On the Relative Importance of the Short-Time Magnitude and Phase Spectra Towards Speaker Dependent Information

Kamil K. Wójcicki, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith University, Nathan QLD 4111, Australia
k.wojcicki@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this work, we investigate the relative contribution of the short-time magnitude and phase spectra towards speaker dependent information. The effect of the analysis window function type is also examined. For this purpose we conduct a human speaker verification experiment that uses phase-only and magnitude-only stimuli. The stimuli are constructed using the analysis-modification-synthesis procedure. The results of our pilot experiment show that the short-time magnitude spectrum contains little speaker information for a low dynamic range analysis window and high amount of speaker information for a large dynamic range window. On the other hand, the short-time phase spectrum contains speaker information predominantly for the low dynamic range analysis window. These suggestive results show that the short-time phase spectrum, commonly discarded in feature extraction for speaker verification, contains useful speaker information. This suggests that further research into feature extraction from the short-time phase spectrum is warranted.

1. Introduction

Although speech is non-stationary, it can be assumed quasi-stationary and therefore can be processed through a short-time Fourier analysis. 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. The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

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

where $w(t)$ is an analysis window function of duration T_w. In speech processing, the Hamming window function is typically used and its width is normally 20–40 ms. The short-time Fourier spectrum, $S(t,f)$, is a complex quantity and can be expressed in polar form as follows,

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

where $|S(t,f)|$ is the short-time magnitude spectrum and $\psi(t,f) = \angle S(t,f)$ is the short-time phase spectrum. The signal $s(t)$ is completely characterised by its magnitude and phase spectra.[1] In our previous work [1, 2, 3], we have investigated the intelligibility resulting from magnitude-only and phase-only stimuli for different analysis window functions. We have found that phase-only stimuli have high intelligibility when constructed using a low dynamic range[2] analysis window. The aim of the present work is to report some preliminary results on the relative importance of the short-time magnitude and phase spectra for speaker verification. For this purpose, we conduct a human speaker verification experiment, where we investigate the Chebyshev analysis window function over three dynamic range settings: 10 dB, 35 dB and 60 dB. The Hamming and the rectangular windows are also included. In our experiment, we use stimuli that contain only the short-time magnitude spectrum information or only the short-time phase spectrum information. We refer to these stimuli as magnitude-only and phase-only stimuli, respectively. The stimuli are constructed using small window durations (32 ms). To retain magnitude-only information the phase spectrum values are randomised. On the other hand, to retain phase-only information the short-time magnitude spectrum values are set to unity. The results of our pilot experiment show that the magnitude-only stimuli contain speaker dependent information for the high dynamic range window (Chebyshev 60 dB), while phase-only stimuli have speaker dependent information predominantly for the lower dynamic range windows (Chebyshev 10 dB and 35 dB). These indicative results are interesting as the long term outlook of this work is to employ the short-time phase spectrum information for machine speaker verification.

This paper is organised as follows. In Section 2, we detail the analysis-modification-synthesis (AMS) procedure used to generate the stimuli. Description of our human speaker verification experiment is given in Section 3. The results and discussion are presented in Section 4.

2. Analysis-modification-synthesis

The aim of this study is to determine the relative contribution of the short-time magnitude and phase spectra towards speaker dependent information. Accordingly, stimuli retaining only the magnitude or phase information are constructed. For this purpose, the analysis-modification-synthesis (AMS) procedure, shown in Fig. 1, is used. In the AMS framework the speech signal is divided into overlapped frames of small duration. The frames are then windowed using an analysis window, $w(t)$, followed by the Fourier analysis, and spectral modification. The spectral modification stage is where only the magnitude or phase information is retained. For example, to construct magnitude-only (MO) stimuli only the magnitude spectrum information is retained. The phase spectrum information is removed by randomising the phase spectrum values.

[1] Throughout our discussions, when referring to the magnitude or phase spectrum, the use of the short-time Fourier transform (STFT) over small window durations (20–40 ms) is implied, unless otherwise stated.

[2] Note that we use the term 'dynamic range' exclusively for the Chebyshev window function to refer to the attenuation (in magnitude response) of the side-lobe level w.r.t. to the main-lobe.

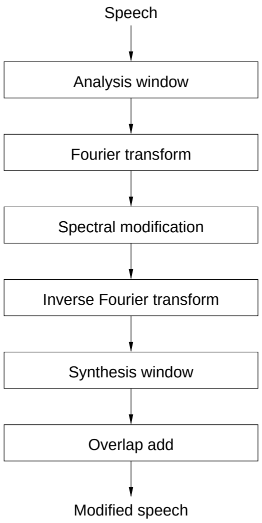

Figure 1: *Block diagram of the speech analysis-modification-synthesis (AMS) procedure used for generation of stimuli files.*

The resulting modified STFT is given by

$$\widehat{S}(t,f) = |S(t,f)|\mathrm{e}^{j\phi(t,f)}, \tag{3}$$

where $\phi(t,f)$ is a random variable uniformly distributed between 0 and 2π. The stimulus, $\hat{s}(t)$, is then constructed by taking the inverse STFT of the $\widehat{S}(t,f)$, followed by an overlap-add (OLA) synthesis [4, 5, 6, 7]. The resulting signal contains all of the information about the short-time magnitude spectra contained in the original signal $s(t)$, but has no information about the short-time phase spectra. Similarly, for generation of phase-only (PO) stimuli the magnitude spectrum values are set to unity while the phase spectrum values are left unchanged, resulting in

$$\widehat{S}(t,f) = \mathrm{e}^{j\psi(t,f)}, \tag{4}$$

where $\widehat{S}(t,f)$ is the modified STFT.

In this study, our goal is to investigate the effect of the analysis window under a human speaker verification task. In particular, we are interested in the effect of the dynamic range of an analysis window on the relative speaker dependent information content of the magnitude and phase spectra. For this purpose, we employ the Chebyshev window characterised by adjustable equi-ripple side-lobe attenuation [8, 9, 10]. These properties allow us to investigate the effect of the dynamic range in a systematic manner. Three dynamic range settings for the Chebyshev window are considered: 10 dB, 35 dB and 60 dB. Since the Hamming window is commonly employed in speech processing and because both the Hamming and the rectangular windows were investigated in the past in related studies [1, 2, 3] we include them in this present work for completeness. Frequency magnitude response comparisons for the window function types employed in our experiment are shown in Fig. 2. For the synthesis window we employ the modified Hanning window [7].

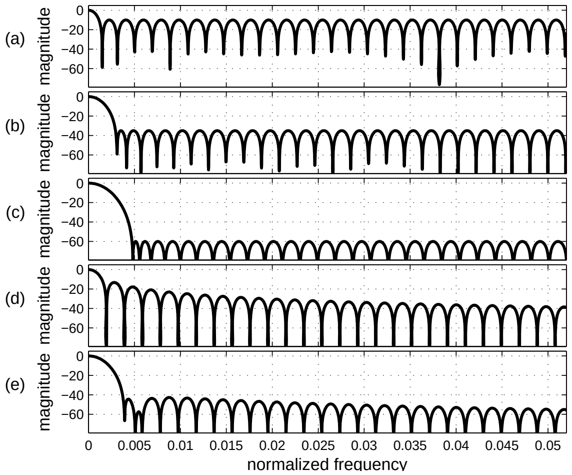

Figure 2: *Magnitude spectrum (in dB) characteristics of the Chebyshev window with: (a) 10 dB, (b) 35 dB, (c) 60 dB side-lobe attenuation; (d) the rectangular window; and (e) the Hamming window. Window function lengths of* 512 *samples were used for this illustration. The FFT analysis length was set to* 32768. *Note that the large FFT size and the limited normalised frequency axis range were used in order to facilitate comparison of both main-lobe and side-lobe characteristics.*

3. Human speaker verification experiment

3.1. Corpus

For the purposes of our experiment we recorded a small gender balanced speaker corpus. The corpus is composed of speech belonging to twelve speakers. Five different sentences are used. Each speaker repeats each sentence six times per session. There are three sessions each separated by two months. The recordings are sampled at 16 kHz with 16 bit precision.

3.2. Stimuli

In our experiment we use a subset of the recorded corpus. The subset consists of two recordings of a continuously spoken digit string “three nine zero two six seven” for each of the twelve speakers. First recording for each speaker is processed using the AMS procedure detailed in Section 2. This is performed to retain only magnitude or phase information. Small window durations, T_w=32 ms, are used throughout. The frame shift is set to T_w/8 to minimise aliasing. The FFT analysis length is set to $2N$, where N is the number of samples in each frame. The resulting stimuli can be grouped as follows: magnitude-only (MO) and phase-only (PO). One of the objectives of this study is to determine the effect of the dynamic range of an analysis window on the speaker dependent information of the magnitude and phase spectra. To achieve this three Chebyshev analysis windows (with 10 dB, 35 dB and 60 dB side-lobe attenuation) as well as the Hamming and the rectangular windows are investigated for each of the above groups. Original (unprocessed) stimuli are also included resulting in total of eleven treatment types. As mentioned before, each of the treatments is applied to one of the two sentences for each speaker, thus producing 132 processed stimuli files. Each of the remaining 12 unprocessed stimuli is played back to the listener against each of the 132 processed stimuli, in a randomised manner, for speaker verification.

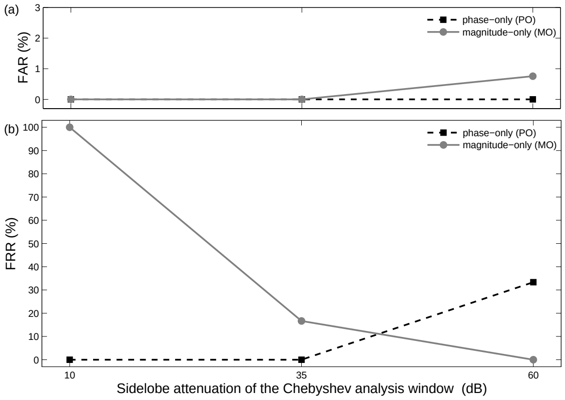

Figure 3: *Human speaker verification results as a function of the dynamic range of the Chebyshev analysis window for magnitude-only (MO) stimuli (dotted line) and phase-only (PO) stimuli (broken line): (a) false acceptance rate (FAR); (b) false rejection rate (FRR). Note that in the above plots, the lines between the 10, 35 and 60 dB points are a simple interpolation and provide only an indicative idea of a general trend.*

3.3. Procedure

The listening tests were conducted in isolation, over a single session, in a quiet room. The task was to verify if two utterances, played one after another, belonged to the same speaker. The stimuli audio file pairs were played in a randomised order and presented over closed circumaural headphones (SONY MDR-V500) at a comfortable listening level. One listener was used in this pilot study. The listener was presented with 1584 speaker verification tests. The entire sitting lasted approximately four hours with many five minute breaks. The responses were collected via a keyboard.

4. Results and Discussion

The results of this pilot study, in terms of standard speaker verification metrics [11], namely the false acceptance rate (FAR) and the false rejection rate (FRR), are shown in Fig. 3 as well as in Table 1. Note that the FAR in our task is close to zero due to the nature of our human speaker verification task. For this reason we limit our discussion to the FRR metric only. The following observations can be made from Fig. 3(b). For the high dynamic range analysis window (Chebyshev 60 dB), the FRR is low for magnitude-only (MO) stimuli. Thus, as commonly accepted, the short-time magnitude spectrum contains speaker dependent information for high dynamic range analysis windows. On the other hand, for the low dynamic range windows (Chebyshev 10 dB and 35 dB), most of the speaker information is suppressed from the short-time magnitude spectrum. On the other hand the phase-only (PO) stimuli contain speaker dependent information predominantly for the low dynamic range analysis windows (Chebyshev 10 dB and 35 dB). The above observations can be explained in terms of how the information is contained in the STFT. Recall, that the STFT given in (1) can be expressed (equivalently) in terms of magnitude and phase spectrum (2). The information contained in the speech signal is thus distributed across these two representations. If for low dynamic range windows $|S(t,f)|$ conveys less information about speech intelligibility then the corresponding $\psi(t,f)$ has to convey more information since the overall information contained in the complex spectrum is constant. An alternate reasoning is as follows. The phase spectrum can be computed using arctangent function (four quadrant version) as

$$\psi(t,f) = \arctan\left(\frac{\mathrm{Im}\{S(t,f)\}}{\mathrm{Re}\{S(t,f)\}}\right), \qquad (5)$$

where $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote real and imaginary parts, respectively. From (5), it is evident that when $|S(t,f)|$ is very small the argument of arctan function will be of $(\frac{0}{0})$ form; hence the phase estimates will not be numerically reliable. Such small values occur when the dynamic range of $w(t)$ (and thus $|S(t,f)|$) is large. On the other hand, low dynamic range analysis windows produce better behaved phase spectra estimates, since higher side-lobes imply that $|S(t,f)|$ is always relatively large.

Table 1: *Human speaker verification results for the original, magnitude-only (MO) and phase-only (PO) stimuli for the Hamming and rectangular analysis window functions.*

Modification	Window	FAR (%)	FRR (%)
Original		0.00	0.00
MO	Rectangular	0.00	0.00
MO	Hamming	0.00	0.00
PO	Rectangular	0.00	0.00
PO	Hamming	0.00	33.33

From Table 1 it can be seen that the phase-only (PO) stimuli contain less speaker dependent information when the Hamming window is used than when the rectangular window is employed. These results can also be explained in terms of the side-lobe attenuation difference between the analysis window functions.

The results of this study suggest that the short-time phase spectrum contains a significant amount of speaker dependent information. As such, feature extraction for speaker verification from the short-time phase spectrum could be considered as a potential research area. It has to be stressed however, that the above are pilot results only. Further tests, with a larger set of listeners, are required in order to draw firmer, statistically relevant conclusions.

5. Conclusion

In this paper, we investigated the effect that the dynamic range of an analysis window has on the relative importance of the short-time magnitude and phase spectra for human speaker verification. Experiment stimuli were constructed by retaining only the short-time magnitude or phase spectra, at short window durations (32 ms) while employing the Chebyshev analysis window at three dynamic range settings: 10 dB, 35 dB and 60 dB. It was shown that magnitude-only stimuli contain speaker dependent information for the analysis window with the large dynamic range, while the short-time phase spectrum was found to contain speaker information predominantly for the low dynamic range window. This suggests that pursuing research into derivation of speaker discriminative features from the short-time phase spectrum is warranted.

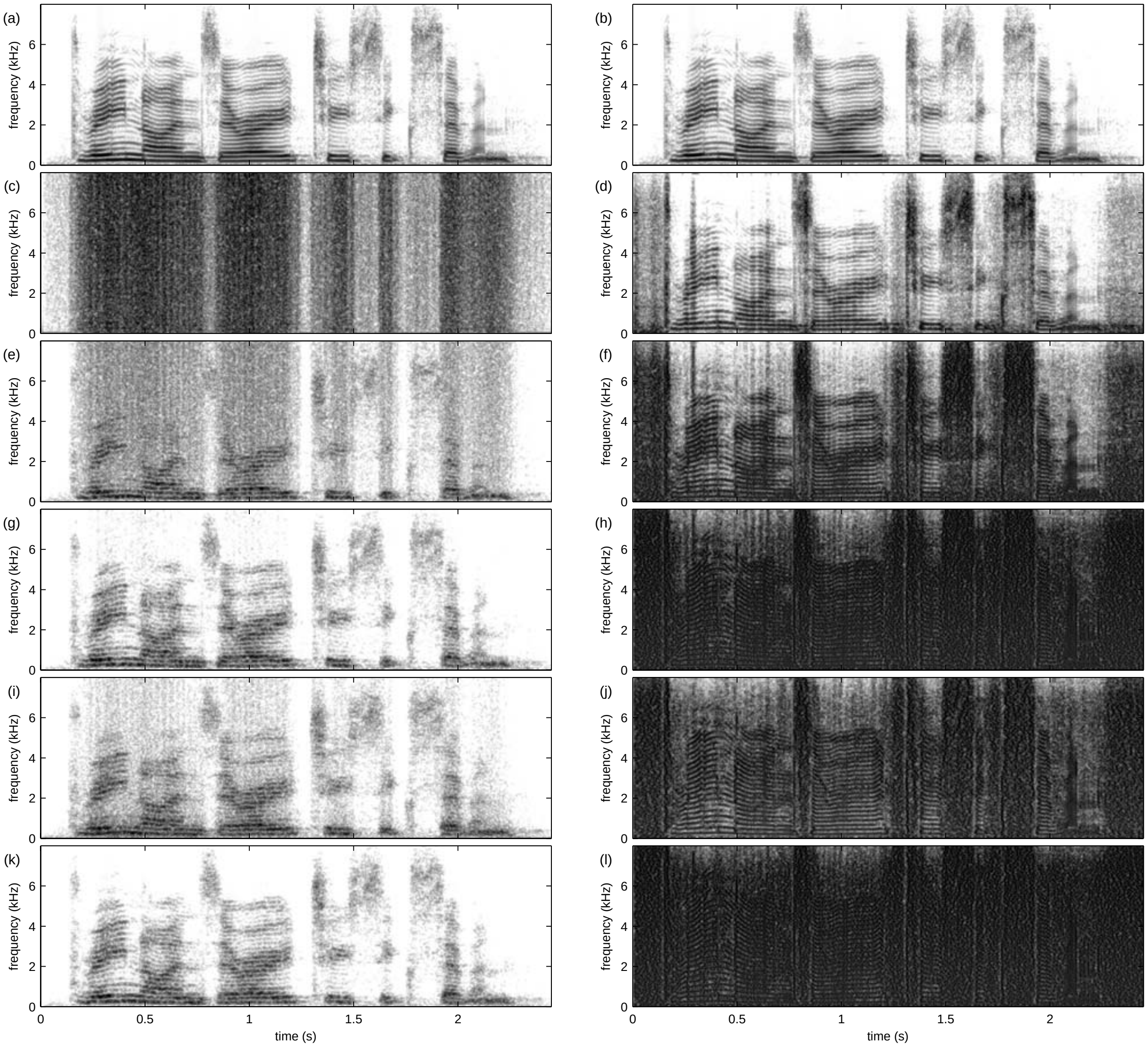

Figure 4: *Stimuli spectrograms for the utterance "three nine zero two six seven" from the speaker corpus by a female speaker: column one (c, e, g, i, k) magnitude-only (MO) stimuli; column two (d, f, h, j, l) phase-only (PO) stimuli; row one (a, b) AMS processed stimuli without spectral modification; row two (c, d) stimuli for the Chebyshev 10 dB analysis window; row three (e, f) stimuli for the Chebyshev 35 dB analysis window; row four (g, h) stimuli for the Chebyshev 60 dB analysis window; row five (i, j) stimuli for the rectangular analysis window; row six (k, l) stimuli for the Hamming analysis window.*

6. References

[1] L. Alsteris and K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, June 2006.

[2] L. Alsteris and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, May 2007.

[3] K. Wojcicki and K. Paliwal, "Importance of the dynamic range of an analysis window function for phase-only and magnitude-only reconstruction of speech," in *Proc. ICASSP*, Apr. 2007, vol. 4, pp. 729–733.

[4] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[5] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis / synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 99–102, Feb. 1980.

[6] M. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 364–373, 1981.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[8] C. Dolph, "A current distribution for broadside arrays which optimizes the relationship between beam width and sidelobe level," *Proc. IRE*, vol. 34, pp. 335–348, 1946.

[9] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[10] P. Kabal, "Time windows for linear prediction of speech, version 2," Tech. Rep., Dept. Electrical & Computer Engineering, McGill University, Dec. 2005.

[11] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, Aug. 1995.