USEFULNESS OF PHASE SPECTRUM IN HUMAN SPEECH PERCEPTION

Kuldip K. Paliwal and Leigh Alsteris

School of Microelectronic Engineering
Griffith University, Brisbane, Australia
e-mail: K.Paliwal@griffith.edu.au, L.Alsteris@griffith.edu.au

ABSTRACT

Short-time Fourier transform of speech signal has two components: magnitude spectrum and phase spectrum. In this paper, relative importance of short-time magnitude and phase spectra on speech perception is investigated. Human perception experiments are conducted to measure intelligibility of speech tokens synthesized either from magnitude spectrum or phase spectrum. It is traditionally believed that magnitude spectrum plays a dominant role for shorter windows (20-30 ms); while phase spectrum is more important for longer windows (128-3500 ms). It is shown in this paper that even for shorter windows, phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum if the shape of the window function is properly selected.

1. INTRODUCTION

Though 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. The short-time Fourier transform (STFT) of speech signal $s(t)$ is given by

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

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:

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

where $|S(\nu, t)|$ is the short-time magnitude spectrum and $\psi(\nu, t) = \angle S(\nu, t)$ is the short-time phase spectrum. Square of magnitude spectrum is called the power spectrum (i.e.; $P(\nu, t) = |S(\nu, t)|^2$). The signal $s(t)$ is completely characterized by its short-time power and phase spectra.

About 150 years ago, Ohm [1] observed that the human auditory system is phase-deaf; i.e., it ignores phase spec-

trum and uses only magnitude spectrum for speech perception. Helmholtz [2] confirmed Ohm's observation by experimenting with stimuli having same magnitude spectrum, but different phase spectra, and hearing no audible difference between these stimuli. Since then, a number of researchers have tried to explore audibility of phase spectrum and have found that Helmholtz's conclusion is not correct. They found [3, 4, 5, 6, 7, 8] that stimuli having same magnitude spectrum and generated as a sum of harmonically-related sinusoids are clearly distinguishable if they have different phase spectra. For speech coding and synthesis applications, it has been shown [9, 10, 11, 12] that the phase spectrum contributes to the quality and naturalness of the synthesized speech.

Though the phase spectrum carries half of the information about the speech signal (as seen from Eq. (2)), it has been totally discarded (or given very little importance) in most of the automatic speech recognition (ASR) applications [13]. For these applications, it is important to know whether phase spectrum provides any information which adds to the intelligibility of speech signal for human speech perception. If this is the case, then it will be useful for ASR applications.

A few studies have been reported in the literature which discuss whether phase spectrum provides any information which can help the human auditory system in identifying phonemes from the speech signal (i.e., whether phase spectrum is contributing to phoneme intelligibility for human speech recognition (HSR).) Schroeder [14] and Oppenheim and Lim [15] have informally observed that the phase spectrum is important for human speech recognition (HSR) for phoneme identification (or, intelligibility) when the window used for STFT is large (greater than 1 sec). When the window duration is small (about 20-40 ms), the short-time phase spectrum conveys no information about the intelligibility of speech. Liu et al. [16] have recently confirmed these findings through a more formal human speech perception study.

For ASR, speech signal is processed frame-wise using a temporal window of duration 20-40 ms. Therefore, if phase spectrum is of any use for ASR application, it should provide some information about phoneme intelligibility using small window durations (20-40 ms) in a human perception experiment.
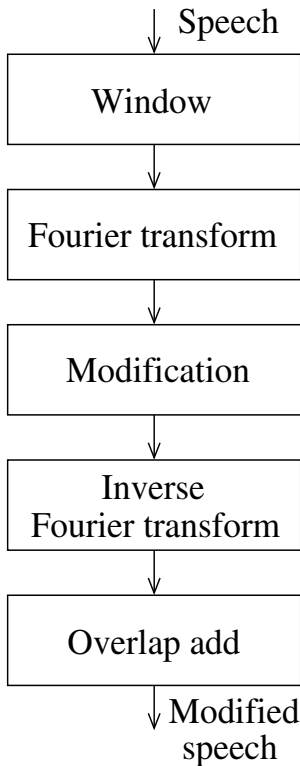
In this paper[1], the usefulness of phase information is explored in human speech perception. Through human listening tests, it is shown that the short-time phase spectrum (with window size of 32 ms) contributes to speech intelligibility as much as the corresponding power spectrum.

2. HUMAN PERCEPTION EXPERIMENTS AND RESULTS

Here, we assess the importance of short-time phase spectrum against the short-time magnitude spectrum through human perception experiments. For this, we record 16 commonly occurring consonants in Australian English in aCa context spoken in a carrier sentence "Hear aCa now". For example, for 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. Each of the 64 utterances are processed through a STFT-based speech analysis-modification-synthesis system (shown in Fig. 1) to retain either only phase information or only amplitude information.

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

In order to get, for example, an utterance with only phase information, the signal is processed through the STFT analysis using Eq. (1) and the short-time magnitude spectrum is

made unity in the modified STFT $\hat{S}(\nu, t)$; i.e.,

$$\hat{S}(\nu, t) = e^{j\psi(\nu, t)}. \tag{3}$$

This modified STFT is then used to synthesize the signal $\hat{s}(t)$ using the overlap-add method [18, 19]. The synthesized signal $\hat{s}(t)$ contains all the information about the short-time phase spectrum contained in the original signal $s(t)$, but will have no information about its short-time magnitude spectrum. We call 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 the short-time magnitude spectrum, but make the short-time phase spectrum totally random; i.e., the modified STFT is computed as follows:

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

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

In the STFT-based speech analysis-modification-synthesis system using the overlap-add method, there are three design issues that have to be addressed. First, what type of window function $w(t)$ should be used for computing STFT (Eq. (1))? Normally, a tapered window function (such as Hanning, Hamming or triangular) has been used in earlier studies [16]. Since these studies have found short-time phase spectrum to be unimportant, we decided to check a window function which is not tapered. Therefore, in our paper, we investigate two window functions: Hamming and rectangular. Second, what should be the duration T_w of the window function? In our study, we investigate the importance of STFT phase spectrum for two different durations: 1) $T_w = 32$ ms and 2) $T_w = 1024$ ms. Third, how often should we compute STFT; i.e., how often should we sample the STFT across time axis? Since we have to synthesize the signal from it, this should be done to avoid the aliasing errors. Thus, it is decided by the window function $w(t)$ used in the analysis. For example, for Hamming window, the sampling period should be at most $T_w/4$ [18]. To be on a safer side, we have used a sampling period of $T_w/8$; i.e., we update our frame every $T_w/8$. Though the rectangular window can be used with larger sampling period, we use the same value of sampling period (i.e., $T_w/8$) to maintain the consistency.

In our human perception (listening) tests, we use 12 subjects; all are native Australian English speakers within the age group of 20-35 years. The magnitude-only, phase-only and original utterances are played in random order to each subject through a headphone and the task of the subject is to identify each utterance as one of the sixteen consonants. This way, we get consonant identification (or, intelligibility) accuracy for each subject for different conditions. These perception tests are done in two sessions. In the first session, we use window duration of 32 ms. Results averaged over the 12 subjects are listed in Table 1. In the second session,

---

[1]Some of the results have been presented earlier in a conference [17].

Table 1. Consonant intelligibility (or, identification accuracy) of magnitude-only and phase-only utterances for Hamming and rectangular windows with window duration of 32 ms.

Type of	Intelligibility (in %) for	
stimuli	Hamming window	Rect. window
Original	88.8	88.8
Magn. only	84.2	78.1
Phase only	59.8	79.9

Table 2. Consonant intelligibility (or, identification accuracy) of magnitude-only and phase-only utterances for Hamming and rectangular windows with window duration of 1024 ms.
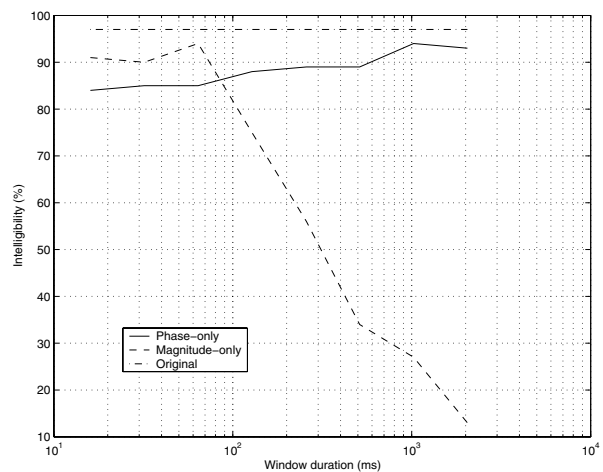
Type of	Intelligibility (in %) for	
stimuli	Hamming window	Rect. window
Original	90.9	90.9
Magn. only	14.1	13.3
Phase only	88.0	89.3

window duration of 1024 ms is used and the results are listed in Table 2.

We can make the following observations from these two tables: For longer window durations ($T_w = 1024$ ms), short-time phase spectrum provides significantly more information than the short-time magnitude spectrum for both the window functions[2]. This observation is consistent with the results reported earlier in the literature [14, 15, 16]. For shorter window durations ($T_w = 32$ ms), intelligibility of magnitude-only utterances is significantly better than the phase-only utterances for Hamming window function, but these are comparable for the rectangular window function. Thus, if we use the rectangular window function in the STFT analysis-modification-synthesis system, the short-time phase spectrum carries as much information about the speech signal as the short-time magnitude spectrum, even for shorter window durations ($T_w = 32$ ms) which are typically used in speech processing applications.

From Tables 1 and 2, we can make another observation. For shorter window durations ($T_w = 32$ ms), Hamming window provides better intelligibility for magnitude-only utterances; while rectangular window is better for the phase-only utterances. This result can be explained as follows. We know that multiplication of speech signal with a window function is equivalent to convolution of speech spectrum $S(f)$ with frequency response $W(f)$ of the window function. Window's spectrum $W(f)$ generally has a big main lobe and a number of side lobes. This causes two problems: frequency

---
[2]Analysis of variance and paired t-tests have been used throughout the paper to check statistical significance of the results at 0.01 level.

resolution problem and spectral leakage problem. The frequency resolution problem is caused by the main lobe of $W(f)$. Wider is the main lobe, larger frequency interval of the speech spectrum gets smoothed and worse becomes the frequency resolution problem. The side lobes cause spectral leakage problem; the amount of spectral leakage increases with the magnitude of side lobes. For magnitude-only utterances, we want to preserve the true magnitude spectrum of the speech signal. For the estimation of magnitude spectrum, frequency resolution as well as spectral leakage are serious problems. Since Hamming window has wider main lobe and smaller side lobes in comparison to rectangular window, the Hamming window provides 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 phase spectrum, the side lobes do not cause serious problem; the smoothing effect caused by the main lobe is more serious. Because of this, the rectangular window results in better intelligibility than the Hamming window for the phase-only utterances.

Fig. 2. Consonant identification performance (or, intelligibility) as a function of window duration for magnitude-only and phase-only utterances. Intelligibility for the original utterances (without any modification) is shown by horizontal dot-dashed line.

So far, we have provided intelligibility results for two window durations only. Now, we investigate how intelligibility varies with window duration for magnitude-only and phase-only utterances. For this, we use Hamming window for magnitude-only utterances and rectangular window for phase only utterances. Results are shown in Fig. 2. It can be observed from this figure that for the magnitude-only utterances, the intelligibility score remains almost constant for 16 to 64 ms window-durations and it decreases sharply for the window duration larger than 64 ms. For phase-only ut-

terances, the intelligibility scores is almost same for all the window durations.

3. CONCLUSION

In this paper, relative importance of short-time magnitude and phase spectra on speech perception is investigated. Human perception experiments are conducted to measure intelligibility of speech tokens synthesized either from magnitude spectrum or phase spectrum. It is traditionally believed that magnitude spectrum plays a dominant role for shorter windows (20-30 ms); while phase spectrum is more important for longer windows (128-3500 ms). It is shown in this paper that even for shorter windows, phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum if the shape of the window function is properly selected.

4. ACKNOWLEDGMENT

5. REFERENCES

[1] G.S. Ohm, "Uber die Definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher tonbildender Vorrichtungen", Ann. Phys. Chem., Vol. 59, pp. 513-565, 1843.

[2] H.L.F. von Helmholtz, *On the Sensations of Tone*, 1875 (English Translation by A.J. Ellis, Longmans, Green and Co., London, 1912).

[3] J.C.R. Licklider, "Effects of changes in the phase pattern upon the sound of a 16-harmonic tone", J. Acoust. Soc. Am., Vol. 29, pp. 780, 1957.

[4] M.R. Schroeder, "New results concerning monaural phase sensitivity", J. Acoust. Soc. Am., Vol. 31, pp. 1579, 1959.

[5] F.A. Bilson, "On the influence of the number and phase of harmonics on the perceptibility of the pitch of complex signals", Acoustica, Vol. 28, pp. 60-65, 1973.

[6] R. Plomp and H.J.M. Steeneken, "Effect of phase on the timbre of complex tones", J. Acoust. Soc. Am., Vol. 46, pp. 409-421, 1969.

[7] R. Carlson, B. Granstrom and D. Klatt, "Vowel perception: The relative perceptual salience of selected acoustic manipulations", Speech Trans. Lab. Q. Prog. Stat. Rep. (TRITA-TLF-79-4), Stockholm, Sweden, pp. 73-83, 1980.

[8] R.D. Patterson, "A pulse ribbon model of monaural phase perception", J. Acoust. Soc. Am., Vol. 82, pp. 1560-1586, 1987.

[9] H. Pobloth and W.B. Kleijn, "On phase perception in speech", Proc. ICASSP, pp. 29-32, 1999.

[10] D.S. Kim, "Perceptual phase redundancy in speech", Proc. ICASSP, pp. 1383-1386, 2000.

[11] H. Banno, K. Takeda and F. Itakura, "The effect of group delay spectrum on timbre", Acoust. Sci. and Tech., Vol. 23, pp. 1-9, 2002.

[12] H. Kawahara, I.M. Katsuse and A.D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, pp. 187-207, 1999.

[13] J.W. Picone, "Signal Modeling techniques in speech recognition", *Proc. IEEE*, Vol. 81, No. 9, pp. 1215-1247, 1993.

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

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

[16] 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.

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

[18] 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.

[19] D.W. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, pp. 236-243, 1984.