FREQUENCY-RELATED REPRESENTION OF SPEECH

Kuldip K. Paliwal †

School of Microelectronic Engineering
Griffith University, Brisbane, Australia

Bishnu S. Atal

AT&T Research Labs.
Florham Park, NJ-07932, USA

ABSTRACT

Cepstral features derived from power spectrum are widely used for automatic speech recognition. Very little work, if any, has been done in speech research to explore phase-based representations. In this paper, an attempt is made to investigate the use of phase function in the analytic signal of critical-band filtered speech for deriving a representation of frequencies present in the speech signal. Results are presented which show the validity of this approach.

1. INTRODUCTION

Currently, the cepstral features are the most commonly used features for speech recognition. These features are derived from the power spectrum of the speech signal. Some of the problems associated with the cepstral features are as follows: 1) Only half of the speech information, power spectrum, is used in their computation. The other half, phase spectrum, is not used. 2) Power spectrum gets affected by the environmental conditions (background noise and channel distortions). Therefore, the cepstral features are not robust with respect to these environmental variations.

In this paper[1], we investigate a representation based on frequencies of the speech signal derived from its phase. In our research work, we measure instantaneous frequencies (IFs) from the phase of the critical-band filtered speech signal and create a representation of measured frequency as a function of the center frequency of the critical-band filters. This representation can be used in different ways for robust speech recognition. One way is to compute a histogram of the measured frequencies and use it (or, its parametric representation) for speech recognition. Note that IF based speech analysis has been used in the past for pitch and formant extraction [3, 4, 5, 6].

† This work was partly supported by ARC (Discovery) grant (No. DP0209283).

[1]This paper is an extended version of our technical report [1]. Some of the results have been presented earlier in a conference [2].

2. REPRESENTATION IN TERMS OF INSTANTANEOUS AMPLITUDES AND FREQUENCIES

Consider a signal $s(t)$. Compute its analytic signal

$$s_a(t) \quad = \quad s(t) + j\hat{s}(t), \tag{1}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$.

We can decompose $s_a(t)$ as follows:

$$s_a(t) \quad = \quad a(t)e^{j\phi(t)}, \tag{2}$$

where

$$a(t) \quad = \quad |s_a(t)| \tag{3}$$

is called the instantaneous amplitude (or envelope) of the signal, and

$$\phi(t) \quad = \angle s_a(t) \tag{4}$$

the phase. The instantaneous frequency (IF) $f(t)$ is computed from the phase $\phi(t)$ as follows:

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt}. \tag{5}$$

The signal is completely represented in terms of its instantaneous amplitude and frequency.

We have used this analytic signal decomposition method for computing the instantaneous frequency (IF) of a signal. However, there are other methods, such as the Teager energy method [7], reported in the literature for computing the IF.

Though the signal $s(t)$ is band-limited, its IF $f(t)$ is not. In fact, the IF values can fluctuate from $-\infty$ to ∞. This is a serious problem with IF. Because of this, the interpretation of IF has been a subject of investigation and debate for years. It still is.

In order to illustrate this problem, we take a frame (of duration $T = 30$ ms) of the vowel sound /i/. We compute the analytic signal and filter it through a 1740–2020 Hz critical-band filter. The real and imaginary parts of the filtered signal are shown in Fig. 1. Its instantaneous amplitude and frequency are shown in Fig. 2.

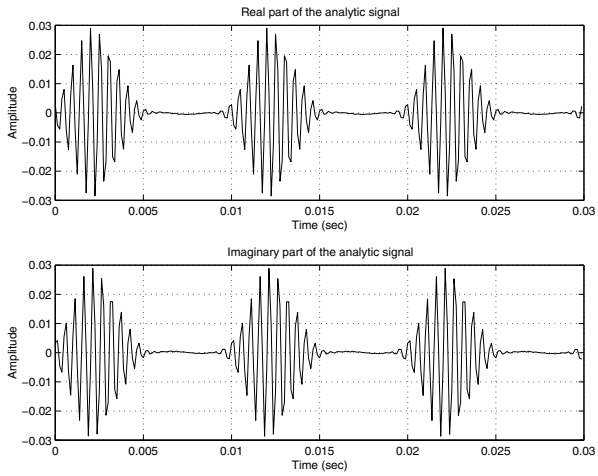

Fig. 1. Real and imaginary parts of the 1740–2020 Hz critical-band filtered speech of vowel sound /i/.

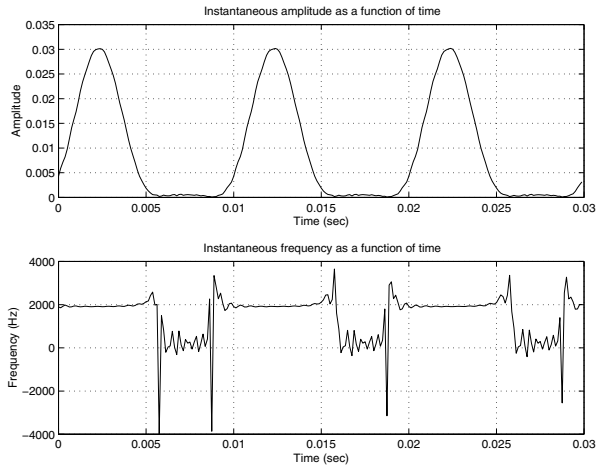

Fig. 2. Instantaneous amplitude and frequency of the 1740–2020 Hz critical-band filtered speech of vowel sound /i/.

It can be seen from Fig. 2 that IF varies a lot. It is not confined within the frequency band 1740–2020 Hz. In addition, it takes negative values. In order to get a meaningful average value of frequency over a frame, we make the observation from Fig. 2 that the measured IF generally misbehaves when the instantaneous amplitude is low. We use only those values of IF for averaging where the instantaneous amplitude is above certain threshold. The resulting average value (called as the thresholded IF) is defined as follows:

$$F_t = \frac{\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\theta(t)dt}{\int_{-\frac{T}{2}}^{\frac{T}{2}} \theta(t)dt} \tag{6}$$

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

$$\theta(t) \quad = \quad \begin{cases} 0, & \text{if } a(t) \leq \Theta \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

The threshold Θ used in our experiments is set to average value of $a(t)$ over the frame duration.

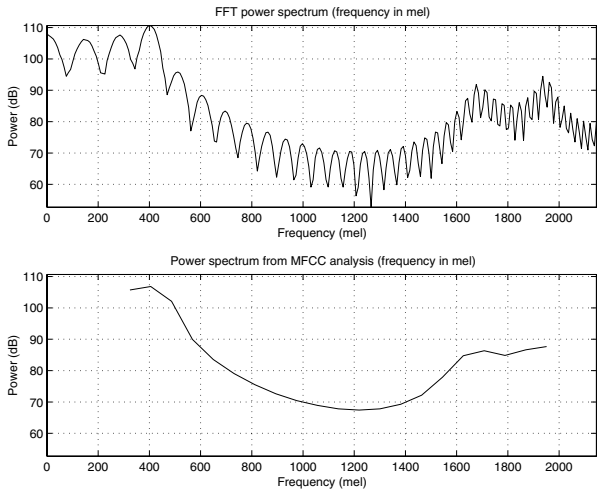

Fig. 3. (a) Power spectrum and (b) Triangular filter-bank power spectrum on mel scale for the vowel sound /i/.

3. SPEECH ANALYSIS

We can outline our speech analysis procedure as follows:

- Step 0: Consider that we are interested in telephone bandwidth speech signal from 200 Hz to 3400 Hz. Sample the frequency range uniformly on mel scale. Using these frequency values as their center frequencies, design $N = 200$ bandpass filters with bandwidths equal to their respective critical bandwidths.
- Step 1: Given the speech signal $s(t)$, $0 < t < T$, for a given frame, construct its analytic signal $s_a(t)$.
- Step 2: Filter $s_a(t)$ through N bandpass filters.
- Step 3: For each of these filtered speech signals, compute instantaneous amplitude $a(t, \nu)$ and IF $f(t, \nu)$, where ν is the center frequency of a given filter.
- Step 4: Compute the average IF estimate $F(\nu)$ using the thresholded definition with a window duration of $T = 30$ ms. Also, compute the mean square (MS) value ($A(\nu)$) of $a(t, \nu)$ for the frame.
- Step 5: $A(\nu)$ as a function of ν contains speech information similar to that in power spectrum. $F(\nu)$ provides a spectrum of IF F as a function of ν.

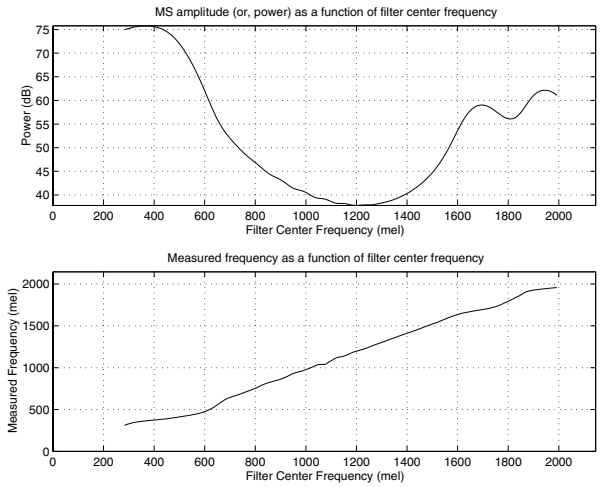

Fig. 4. (a) MS amplitude A and (b) Measured frequency F as a function of critical-band filter's center frequency ν for the vowel sound /i/.

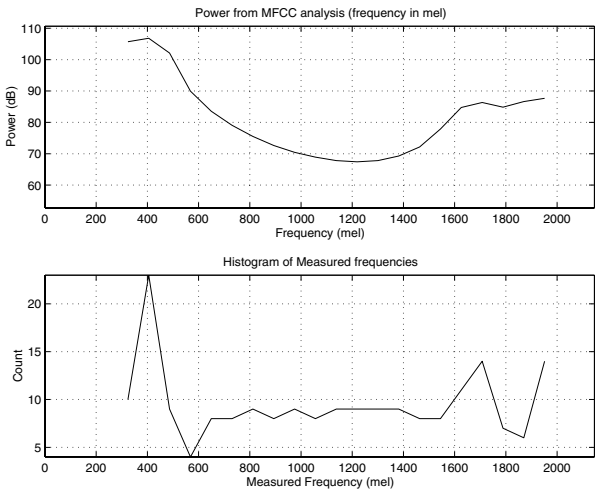

Fig. 5. (a) Triangular filter-bank power spectrum on mel scale and (b) Histogram of measured frequency F for the vowel sound /i/.

- Step 6: In order to use $F(\nu)$ information for speech recognition, compute the histogram of the measured frequency F. This (or, its parameters) are used as features for speech recognition.

In order to illustrate the results of our speech analysis procedure, we use the same speech signal as used in Fig. 1. In Fig. 3(a), we show the power spectrum on a mel-warped frequency scale. A triangular filter-bank power spectrum used in conventional cepstral analysis for mel-frequency cepstral coefficients (MFCCs) is shown in Fig. 3(b).

In Fig. 4(a), we show MS amplitude A (computed in dB from the instantaneous amplitude $a(t)$) as a function of critical-band filter's center frequency ν. By comparing Fig. 4(a) with Fig. 3(b), we can see that the MS amplitude spectrum contains information similar to that in the MFCC power spectrum.

In Fig. 4(b), we show the (measured) average frequency F as a function of critical-band filter's center frequency ν. If the power spectrum were flat (i.e., it had no information), this plot would have looked like a diagonal line. However, for the speech signal, it shows flat regions where-ever there is formant activity in the power spectrum. This plot is the basic source of speech information contained in the measured frequencies derived from the phase of the critical-band filtered speech signals. It can be used in various ways for speech recognition. One particular way is to compute its histogram. It is shown in Fig. 5(b). The peaks in this histogram correspond to the formants. We have carried out this type of frequency analysis for different vowel and consonant sounds of speech and observed that the frequency distribution (histogram) contains meaningful information about the speech signal. This can be used either directly or in some parametric form for speech recognition.

In order to illustrate its robustness to the additive background noise distortion, we add white Gaussian noise to the speech signal. The amount of noise is adjusted such that the signal-to-noise ratio (SNR) is 15 dB. We show in Fig. 6(b) the histogram of measured frequencies for this noisy signal. By comparing this figure with Fig. 5(b), we can see that this histogram does not change much by the additive noise distortion and it is more robust to additive noise than the power spectrum.

In Fig. 7(b), we show the the histogram of measured frequencies for the channel-distorted speech signal. We use differencing of speech signal to simulate the channel distortion effect. Again, we observe that the frequency histogram is more robust to channel distortion than the power spectrum.

4. 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. These 10 IFs (called as mel frequency instan-

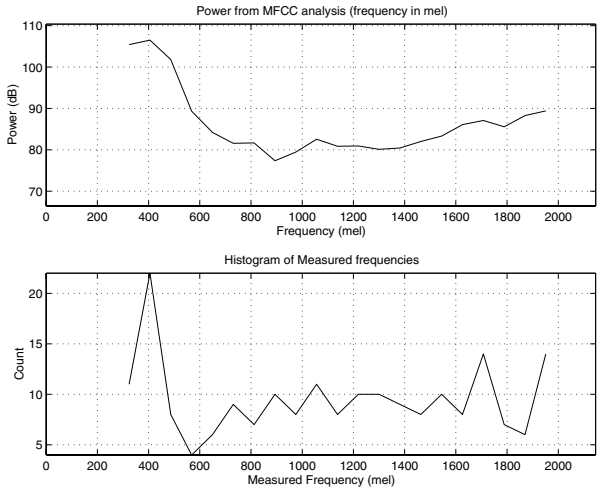

Fig. 6. (a) Triangular filter-bank power spectrum on mel scale and (b) Histogram of measured frequency F for the noisy vowel sound /i/ with SNR $= 15$ dB.

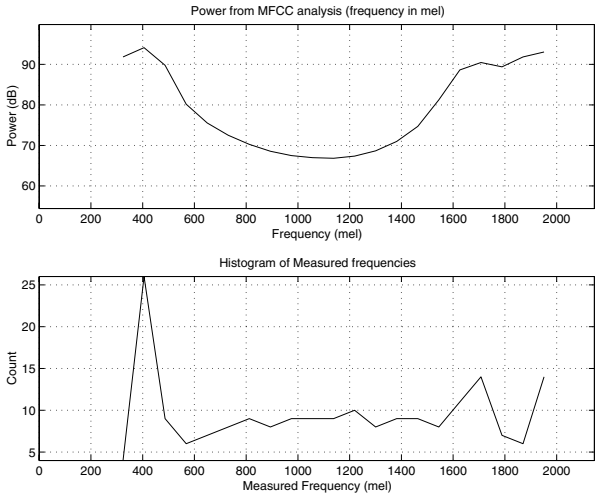

Fig. 7. (a) Triangular filter-bank power spectrum on mel scale and (b) Histogram of measured frequency F for the channel distorted vowel sound /i/.

taneous 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 1. 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.

Table 1. 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

5. CONCLUSIONS

In this paper, we have developed an analysis procedure which can be used compute frequency related features from the phase of the speech. We have shown that these features contain meaningful linguistic information. We have utilized them successfully for speech recognition.

6. REFERENCES

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

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

[3] T. Abe, T. Kobayashi and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency", *Proc. ICASSP*, pp. 756-759, 1995.

[4] 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.

[5] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM-FM modulation model", *Speech Communication*, Vol. 28, pp. 195-209, 1999.

[6] 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.

[7] 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.