音声言語情報処理 45－1
（２００３．２．７）

Usefulness of Phase in Speech Processing

Kuldip K. Paliwal
School of Microelectronic Engineering
Griffith University, Brisbane, QLD 4111, Australia

Abstract: It is a common belief in speech community that the short-time phase spectrum plays very little (or, no) role in human perception tasks as well as in automatic speech recognition systems. In this paper, the usefulness of phase information is explored in human speech perception as well as in automatic speech recognition. Through human perception experiments, it is shown that the short-time phase spectrum (with widow size of 32 ms) contributes to speech intelligibility as much as the corresponding power spectrum. A representation based on frequencies of the speech signal derived from its short-time phase is developed and is found to be as good as cepstral representation (derived from power spectrum) for automatic speech recognition.

Keywords: Short-time phase spectrum, speech recognition, speech perception

音声処理における位相の有用性

Kuldip K. Paliwal
School of Microelectronic Engineering
Griffith University, Brisbane, QLD 4111, Australia

Abstract: 音声知覚、音声認識において短時間位相スペクトルが重要な役割を持たないと言うことはこれまで常識とされてきた。本稿では、その音声知覚、音声認識における有用性を調査する。音声知覚実験を通し、32msec 長の窓で得られた短時間位相スペクトルがパワースペクトルと同様に音声明瞭度に貢献することを示す。位相スペクトルから導出された周波数上の表現形式を提案し、パワースペクトルから導出されたケプストラムと同等の性能が得られることを音声認識実験から明らかにする。

key words

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

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

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:

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

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.

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 im-

portance) in most of the speech processing applications (such as speech recognition [1, 2] and enhancement [3, 4]). It is perhaps due to our common understanding derived through psychoacoustic experiments (done as early as in the nineteenth century by Helmholtz [5]) that the human ear is almost insensitive to phase. Even the recent human perception studies [6] have indicated that the short-time phase spectrum (window duration of about 30 ms) conveys no information about the intelligibility of speech. In the current automatic speech recognition systems [1, 2], the cepstral features are the most commonly used features. These features are derived using only power spectrum (phase spectrum is totally ignored). Similarly, in speech enhancement systems [3], only power spectrum is enhanced; phase spectrum of noisy speech is left untouched.

In this paper, the usefulness of phase information is explored in human speech perception as well as in automatic speech recognition. Through human perception experiments, it is shown that the short-time phase spectrum (with widow size of 32 ms) contributes to speech intelligibility as much as the corresponding power spectrum. A representation based on frequencies of the speech signal derived from its short-time phase is developed and is found to be as good as cepstral representation for automatic speech recognition.

Short-time phase spectrum in human speech perception [7]

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 to retain either only phase information or only amplitude information.

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.,

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

This modified STFT is then used to synthesize the signal $\hat{s}(t)$ using the overlap-add method [8]. 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:

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

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 [6, 4]. 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 $T_w/4$ [8]. 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 and phase-only 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. We list in Table 1 our results averaged over the 12 subjects. We can make the following observations from this table: 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. 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.

表 1: Consonant intelligibility (or, identification accuracy) of magnitude-only and phase only utterances for Hamming and rectangular windows with window durations of 32 ms and 1024 ms.

Window type	Intelligibility (in %) for			
	magnitude-only		phase-only	
	32 ms	1024 ms	32 ms	1024 ms
Hamming	84.2	14.1	59.8	88.0
Rectangular	78.1	13.2	80.0	89.3

Short-time phase spectrum in automatic speech recognition [9]

As mentioned earlier, the cepstral features used in current speech recognition systems are obtained from the power spectrum $P(\nu, t)$. They do not use any information from the phase spectrum $\psi(\nu, t)$. In this paper, we propose to use frequency-related features derived from the short-time phase spectrum $\psi(\nu, t)$ for speech recognition. For this, a short-time instantaneous frequency (IF) spectrum is computed as follows [10]:

$$F(\nu, t) \quad = \quad \nu + \frac{1}{2\pi} \frac{d\psi(\nu, t)}{dt}. \qquad (5)$$

We use in the present paper this short-time IF spectrum for deriving the features for speech recognition. Note that this IF spectrum has been used in the past for extracting fundamental frequency [10, 11] and formants [12, 13].

Instead of using the STFT analysis, we use a procedure which employs a bank of bandpass filters for frequency decomposition. We describe below our procedure 1) for computing the short-time power and IF spectra, and 2) for extracting the frequency-related features from the short-time IF spectrum.

Consider that we are interested in telephone bandwidth speech signal from 200 Hz to 3400 Hz. We sample the frequency range uniformly on mel scale at $N = 200$ points. Using these frequency values as their center frequencies, design $N = 200$ bandpass filters with bandwidths equal to their respective critical bandwidths [1]. Our analysis procedure can be described in terms of the following steps:

- **Step 1:** Apply the speech signal $s(t)$ to each of the N bandpass filters. Let the output of the i-th bandpass filter be $s(\nu_i, t)$, where ν_i is the center frequency of the i-th bandpass filter. For illustration, we consider a speech signal corresponding to vowel /i/ and apply it to the i-th bandpass filter ($i = 139$) with center frequency $\nu_i = 1880$ Hz and bandwidth = 280 Hz, and its filtered output $s(\nu_i, t)$ is shown in Fig. 1(a).

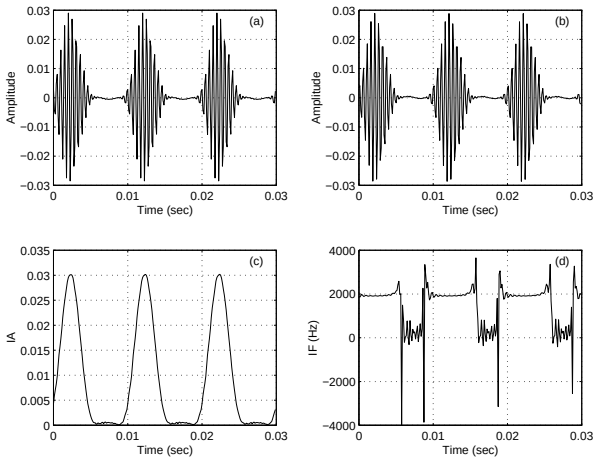

図 1: (a) Filtered signal $s(\nu_i, t)$ from the i-th bandpass filter with center frequency $\nu_i = 1880$ Hz and bandwidth = 280 Hz, (b) Hilbert transform $\hat{s}(\nu_i, t)$, (c) IA function $a(\nu_i, t)$, and (d) IF function $f(\nu_i, t)$.

- **Step 2:** For each filtered signal $s(\nu_i, t)$, compute the Hilbert transform $\hat{s}(\nu_i, t)$. This is shown in

Fig. 1(b).

- **Step 3:** For each filtered signal $s(\nu_i, t)$, construct an analytic signal

$$s_a(\nu_i, t) \quad = \quad s(\nu_i, t) + j\hat{s}(\nu_i, t), \tag{6}$$

and decompose it as follows:

$$s_a(\nu_i, t) \quad = \quad a(\nu_i, t)e^{j\phi(\nu_i, t)}, \tag{7}$$

where $a(\nu_i, t) = |s_a(\nu_i, t)|$ is the instantaneous amplitude (IA) of the filtered signal $s(\nu_i, t)$, and $\phi(\nu_i, t) = \angle s_a(\nu_i, t)$ the instantaneous phase. The instantaneous frequency (IF) $f(\nu_i, t)$ is computed from the instantaneous phase $\phi(\nu_i, t)$ as follows [14]:

$$f(\nu_i, t) = \frac{1}{2\pi}\frac{d\phi(\nu_i, t)}{dt}. \tag{8}$$

The IA and IF functions for the i-th filter ($i = 139$) are shown in Fig. 1(c) and 1(d), respectively[1].

- **Step 4:** For each filtered signal $s(\nu_i, t)$, compute a short-time power estimate from the IA function as follows:

$$P(\nu_i, t) = \frac{\int_{-\infty}^{\infty}[a(\nu_i, \tau)]^2 w(t-\tau)d\tau}{\int_{-\infty}^{\infty} w(t-\tau)d\tau}, \tag{9}$$

where $w(t)$ is a window function, similar to the one used in STFT. $P(\nu_i, t)$ as a function of ν_i provides an estimate of the short-time power spectrum for the frame centered at time t. This power spectrum is shown in Fig. 2(a).

- **Step 5:** It can be observed from Fig. 1 that the filtered signal $s(\nu_i, t)$ is band-limited (between 1740 and 2020 Hz), but its IF $f(\nu_i, t)$ is not confined within the band boundaries. This is quite counter-intuitive and a number of methods are reported in the literature to overcome this problem [13]. In the present paper, we handle this problem by observing the fact that the IF misbehaves only when the corresponding IA is low. Therefore, we use only those values of IF for computing the short-time IF estimate for which the corresponding IA is above certain threshold. For each filtered signal $s(\nu_i, t)$, we define a short-time IF estimate as follows:

$$F(\nu_i, t) = \frac{\int_{-\infty}^{\infty} f(\nu_i, \tau)\theta(\nu_i, \tau)w(t-\tau)d\tau}{\int_{-\infty}^{\infty} \theta(\nu_i, \tau)w(t-\tau)d\tau}, \tag{10}$$

where the threshold function $\theta(\nu_i, t)$ is defined as follows:

$$\theta(\nu_i, t) \quad = \quad \begin{cases} 0, & \text{if } a(\nu_i, t) \leq \Theta(\nu_i), \\ 1, & \text{otherwise.} \end{cases} \tag{11}$$

The value of threshold $\Theta(\nu_i)$ can be selected through experimentation. In our experiments, we have set it to the average value of $a(\nu_i, t)$ over the frame duration. $F(\nu_i, t)$ as a function of ν_i provides an estimate of the short-time IF spectrum for the frame centered at time t. This IF spectrum is shown in Fig. 2(b).

It can be observed from Fig. 2(b) that the short-time IF spectrum captures the formant structure in the form of flat (or, low slope) portions. That is, it shows flat regions where-ever there is formant activity in the power spectrum. Thus, it contains useful information for speech recognition. However, it is not clear how to use this spectrum to extract features for speech recognition. One possible method is to take the first derivative of this spectrum with respect to frequency ν_i and compute the cepstral coefficients from the resulting derivative through DCT. However, we have experimentally found this method to be unsatisfactory for speech recognition.

We use another method to derive recognition features from the short-time IF spectrum $F(\nu_i, t)$. In

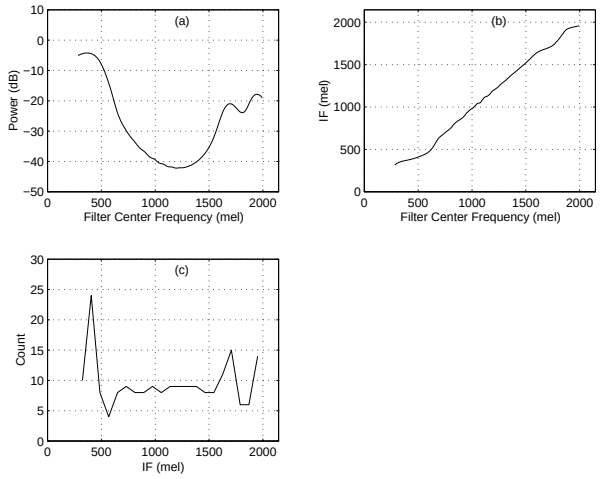

図 2: (a) Short-time power spectrum $P(\nu_i, t)$, (b) Short-time IF spectrum $F(\nu_i, t)$, and (c) Short-time IF histogram $H(F, t)$.

[1]We have used here the analytic signal decomposition method for computing the instantaneous frequency (IF) of a signal. However, there are other methods, such as the Teager energy method [15], reported in the literature for computing the IF.

this method, we completely ignore the information about the center frequencies ν_i of the bandpass filters and pool all the short-time IF values for the frame centered at time t. We use this pool of IF values to generate a histogram. We call it the short-time IF histogram and denote it by $H(F,t)$. This histogram is shown in Fig. 2(c). Note that this short-time IF histogram has got formant peaks similar to that in the short-time power spectrum shown in Fig. 2(a). We have carried out this type of frequency analysis for different vowel and consonant sounds of speech and observed that the short-time IF histogram $H(F,t)$ contains meaningful formant-like information about the speech signal.

- Step 6: We have seen that the short-time IF histogram $H(F,t)$ contains useful information for speech recognition. In order to use it for speech recognition, we parameterize it into cepstral coefficients through DCT. We call these cepstral coefficients the frequency-related features and use them for speech recognition.

In order to test the effectiveness of the short-time IF spectrum, we use a very simple multi-speaker vowel recognition system. The data base consists of 10 Hindi vowels spoken 30 times in /b/-V-/b/ context by three speakers (2 males and one female). Sampling rate of speech signal is 8 kHz. A 30 ms segment is excised from the central steady-state vowel portion of each utterance. We use 15 repetitions from each speaker for training the recognizer and the remaining 15 for testing. Thus, we have 450 vowel segments as training data and another 450 as test data. For the recognition experiments reported in this paper, we use only 10 bandpass filters uniformly spaced on mel frequency scale over the range of 200 Hz to 3400 Hz. ¿From each vowel segment, we extract 10 short-time IFs (using Eq. 10). These 10 IFs (called as mel frequency instantaneous frequencies (MFIFs)) form a feature vector for each vowel segment. For vowel recognition, we use a Bayesian classifier with the maximum posterior probability decision rule. We train our recognition system with clean speech, but test it on clean speech as well as on speech distorted by additive white noise with signal-to-noise ratio (SNR) of 20 dB. Recognition results are listed in Table 2. To provide comparison with features used in current speech recognition systems, we also provide in this table results obtained by using 10 linear prediction cepstral coefficients (LPCCs) and 10 mel-frequency cepstral coefficients (MFCCs). It can be seen from this table that the MFIF features provide recognition results comparable to the LPCC and MFCC features.

表 2: Speech recognition performance of the LPCC, MFCC and MFIF features in presence of additive noise distortion.

SNR	Recognition accuracy (in %)		
(dB)	LPCC	MFCC	MFIF
∞	80.9	80.4	78.7
20	62.4	68.4	69.5

Conclusions

In this paper, the usefulness of phase information is explored in human speech perception as well as in automatic speech recognition. Through human perception experiments, it is shown that the short-time phase spectrum (with widow size of 32 ms) contributes to speech intelligibility as much as the corresponding power spectrum. A representation based on frequencies of the speech signal derived from its short-time phase is developed and is found to be as good as cepstral representation (derived from power spectrum) for automatic speech recognition.

Acknowledgment

The author wishes to thank Dr. B.S. Atal and Mr. L. Alsteris for fruitful collaboration. This paper is based on two papers [9, 7] coauthored with these collaborators.

参考文献

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

[2] S. Young, "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, Vol. 13, pp. 45-57, Sept. 1996.

[3] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", Proc. IEEE, Vol. 67, pp. 1586-1604, 1979.

[4] D.L. Wang and J.S. Lim, "The unimportance of phase in speech enhancements", IEEE Trans. Acoust., Speech and Signal Process., Vol. 30, pp. 679-681, Aug. 1982

[5] H. Helmholtz, *On the Sensations of Tone*, Dover, New York, 1954.

[6] 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.

[7] K.K. Paliwal and Leigh Alsteris, "On the importance of short-time phase spectrum in speech perception", paper under preparation.

[8] 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.

[9] K.K. Paliwal and B.S. Atal, "Representing frequencies in speech", Techn. Report, AT&T Research Labs., Florham Park, NJ, Jan. 2000.

[10] T. Abe, T. Kobayashi and S. Imai, "Robust pitch estimation with harmonic enhancement in noisy environments based on instantaneous frequency", *Proc. ICSLP*, pp. 1277-1280, 1996.

[11] 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.

[12] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation", *J. Acoust. Soc. Am.*, Vol. 99, pp. 3795-3806, 1996.

[13] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", *J. Acoust. Soc. Am.*, Vol. 105, pp. 1912-1924, 1999.

[14] L. Cohen, "Time-frequency analysis – A review", *Proc. IEEE*, Vol. 77, pp. 941-981, 1989.

[15] P. Maragos, J.F. Kaiser and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Signal Processing*, Vol. 41, pp. 3024-3051, 1993.