Split-band Perceptual Harmonic Cepstral Coefficients as Acoustic Features for Speech Recognition

Liang Gu and *Kenneth Rose*

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

ABSTRACT

This paper presents a significant modification of our previously proposed speech recognizer's front-end based on perceptual harmonic cepstral coefficients. The spectrum is split into two frequency bands, which correspond to the harmonic and non-harmonic components. A weighting function, which depends both on the voiced/unvoiced/ transitional classification and on the prominence of harmonic structures, is applied to the harmonic band, and ensures accurate representation of the voiced and transitional speech spectral envelope. Conventional smoothed spectrum is used in the non-harmonic band. The mixed spectrum undergoes mel-scaled band-pass filtering, and the log-energy of the filters' output is discrete cosine transformed to produce cepstral coefficients. Experiments with Mandarin digit and E-set databases show significant recognition gains over plain perceptual harmonic cepstral coefficients and considerable gains over standard techniques.

1. INTRODUCTION

A main difficulty that plagues conventional spectral acoustic analysis concerns the vocal tract transfer function whose accurate description is crucial to effective speech recognition [1]. While the information on the vocal transfer function could be extracted from the gross spectral shape of the input speech, the irrelevant information of excitation signals must be removed for accurate spectral representation. One approach is to estimate a smoothed version of the short-time spectrum by mel-scaled band-pass filters, which results in the well-known mel-frequency cepstral coefficients (MFCC) [2]. Although such a procedure is fast and efficient, it is sub-optimal as the vocal tract transfer function information is known to reside in the spectral envelope, which is mismatched with the smoothed spectrum, especially for voiced speech. Consequently, major gains may be recouped by alternative approaches based on direct spectral envelope estimation [3][4].

Another difficulty encountered in conventional acoustic analysis is that of appropriate spectral amplitude transformation for better recognition performance. The logarithmic power spectrum representation in MFCC has obvious merits due to its gain-invariance properties and the approximate Gaussian distributions it thus provides. Cubic root representation is used in the perceptual linear prediction (PLP) representation [5] for psychophysical considerations, at the cost of compromising the level-invariance properties and hence robustness.

Recently, we proposed a new approach to overcome the above shortcomings [6] and investigated extensions for noisy environments [7]. Rather than average the energy within each filter, the harmonic cepstral coefficients were derived from a spectrum envelope estimate which was weighted at harmonic locations for voiced and transitional speech. They were left similar to MFCC for unvoiced speech and silence. The intensity-loudness power-law was applied within each filter, along with logarithmic energy across filters, to reduce the spectral amplitude variation within each filter without degradation of the gain-invariance properties. The resulting acoustic features formed the perceptual harmonic cepstral coefficients (PHCC) representation. Experiments showed that PHCC significantly outperformed conventional MFCC under both clean [6] and noisy speech environments [7].

In this paper, the new *split-band PHCC* (SB-PHCC) approach is proposed to enhance and extend PHCC via split-band spectral analysis. The speech spectrum is split, at a cutoff frequency, into two spectral bands corresponding to harmonic and non-harmonic components. The harmonic weighted spectrum is used in the harmonic band, and traditional smoothed spectrum is adopted for the non-harmonic band. The cutoff frequency selection is optimized by maximizing the average voicing strength ratio of harmonic to non-harmonic bands. Experiments with Mandarin digit and E-set databases show that SB-PHCC significantly outperforms plain PHCC and yields greater gains over conventional MFCC.

2. PERCEPTUAL HAMONIC CEPSTRAL COEFFICIENTS

PHCC employs a framework similar to MFCC except that it attempts to closely approximate the perceptually compressed spectral envelope weighted at pitch harmonics. The procedure consists of the following steps:

1) The speech frame is processed by DFT to obtain the short-time power spectrum;

2) The intensity-loudness power law is applied to the original spectrum to obtain the root-power compressed spectrum;

3) Robust pitch estimation and voiced/unvoiced/transition (V/UV/T) classification are performed (We employ the spectro-temporal auto-correlation (STA) [8] followed by a peak-picking algorithm [6]);

4) Class-dependent harmonic weighting is applied to obtain the harmonics weighted spectrum (HWS). For voiced and transitional speech, HWS is dominated by the harmonic spectrum (i.e. upper envelope of the short-term spectrum). For unvoiced sounds, HWS degenerates to the conventional smoothed spectrum.

This work was supported in part by the National Science Foundation under grants no. IIS-9978001, EIA-9986057, the University of California MICRO Program, Conexant Systems, Inc., Dolby Laboratories, Inc., Lucent Technologies, Inc., Medio Stream, Inc., and Qualcomm, Inc.

5) Mel-scaled filters are applied to the HWS and the log energy output is computed and transformed into cepstrum by the discrete cosine transform (DCT).

More details of PHCC can be found in [6].

The advantage of PHCC over conventional acoustic analysis methods is mainly attributed to its spectral envelope estimation. It is widely recognized in the speech coding community that it is the spectral envelope and not the gross spectrum that represents the shape of the vocal tract [9]. However, spectral envelope estimation may greatly reduce the representation accuracy and robustness in the case of non-harmonic sounds. In our early PHCC approach [6], the possible distortion due to spectral envelope extraction was mitigated by effective V/UV/T detection. Nevertheless, significant distortion was observed in voiced and transitional speech since the spectral envelope was estimated by HWS throughout the frequency domain. While HWS performs well in the harmonic region of the speech spectrum, it tents to impart an undesirable effect to noise-like non-harmonic regions and hence reduce robustness. To overcome this drawback, we propose the split-band PHCC (SB-PHCC), in which spectral envelope extraction is restricted to the harmonic band where the harmonic structure is rich and reliable, while conventional smoothed spectral estimation is applied to the non-harmonic band for higher representation robustness and accuracy.

3. SPLIT-BAND PHCC

A flowchart of the SB-PHCC algorithm is shown in Figure 1. The speech signal undergoes discrete Fourier transformation, followed by root-power compression, as in plain PHCC. However, in SB-PHCC the STA algorithm is not only adopted for robust V/UV/T detection and pitch estimation, but also for split-band analysis by computing three new parameters, namely, harmonic confidence, voicing strength and cutoff frequency, which reflect the prominence of harmonic structures in the speech spectrum. These parameters, as well as the peak-picked harmonic locations, form the basis of split-band HWS estimation. The extracted mixed spectrum passes through mel-scaled band-pass filters, followed by discrete cosine transform, and results in the split-band perceptual harmonic cepstral coefficients.

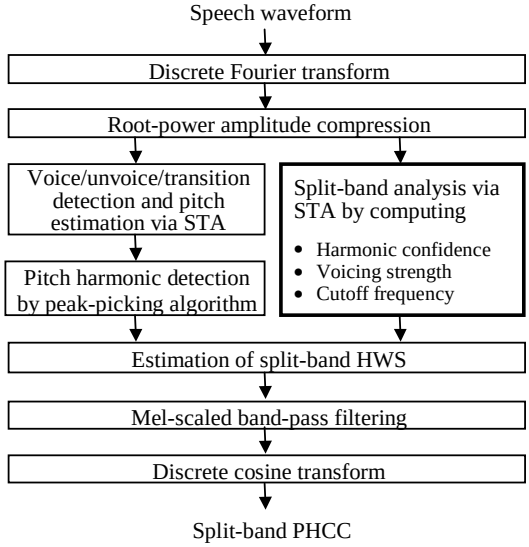

Figure 1. Flowchart of split-band perceptual harmonic cepstrum coefficient (SB-PHCC) analysis

Next, we describe the SB-PHCC procedure in more detail.

A. Spectro-temporal autocorrelation (STA)

Robust pitch estimation is critical for harmonic-based spectral envelope representation. Although small errors could be corrected by a peak-picking algorithm as described in [9], pitch multiple and sub-multiple errors can greatly reduce the accuracy of the spectral envelope for voiced sounds. One effective approach to eliminate such errors is the STA algorithm, which was first proposed for design of harmonic speech coders [8]. In this paper, STA is further harnessed to measure the harmonic characteristics of the speech spectrum, via the computation of three new parameters that will be defined in subsection 3.C.

Given a speech signal $s_t(n)$, the temporal auto-correlation (TA) for candidate pitch τ is defined as

$$R_{TA}(\tau)=\frac{\sum_{n=0}^{N-\tau-1}[\tilde{s}_t(n)\cdot\tilde{s}_t(n+\tau)]}{\sqrt{\sum_{n=0}^{N-\tau-1}\tilde{s}_t^2(n)\cdot\sum_{n=0}^{N-\tau-1}\tilde{s}_t^2(n+\tau)}}$$

where $\tilde{s}_t(n)$ is the zero-mean version of $s_t(n)$, and N is the number of samples for pitch estimation.

Motivated by the pitch multiple errors that were observed in the conventional TA method, the spectral auto-correlation (SA) criterion was introduced and defined as

$$R_{SA}(\tau)=\frac{\int_0^{\pi-\omega_\tau}\tilde{S}_f(\omega)\tilde{S}_f(\omega+\omega_\tau)d\omega}{\sqrt{\left[\int_0^{\pi-\omega_\tau}\tilde{S}_f^2(\omega)d\omega\right]\left[\int_0^{\pi-\omega_\tau}\tilde{S}_f^2(\omega+\omega_\tau)d\omega\right]}}$$

where $\omega_\tau=2\pi/\tau$, $S_f(\omega)$ is the power spectrum of $s_t(n)$, and $\tilde{S}_f(\omega)$ is the zero-mean version of $S_f(\omega)$. However, pitch sub-multiple may occur in SA. STA was devised to reduce both pitch multiple and sub-multiple errors, and is defined as:

$$R_{STA}(\tau)=\beta\cdot R_{TA}(\tau)+(1-\beta)\cdot R_{SA}(\tau) \quad (1)$$

where $\beta=0.5$ was reported to yield the best results in [8].

B. Harmonic weighted spectrum (HWS)

Spectral envelope representation as above has been previously proposed and is currently widely used in harmonic speech coding [9], and more recently in speech recognition [3][4]. However, the number of harmonics varies significantly from speaker to speaker. This suggests that some processing must be applied to the harmonic spectrum prior to its utilization to effective speech recognition. If the speech spectrum between pitch harmonics is smooth, interpolation or normalization methods [3] can be used to retrieve the spectrum spline, albeit with high sensitivity to pitch estimation errors. Instead, we proposed an approach called harmonic weighted spectrum

(HWS) estimation [6]. Given $S_f(\omega)$, the magnitude spectrum of input speech, HWS is defined as

$$HWS(\omega) = w_h(\omega) \cdot S_f(\omega)$$

where $w_h(\omega)$ is the harmonic weighting function which was originally defined in [6] as

$$w_h(\omega) = \begin{cases} W_H, & \omega \text{ is pitch harmonic} \\ 1, & otherwise \end{cases}, \quad (2)$$

where W_H was adjusted depending on the V/UV/T classification. It was set to a high value for voiced sounds and intermediate value for transitional sounds. The harmonic weighting function is modified in this paper, as will be explained next.

C. Split-band Analysis

The PHCC harmonic weighting function (2) does not take into account the distortion of spectral envelope estimation at non-harmonic locations for both voiced and transitional speech. A split-band analysis is hence proposed here to eliminate this drawback. The underlying premise of this technique is that there exists a single transition frequency (the voicing cutoff frequency) below which the harmonic structure is rich and clear, and above which the spectrum is essentially non-harmonic. Therefore, for voiced and transitional sounds, the original spectrum is split into two bands – the (low frequency) harmonic band and the (high frequency) non-harmonic band. Given the differing characteristics of the two bands, potential gains are expected if they are treated separately. In the proposed SB-PHCC, HWS is implemented in the harmonic band, while MFCC is used in the non-harmonic band. Thus, the accuracy of the spectral envelope representation is maintained by harmonic-weighted spectral estimation, while the noise-sensitivity in the non-harmonic band is reduced by the smoothing procedure, where no harmonic-based analysis is necessary.

To carry out the split-band analysis, three new parameters are defined and computed to measure the prominence of the harmonic structures observed in the speech spectrum.

The prominence of the harmonic structure over the full-band may be measured by the *harmonic confidence*, which is defined as

$$H_a = \max_{\tau} R_{STA}(\tau),$$

where $R_{STA}(\tau)$ is the STA defined in (1).

The prominence of the harmonic structure about frequency Ω can be measured by the *voicing strength*, which is defined as

$$V_S(\Omega) = \frac{\int_{\Omega-\omega_0}^{\Omega} \tilde{S}_f(\omega)\tilde{S}_f(\omega+\omega_0)d\omega}{\sqrt{\left[\int_{\Omega-\omega_0}^{\Omega} \tilde{S}_f^2(\omega)d\omega\right]\left[\int_{\Omega}^{\Omega+\omega_0} \tilde{S}_f^2(\omega)d\omega\right]}}$$

where ω_0 is the fundamental frequency.

The boundary between harmonic band and non-harmonic band is specified by a *voicing cutoff frequency*. The voicing cutoff frequency is recognized as an important quantity in speech coding, where a number of relevant techniques have been developed [10][11]. Here we propose an algorithm based on average voicing strength ratio between the harmonic band and non-harmonic band, which can be described as

$$\omega_T = \arg \max_{\omega_{T_l} < \omega < \omega_{T_h}} \frac{\left[\int_{\omega_{T_l}}^{\omega} V_s(\Omega)d\Omega\right] \Big/ (\omega - \omega_{T_l})}{\left[\int_{\omega}^{\omega_{T_h}} V_s(\Omega)d\Omega\right] \Big/ (\omega_{T_h} - \omega)} \quad (3)$$

where ω_{T_l} and ω_{T_h} delimit the allowed interval for the cutoff frequency. In our experiment, we set ω_{T_l} = 2000π and ω_{T_h} = 6000π.

We hence propose a new harmonic weighting function, which is substantially different from the one we used in plain PHCC [6]. The SB-PHCC harmonic weighing function is defined as

$$w_h(\omega) = \begin{cases} \max\left(1, e^{(H_a-\eta)\cdot\gamma}\right), & \text{if } \omega \le \omega_T \text{ is pitch harmonic} \\ 1, & otherwise \end{cases},$$

where η is the harmonic confidence threshold, γ is the weight factor, and ω_T is the cut-off frequency. For voiced sounds, ω_T is obtained from (3). For transitional sounds, ω_T is fixed due to the reduced reliability of (3) which is compromised by low average voicing strength in the harmonic band. In our experiments, η and γ are set to 0.5 and 10, respectively, and ω_T is set to 4000π for transitional sounds.

D. Within-filter amplitude compression

The perceptual amplitude compression procedure we developed for plain PHCC [6] is applied in SB-PHCC to reduce amplitude variation, and is summarized here for completeness.

It is widely recognized that auditory properties can be exploited to improve automatic speech recognition. One example is the perceptual transformation of the spectrum amplitude, which is handled in radically different ways by the leading acoustic analysis systems. PLP applies the equal-loudness curve and the intensity-loudness power law to better exploit knowledge about the auditory system [5], but requires scale normalization, which was experimentally found to have a critical impact on the overall recognition performance. MFCC sacrifices some perceptual precision and circumvents this difficulty by approximating the auditory curve with a logarithmic function that offers the elegant level-invariance properties.

In an effect to enjoy the best of both approaches, we apply the intensity-loudness power-low within each filter and compute the log energy over all filters. Hence,

$$\hat{S}(\omega) = [S(\omega)]^{\theta}$$

$$\hat{E}_i = \log(E_i), \quad 1 \le i \le M$$

where $\hat{S}(\omega)$ is the compressed spectrum, $\hat{E}_i$ is the log energy for band-pass filter i, and θ is the root-power factor. The resulting spectrum representation can significantly reduce the amplitude variation within each filter, without degradation of

the gain-invariance properties and, since the filter energy levels are still represented in logarithmic scale, without recourse to normalization.

The cubic-root amplitude compression (θ =1/3) selected in [5] was found to perform best in our clean speech experiment [6]. It was, however, not optimal in our noise speech experiment [7]. Instead, we vary θ with SNR to achieve improve performance (θ is set to 2/3 for very low SNR).

4. EXPERIMENTAL RESULTS

To test the performance of SB-PHCC, experiments were first carried out on a database of speaker-independent isolated Mandarin digits collected in an office environment. The recognition task consists of 11 pronunciations representing 10 Mandarin digits from 0 to 9, with 2 different pronunciations for the digit "1" ([i] and [iao]). The database includes 150 speakers (75 male and 75 female), one utterance per speaker. Of the 150 speakers, 60 male and 60 female speakers were selected at random for training, and the remaining 30 speakers were set aside for the test set.

In our experiments, 39-dimension speech features were used, including 12 cepstral parameters, log energy, and their first-order and second-order dynamics (time derivatives). We used an analysis frame of width 30ms and step of 10ms, and a Hamming window. 9-state continuous density HMM was used with single Gaussian pdf per state. The experimental results for MFCC, PHCC and SB-PHCC are summarized in Table 1. It shows substantial decrease in error rate from MFCC, through PHCC, to SB-PHCC, for both male and female speakers.

To further test the performance of SB-PHCC on unvoiced sounds, additional experiments were carried out on OGI's E-set database. The recognition task is to distinguish between nine highly confusable English letters {b, c, d, e, g, p, t, v, z}, where the vowels are of minimal significance to the classification task. The database was generated by 150 speakers (75 male and 75 female) and includes one utterance per speaker. The experimental results are summarized in Table 2. SB-PHCC achieved consistently better results than PHCC over a range of acoustic model complexities, and offers over 15% error reduction relative to MFCC.

Speaker Gender	Male	Female	Male & Female
MFCC	0.6 %	3.9 %	2.9 %
PHCC	0.4 %	2.4 %	1.8 %
SB-PHCC	0.3 %	1.9 %	1.4 %

Table 1. Test-set error rate of MFCC, PHCC and SB-PHCC on isolated Mandarin digit recognition

Acoustic Models	7-state CHMM	13-state CHMM	21-state TMHMM
MFCC	15.3%	11.0 %	7.3 %
PHCC	12.2 %	9.0 %	6.2 %
SB-PHCC	11.3 %	8.5 %	5.8 %

Table 2. Test-set error rate of MFCC, PHCC and SB-PHCC on the E-set

5. CONCLUSION

The perceptual harmonic cepstral coefficients (PHCC) were previously proposed as promising acoustic features for speech recognition. The spectral envelope is represented via a weighted version of the power spectrum that emphasizes pitch harmonics, where the weighting function depends on the voice/unvoice/transition classification. A new split-band PHCC (SB-PHCC) is proposed to reflect the spectral distribution of harmonic structure in the design of the weighting function. The speech spectrum is split into harmonic and non-harmonic bands, at the voicing cutoff frequency whose selection is determined by a maximum average voicing strength ratio criterion. The harmonic weighted spectrum is applied in the harmonic band, while smoothed spectrum is used in the non-harmonic band to increase robustness. Experiments on the Mandarin digit and E-set speech databases show significant performance improvement of SB-PHCC over PHCC and hence over MFCC.

6. REFERENCES

[1] M. J. Hunt, "Spectral signal processing for ASR", *Proc. ASRU'99*, Dec. 1999.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 357-366, vol. 28, Aug. 1980.

[3] H. K. Kim and H. S. Lee, "Use of spectral autocorrelation in spectral envelope linear prediction for speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, pp.533-541, 1999.

[4] Q. Zhu and A. Alwan, "AM-demodulation of speech spectra and its application to noise robust speech recognition", *Proc. ICSLP'2000*, Oct. 2000.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust.. Soc. America*, pp. 1738-1752, vol. 87, no. 4, Apr. 1990.

[6] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients as the front-end for speech recognition", *Proc. ICSLP'2000*, Oct. 2000.

[7] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment", *Proc. ICASSP'2001*, May. 2001.

[8] Y. D. Cho, M. Y. Kim and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination", *Proc. ICASSP'98*, pp. 601-604, 1998.

[9] M. Jelinek and J. P. Adoul, "Frequency-domain spectral envelope estimation for low rate coding of speech", *Proc. ICASSP'99*, pp. 253-256, 1999.

[10] D. W. Griffin and J. S. Lim, "Multiband Excitation Coder", *IEEE Trans. ASSP*, vol. 36, pp.1223-1235, 1988.

[11] E. K. Kim and Y. H. Oh, "New analysis method for harmonic plus noise model based on time-domain periodicity score", *Proc. ICSLP*, 2000.