6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition

Liang Gu and *Kenneth Rose*

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email: {liang, rose}@scl.ece.ucsb.edu

ABSTRACT

Perceptual harmonic cepstral coefficients (PHCC) are proposed as features to extract for speech recognition. Pitch estimation and classification into voiced, unvoiced, and transitional speech are performed by a spectro-temporal auto-correlation technique. A peak picking algorithm is then employed to precisely locate pitch harmonics. A weighting function, which depends on the classification and the pitch harmonics, is applied to the power spectrum and ensures 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. For perceptual considerations, within-filter cubic-root amplitude compression is applied to reduce amplitude variation without compromise of the gain invariance properties. Experiments show substantial recognition gains of PHCC over MFCC, with 48% and 15% error rate reduction for the Mandarin digit database and E-set, respectively.

1. INTRODUCTION

Most modern speech recognition systems focus on the speech short-term spectrum for feature analysis-extraction (also referred to as the "front-end" analysis). The technique attempts to capture information on the vocal tract transfer function from the gross spectral shape of the input speech, while eliminating as much as possible the irrelevant effects of excitation signals. Over the last several decades, a number of speech spectral representations have been developed [1], among which the mel-frequency cepstral coefficients (MFCC) [2] have become most popular. Nevertheless, it is believed that existing front ends are considerably sub-optimal, and that major gains may be recouped by advances in this area.

One main difficulty that plagues conventional front-end analysis is concerned with the vocal tract transfer function whose accurate 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 sounds. Experiments show that this mismatch substantially increases the feature variance within the same utterance.

Another difficulty encountered in conventional front-end analysis is that of appropriate spectral amplitude transformation for higher recognition performance. The log power spectrum representation in MFCC is clearly attractive because of its gain-invariance properties and the approximate Gaussian distributions it thus provides. Cubic root representation is used in the perceptual linear prediction (PLP) representation [3] for psychophysical considerations, at the cost of compromising the level-invariance properties and hence robustness.

In this paper, a new approach is proposed to overcome the above shortcomings, which is inspired by ideas borrowed from speech coding [4]. Rather than average the energy within each filter, the harmonic cepstral coefficients (HCC) are derived for voiced speech from the spectrum envelope sampled at harmonic locations for voiced speech. They are similar to MFCC for unvoiced sounds and silence. We adopt the spectro-temporal auto-correlation (STA) method for accurate and robust pitch estimation that was previously developed for sinusoidal speech coders [5]. The HCC representation is further improved by applying the intensity-loudness power-law 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 form the "perceptual" HCC (PHCC) representation. Experiments with the Mandarin digit and the E-set databases show that PHCC significantly outperforms conventional MFCC for both voiced and unvoiced speech.

2. SPECTRAL ENVELOPE ESTIMATION

A. Spectral envelope vs. smoothed spectrum

Modern speech recognition systems retrieve information on the vocal tract transfer function from the gross spectral shape. The 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 two drawbacks that significantly deteriorate their accuracy.

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

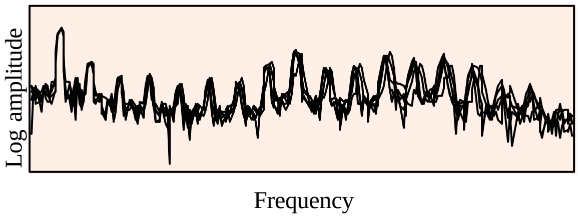

Figure 1. Power spectrum curves (512-point FFT) for 5 consecutive frames in speech segment [a]

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 [4]. 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. This phenomenon is illustrated in Figure 1 with the stationary part of the voiced sound [a]. The figure demonstrates that the upper envelope of the power spectrum sampled at pitch harmonics is nearly unchanged, while the variation of the lower envelope is considerable. The conventional smoothed spectrum representation may be roughly viewed as averaging the upper and lower envelopes. It therefore exhibits much more variation than the upper spectrum envelope alone.

Although some of the loss caused by the imprecision of spectrum smoothing may be compensated for and masked by higher complexity statistical modeling, the recognition rate eventually reaches saturation at high model complexity. The premise of this paper is that the sub-optimality of the front-end is currently a major performance bottleneck of powerful, high complexity speech recognizers. We therefore propose the alternative of *Harmonic Cepstral Coefficients* (HCC) as a more accurate spectral envelope representation.

B. Harmonic Cepstral Coefficient Computation

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

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

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

3) 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 becomes equivalent to the conventional smoothed spectrum.

4) 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).

A block diagram of HCC is shown in Figure 2. Next we will describe steps 2) to 4) and our implementation of them in greater detail.

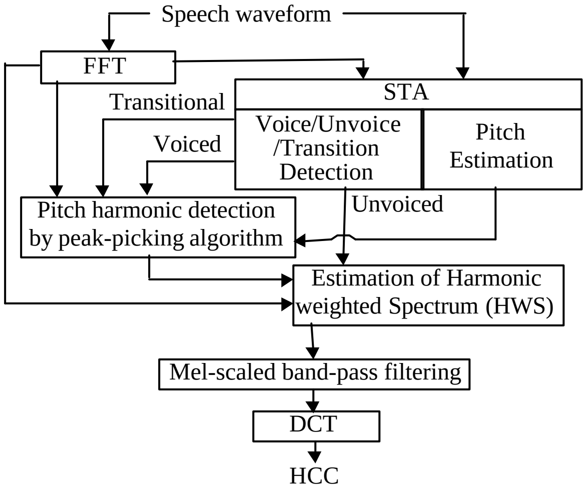
Speech waveform
FFT
STA
Transitional
Voice/Unvoice /Transition Detection
Pitch Estimation
Voiced
Unvoiced
Pitch harmonic detection by peak-picking algorithm
Estimation of Harmonic weighted Spectrum (HWS)
Mel-scaled band-pass filtering
DCT
HCC

Figure 2. Block diagram of Harmonic Cepstrum Coefficients (HCC) analysis

C. Robust pitch estimation by the spectro-temporal auto-correlation (STA) algorithm

Spectral envelope representation requires robust pitch estimation. Minor errors are easily corrected by the peak-picking algorithm to be described later. However, errors due to pitch multiples or sub-multiples will greatly corrupt the HWS computation for voiced speech frames. To mitigate the latter error types, our implementation adopts the STA algorithm that was first proposed for the design of harmonic speech coders [5].

Temporal auto-correlation (TA) has been traditionally used for pitch estimation. Given a speech signal $s_t(n)$, the TA criterion for candidate pitch $\boldsymbol{t}$ is defined as

$$R^T(\boldsymbol{t}) = \frac{\sum_{n=0}^{N-t-1}\left[\tilde{s}_t(n)\cdot\tilde{s}_t(n+\boldsymbol{t})\right]}{\sqrt{\sum_{n=0}^{N-t-1}\tilde{s}_t^2(n)\cdot\sum_{n=0}^{N-t-1}\tilde{s}_t^2(n+\boldsymbol{t})}}$$

where $\tilde{s}_t(n)$ is the zero-mean version of $s_t(n)$, and N is the number of samples for pitch estimation. The pitch estimate is obtained by maximizing TA. Unfortunately, TA occasionally selects pitch multiples, especially when the speech signal is highly periodic with a short pitch period. This error is disastrous for our purposes as it corrupts the estimated harmonic spectral envelope. Spectral auto-correlation (SA) has been proposed to circumvent the pitfall of pitch multiples, and is defined as

$$R^{S}(t)=\frac{\int_{0}^{p-w_{t}}\tilde{S}_{f}(w)\tilde{S}_{f}(w+w_{t})}{\sqrt{\int_{0}^{p-w_{t}}\tilde{S}_{f}^{2}(w)\int_{0}^{p-w_{t}}\tilde{S}_{f}^{2}(w+w_{t})}}$$

where $w_t = 2p/t$, $S_f(w)$ is the magnitude spectrum of $s_t(n)$, and $\tilde{S}_f(w)$ is the zero-mean version of $S_f(w)$.

Clearly, the danger here is of pitch sub-multiples. To mitigate both error types, STA is defined as an average criterion:

$$R(t)=b\cdot R^{T}(t)+(1-b)\cdot R^{S}(t)$$

where $b = 0.5$ was found to yield good results in practice [5].

The STA criterion $R(t)$ may also be used to perform V/UV/T detection. If $R(t) > a_V$, the speech frame is classified as voiced, if $R(t) < a_U$, it is classified as unvoiced, and if $a_V \geq R(t) \geq a_U$, it is declared transitional. The two thresholds can be determined based on experiments and we used $a_V = 0.8$ and $a_U = 0.5$.

D. 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 discussed next.

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).

E. Harmonics weighted spectrum (HWS)

Spectral envelope representation as above has been previously proposed and is currently used in harmonic speech coding, where the spectrum amplitude sampled at pitch harmonics is vector quantized. However, the number of harmonics varies significantly from speaker to speaker (a problem that led to growing interest in variable dimension vector quantization). This also implies that some processing must be applied to the harmonic spectrum prior to its applicability to speech recognition. We propose to 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 harmonics weighted spectrum (HWS) estimation. Given $S_f(w)$, the magnitude spectrum of input speech, HWS is defined as

$$HWS(w)=w_{h}(w)\cdot S_{f}(w)$$

where $w_h(w)=\begin{cases} W_H, & w\, is\ pitchharmonic \\ 1, & otherwise \end{cases}$

As shown in Figure 2, the filter log-energy is calculated from the HWS and followed by DCT to generate the cepstral coefficients.

In our 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.

3. PERCEPTUAL HAMONIC CEPSTRAL COEFFICIENTS

A. 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, but requires scale normalization, which was experimentally found critical for 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 propose a new approach which applies intensity-loudness power-low (here we use cubic-root amplitude compression) within each filter and computes the log energy over all filters. Hence,

$$\hat{S}(w)=[S(w)]^{1/3}$$

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

where $\hat{S}(w)$ is the compressed spectrum and $\hat{E}_i$ is the log energy for band-pass filter *i*. The resulting spectrum representation can

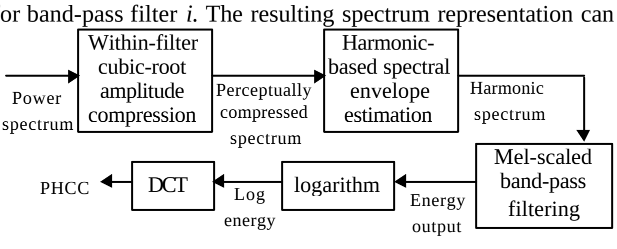
Within-filter cubic-root amplitude compression
Power spectrum
Perceptually compressed spectrum
Harmonic-based spectral envelope estimation
Harmonic spectrum
Mel-scaled band-pass filtering
Energy output
logarithm
Log energy
DCT
PHCC

Figure 3. Perceptual Harmonic Cepstrum Coefficients (PHCC) speech analysis 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.

B. Perceptual harmonic cepstrum coefficients

We incorporate the above perceptual amplitude transformation within the HCC framework to obtain the proposed *perceptual* HCC (PHCC), as is shown in Figure 3. We have noted that the within-filter amplitude compression reduces envelope corruption damage caused by pitch harmonic errors in the case of voiced sounds, and decreases amplitude variation due to white-noise in unvoiced sounds. It thus improves the accuracy and robustness of spectral envelope estimation.

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 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 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 tied-mixture HMM was used with 99 single Gaussian pdfs. The experiment results for PHCC and MFCC are summarized in Table 1.

Table 1 shows that the error rate has been decreased by nearly 50% for both male and female speakers, and demonstrates the consistent superiority of PHCC over speakers with differing pitch levels. 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 gains observed.

Speaker Gender	Male	Female	Male & Female
MFCC	0.5 %	3.0 %	2.1 %
PHCC	0.2 %	1.4 %	1.1 %

Table 1. Test-set error rate based on PHCC and MFCC for speaker-independent isolated Mandarin digit recognition

Acoustic Models	7-state CHMM	13-state CHMM	21-state TMHM M
MFCC	15.3%	11.0 %	7.3 %
PHCC	12.2 %	9.0 %	6.2 %

Table 2. Test-set error rate based on PHCC and MFCC for English E-set recognition

To critically test the performance of PHCC on unvoiced sounds, experiments were further carried out on OGI's E-set database. The recognition task is to distinguish between nine 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 experiment results are summarized in Table 2. PHCC achieved better results than MFCC over a range of acoustic model complexities, and offers over 15% error reduction relative to MFCC. As the utterances in the E-set database mainly differ in the unvoiced sounds, the improvement is attributed to the new perceptual amplitude transformation and the handling of transition sounds in the harmonic spectrum estimation.

5. CONCLUSION

The proposed harmonic cepstral coefficients (HCC) offer a representation of the spectral envelope based on the harmonic spectrum, which is a weighted version of the power spectrum that emphasizes pitch harmonics. The weighting function depends on the frame's V/UV/T classification. In order to exploit both the psychophysical and gain-invariance properties of PLP and MFCC, respectively, the method employs within-filter cubic root amplitude compression and logarithmic level-scaled band-pass filtering. Experiments on the Mandarin digit and E-set databases show substantial performance gains of PHCC over MFCC. Future work will focus on the extension of PHCC to perceptual harmonic linear prediction.

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

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

[5] 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.