Perceptual Harmonic Cepstral Coefficients for Speech Recognition in Noisy Environment

Liang Gu and *Kenneth Rose*

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

ABSTRACT

Perceptual harmonic cepstral coefficients (PHCC) are proposed as features to extract from speech for recognition in noisy environments. A weighting function, which depends on the prominence of the harmonic structure, is applied to the power spectrum to ensure accurate representation of the voiced speech spectral envelope. The harmonics weighted power spectrum undergoes mel-scaled band-pass filtering, and the log-energy of the filters' output is discrete cosine transformed to produce cepstral coefficients. Lower spectral clipping is applied to the power spectrum, followed by within-filter root-power amplitude compression to reduce amplitude variation without compromise of the gain invariance properties. Experiments show significant recognition gains of PHCC over MFCC, with 23% and 36% error rate reduction for the Mandarin digit database in white and babble noise environments.

1. INTRODUCTION

Noise robust feature extraction poses one of the greatest challenges in the design of high performance automatic speech recognition systems. While most feature extraction techniques attempt to capture information on the vocal tract transfer function from the gross spectral shape of the input speech, the accuracy and robustness of the speech representation may deteriorate dramatically due to the spectral distortion caused by the additive background noise. The well-known mel-frequency cepstral coefficients (MFCCs) [1], though adopted by most ASR systems for its superiority in clean speech recognition, do not cope well with noisy speech. The alternative perceptual linear prediction (PLP) coefficients [2] promise improvement over MFCC in noisy conditions by incorporating perceptual features of the human auditory mechanism. Nevertheless, it is believed that the existing front ends are sub-optimal, and the discovery of new noise-immune or noise-insensitive features is anticipated.

One main difficulty in conventional spectrum-based feature extraction algorithms is concerned with the vocal tract transfer function whose accurate and robust description is crucial to effective speech recognition. In the MFCC approach, a smoothed version of the short-term speech spectrum is computed from the output energy of a bank of filters. While such a procedure is fast and efficient, it is inaccurate 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 and transitional speech. Moreover, the spectrum envelope tends to have much higher SNR than smoothed spectrum under the same noise conditions, which leads to a more robust representation of the vocal tract transfer function. Hence, speech features derived from the spectral envelope are expected to provide better performance in noisy environments compared with traditional front ends based on smoothed spectrum [3].

Another difficulty encountered in conventional feature extraction algorithms is that of appropriate spectral amplitude transformation for higher recognition accuracy and robustness. The log power spectrum representation in MFCC is clearly attractive because of its gain-invariance properties and the approximate Gaussian distributions it thus produces. Root power representation is used in PLP 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 [4]. Rather than average the energy within each filter, the harmonic cepstral coefficients were derived from the spectrum envelope sampled at the harmonic locations for voiced and transitional speech. They were similar to MFCC for unvoiced sounds 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 features formed the perceptual harmonic cepstral coefficients (PHCC) representation. Experiments under clean speech environment showed that PHCC significantly outperformed conventional MFCC for both voiced and unvoiced speech [4].

In this paper, the PHCC front end is extended for speech recognition in noisy environments by incorporating several "anti-noise" techniques. A weight function is designed for the computation of the harmonic weighted spectrum to mitigate the distortion of harmonic structures caused by background noise. The power spectrum is lower clipped before root-power compression to reduce the noise sensitivity associated with small spectral values. The root-power function is adjustable to the noisy environment characteristics. Experiments with the Mandarin digit database under varied noisy environments show that PHCC does provide significant improvement over conventional MFCC under noisy conditions.

2. PERCEPTUAL HAMONIC CEPSTRAL COEFFICIENTS

A. Spectral envelope vs. smoothed spectrum

Modern speech recognition systems retrieve information on the vocal tract transfer function from the gross spectral shape. The

This work was supported in part by the National Science Foundation under grant no. IIS-9978001, EIA-9986057, the University of California MICRO Program, Conexant Systems, Inc., Lernout & Hauspie Speech Products, Lucent Technologies, Inc., Medio Stream, Inc., and Qualcomm, Inc.

0-7803-7041-4/01/$10.00 ©2001 IEEE

speech signal is generated via modulation by an excitation signal that is quasi-periodic for voiced sounds, and white noise for unvoiced sounds. A typical approach, employed in MFCC and PLP, is to compute the energy output of a bank of band-pass mel-scaled or bark-scaled filters, whose bandwidths are broad enough to remove fine harmonic structures caused by the quasi-periodic excitation of voiced speech. The efficiency and effectiveness of these spectral smoothing approaches led to their popularity. However, there are several drawbacks that significantly decrease their accuracy and robustness.

The first drawback is the limited ability to remove undesired harmonic structures. In order to maintain adequate spectral resolution, the standard filter bandwidth in MFCC and PLP is usually in the range of 200Hz-300Hz in the low frequency region. It is hence sufficiently broad for typical male speakers, but not broad enough for high pitch (up to 450Hz) female speakers. Consequently, the formant frequencies are biased towards pitch harmonics and their bandwidth is misestimated.

The second drawback concerns information extraction to characterize the vocal tract function. It is widely agreed in the speech coding community that it is the spectral envelope and not the gross spectrum that represents the shape of the vocal tract [5]. Although the smoothed spectrum is often similar to the spectral envelope of unvoiced sounds, the situation is quite different in the case of voiced and transitional sounds. Experiments show that this mismatch substantially increases the spectrum variation within the same utterance [4].

The third drawback is the high spectral sensitivity to background noise. The conventional smoothed spectrum representation may be roughly viewed as averaging the upper and lower envelopes. It therefore exhibits much higher SNR than the upper spectrum envelope alone in noisy conditions.

Although some of the loss caused by the imprecision and low robustness of spectrum smoothing may be compensated for and masked by higher complexity statistical modeling, the recognition rate eventually reaches saturation at high model complexity. This motivated the development of the alternative of *Perceptual Harmonic Cepstral Coefficients* (PHCC) as a more accurate and robust spectral representation [4].

b. Computation of Perceptual Harmonic Cepstral Coefficient

PHCC computation is similar to that of MFCC except that it attempts to closely approximate the perceptually compressed spectral envelope sampled at pitch harmonics. The procedure consists of the following steps:

1) The speech frame is processed by DFT to obtain the short-term 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) [4][5] followed by the peak-picking algorithm);

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.

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).

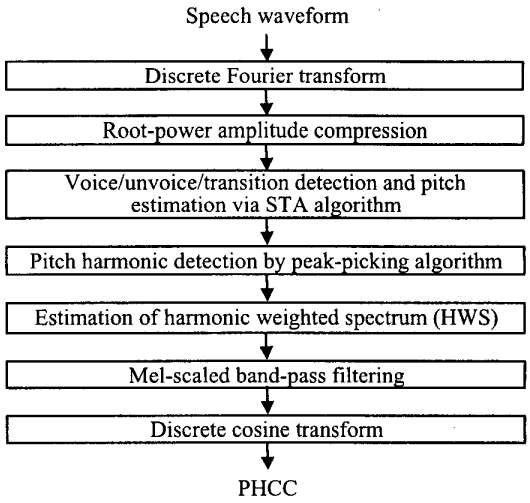

Figure 1. Flowchart of Perceptual Harmonic cepstrum Coefficient (PHCC) analysis

A flowchart of PHCC computation is shown in Figure 1. Next, we will describe steps 3 and 4 in more detail.

C. The peak-picking algorithm

In the case of voiced speech frames, more accurate determination of the harmonic frequencies is obtained by applying the peak-picking algorithm to the power spectrum, which corrects minor pitch estimation errors or non-integer pitch effects. The initial estimated harmonics obtained from STA are refined by looking for local maxima in a search interval that excludes neighboring harmonics. Once the peaks are found, the power spectrum value at pitch harmonics is given emphasis by appropriate weighting as will be explained below.

The peak-picking algorithm is also useful for transitional speech frames as they contain some quasi-harmonic structures. Since there are no well-defined initial harmonic frequencies, they are set to fixed values (multiples of 100Hz were quite effective in experiments).

D. Harmonic weighted spectrum (HWS)

Spectral envelope representation as above has been previously proposed and is currently used in harmonic speech coding [6], where the spectrum amplitude sampled at pitch harmonics is vector quantized. However, the number of harmonics varies significantly from speaker to speaker. This also implies that some processing must be applied to the harmonic spectrum prior to its applicability to speech recognition. We use the harmonics weighted energy output of mel-scale filters instead of the harmonic spectrum directly.

In the case of voiced speech, the most important information available about the spectral envelope is captured by the spectrum sampled at pitch harmonic frequencies. If the spectrum between pitch harmonics is smooth, interpolation methods can be used to retrieve the spectrum spline, albeit with high sensitivity to pitch estimation errors. Instead, we propose a different approach called harmonic weighted spectrum (HWS) estimation. Given $S_f(\omega)$, the magnitude spectrum of input speech, HWS is defined as

$$HWS(\omega) = w_h(\omega) \cdot S_f(\omega) \quad (1)$$

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

As shown in Figure 1, the filter log-energy is calculated from the HWS and followed by DCT to generate the cepstral coefficients.

In our clean speech simulations, W_H was set to 100 for voiced sounds and 10 for transitional sounds. The HWS of voiced speech reflects the spectrum spline at harmonic points. In the case of unvoiced speech, HWS is simply the power spectrum. The HWS of transitional speech represents the power spectrum with emphasis on quasi-harmonic points. Therefore, when combined with mel-scaled band-pass filtering, HWS can be effectively used to extract parameters that characterize the spectral envelope for the three classes of speech frames.

E. Within-filter amplitude compression

It is widely recognized that auditory properties can be exploited to improve automatic speech recognition. Perhaps the most notable example is the common use of band-pass filters of broader bandwidth at high frequencies, according to the frequency resolution of the human ear. MFCC implements this by mel-scaled spacing, and PLP employs critical-band spectral resolution. Another important aspect, the perceptual transformation of the spectrum amplitude, is handled in radically different ways by the leading front-end systems. PLP applies the equal-loudness curve and the intensity-loudness power law to better exploit knowledge about the auditory system [7], 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 attempt to "enjoy the best of both worlds", we apply the intensity-loudness power-low (here we use cubic-root amplitude compression for clean speech) within each filter and compute the log energy over all filters. Hence,

$$\hat{S}(\omega) = [S(\omega)]^{1/3} \quad (2)$$

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

where $\hat{S}(\omega)$ is the compressed spectrum and $\hat{E}_i$ is the log energy for band-pass filter i. 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.

3. EXTENSION OF PHCC TO NOISY ENVIRONMENTS

Although PHCC was initially proposed as a more accurate front end for clean speech recognition [4], it can also be extended to noisy speech recognition. To achieve this goal, several anti-noise modifications are applied to our previously proposed PHCC method.

A. Modified weight function for the HWS

The advantages of spectral envelope representation over conventional smoothed spectrum representation are less obvious in noisy environments. On the one hand, the harmonic spectrum estimation discards the variations in the valleys between harmonic locations caused by the background noise, which leads to more robust spectral representation. On the other hand, the original harmonic structure in voiced and transitional speech may be blurred significantly by the input additive noise, especially in high frequency regions. A solution to these problems calls for a more effective weight function for the HWS.

Here we propose a modified weight function for HWS estimation in noisy environments. A new parameter, *harmonic confidence*, is defined as

$$H_a = \max_{\tau} R(\tau) ,$$

where $R(\tau)$ is the spectro-temporal autocorrelation criterion defined in [4][5].

The harmonic weight of (1) is now modified to

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

where ω_T is the cut-off frequency. In the modified HWS computation, the harmonic-based spectral envelope is emphasized in the low frequency zone below ω_T, whose harmonic structure is more robust to additive noise. The conventional smoothed spectrum is retained in the high frequency zone above ω_T. In addition, the weight value depends on the harmonic confidence H_a, to account for the effect of noise signals, where η is the harmonic confidence threshold, and γ is the weight factor. In our experiment, ω_T, η and γ are set to 2.5 kHz, 0.5 and 10, respectively.

B. Pre-compression spectral masking

One major shortcoming of logarithm-based approaches (including MFCC and PLP) is that the logarithm function is unbounded as its argument tends to zero. It is thus very sensitive to small input values. This may greatly deteriorate the representation robustness, as these low energy parts hold the worst SNR under noisy environments. A common noise reduction technique is to apply a lower bound to the original spectrum (so-called "masking" [8]) before the logarithm operation. We found that this technique may be beneficially applied to the within-filter amplitude compression.

If $S(\omega)$ is the original spectrum, the masking operation can be defined as

$$\tilde{S}(\omega) = \max\left(S(\omega), c\right) \quad ,$$

where c is a very small value, which may either be a fixed number or vary depending on noise conditions.

C. Root-power representation

Another modification to improve the performance of PHCC representation in noisy environments consists of replacing the intensity-loudness power-low of (2) by

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

where θ is the root-power factor. While it was previously set to a fix value in clean speech recognition, it may now be adjusted to the noise environment.

4. EXPERIMENT RESULTS

To test the performance of PHCC, experiments were first carried out on a database of speaker-independent isolated Mandarin digits collected in white and babble noise 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) with 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 experiment, 26-dimension speech features were used, including 12 cepstral (MFCC or PHCC) parameters, log energy, and their 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 experiment results for PHCC and MFCC are summarized in Table 1 and 2.

Table 1 shows that the error rate decreased by nearly 50% in clean speech environment and by 23% to 36% in white noise environment, and demonstrates consistent superiority of PHCC over MFCC at differing noise levels. Table 2 shows that similar improvement of PHCC is achieved in babble noise environment. The main source of errors in recognizing Mandarin digits is the confusion between vowels such as [a] and [e]. This is where the spectral envelope based PHCC substantially outperforms conventional MFCC, hence the significant and consistent gains observed in clean speech and noisy environments. The improvement in noisy environment is also attributed to modified weight function for HWS, and the within-filter root-power amplitude compression following low-bound masking procedure.

Front-end	Clean	20 dB	10 dB	0 dB
MFCC	2.1 %	4.8 %	16.9 %	45.6
PHCC	1.1 %	2.9 %	13.0 %	29.1

Table 1. Test-set error rates of PHCC and MFCC for speaker-independent isolated Mandarin digit recognition under white noise environment

Front-end	Clean	20 dB	10 dB	0 dB
MFCC	2.1 %	4.1 %	13.3 %	35.2 %
PHCC	1.1 %	2.3 %	10.5 %	27.4 %

Table 2. Test-set error rates of PHCC and MFCC for speaker-independent isolated Mandarin digit recognition under babble noise environment

5. CONCLUSION

The perceptual harmonic cepstral coefficients (PHCC) were previously proposed as accurate features for clean speech recognition. The spectral envelope is represented based on the harmonic spectrum, which is a weighted version of the power spectrum that emphasizes pitch harmonics. The method also employs within-filter cubic root amplitude compression and logarithmic level-scaled band-pass filtering to exploit both the psychophysical and gain-invariance advantages of PLP and MFCC, respectively. The PHCC front-end is modified for noise speech recognition. The weighting function depends on the prominence of harmonic structure in the frequency domain, instead of on the voice/unvoice/transition classification. The weakness of harmonic structures in the high frequency spectrum is also considered in the weighting function design. A lower-clipping of the power spectrum is enforced prior to amplitude compression to enhance SNR , and hence noise robustness. Experiments on Mandarin digit speech recognition in noisy environments show significant performance gains of PHCC over MFCC. Future work will focus on the extension of PHCC to perceptual harmonic linear prediction.

6. REFERENCES

[1] 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.

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

[3] Q. Zhu and A. Alwan, "AM-demodulation of speech spectra and its application to noise robust speech recognition", *Proc. ICSLP2000* , Oct. 2000.

[4] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients as the front-end for speech recognition", *Proc. ICSLP2000* , Oct. 2000.

[5] Y. D. Cho, M. Y. Kim and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination", *Proc. ICASSP98* , pp. 601-604, 1998.

[6] M. Jelinek and J. P. Adoul, "Frequency-domain spectral envelope estimation for low rate coding of speech", *Proc. ICASSP99* , pp. 253-256, 1999.

[7] M. J. Hunt, "Spectral signal processing for ASR", *Proc. ASRU99* , Dec. 1999.

[8] D. H. Klatt, "A digital filter bank for spectral matching", *Proc. ICASSP79* , pp. 573-576, 1979.