constructed with analysis window duration of 1024 ms.*

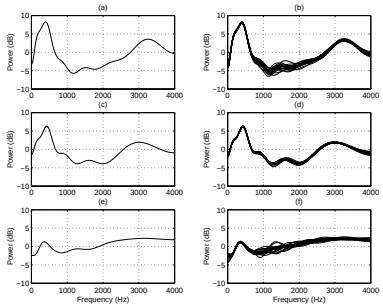

Figure 6: *The left £gures present magnitude spectra produced by inverse DCT of an MFCC vector from (a) original speech and its (c) MO32 and (e) PO32 stimuli. The right £gures present the magnitude spectra for 20 observations of white noise at 20dB SNR for (b) original speech and its (d) MO32 and (f) PO32 stimuli.*

6. Acknowledgements

This work was partly supported by ARC (Discovery) grant (No. DP0209283). Thanks to the volunteers who took part in the perception tests.

7. References

[1] K.K. Paliwal, "Usefulness of phase in speech processing", Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan, pp. 1-6, Feb. 2003.

[2] K.K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception", Proc. Eurospeech, Geneva, Switzerland, pp. 2117-2120, Sept. 2003.

[3] J.B. Allen and L.R. Rabiner, "A uni£ed approach to short-time Fourier analysis and synthesis" Proc. IEEE, Vol. 65, No. 11, pp. 1558-1564, 1977.

[4] D.W. Grif£n and J.S. Lim, "Signal estimation from modi£ed short-time Fourier transform", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, pp. 236-243, 1984.

[5] M.R. Portnoff "Short-time Fourier analysis of sampled speech" IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364-373, 1981.

[6] T.F. Quatieri, *Discrete-time speech signal processing*, Prentice Hall, Upper Saddle River, NJ, 2002.

[7] L.R. Rabiner and R.W. Schafer, *Discrete-time speech signal processing, principles and practice*, Prentice Hall, Englewood Cliffs, NJ, 1978.

[8] 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.

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

[10] M.R. Schroeder, "Models of hearing", Proc. IEEE, Vol. 63, pp. 1332-1350, 1975.

[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] S. Young, *The HTK Book*, Cambridge University Engineering Department, Cambridge, England, 2001.