[[1]](#footnote-1)

A Review on Image Haze Removal Using Dark Channel Prior

Zhuowen Lin, 11510818, *Information Engineering, SUSTech*

*Abstract*—The quality of images captured by digital camera sensors can be degraded by a lot of reasons, the haze present in the atmosphere is one of which. The removal of haze, called dehazing, is typically performed upon the basis of the atmospheric light degradation model. To accomplish the task of dehazing, a statistical knowledge called dark channel prior (DCP) was proposed and later received different improvements from different research. The DCP is derived from the characteristic of outdoor haze free images that the intensity value of at least on color channel within a local image window is close to zero. Based on the DCP, the removal of haze can be finished in five steps: dark channel construction, atmospheric light estimation, transmission map construction, transmission map refinement, and image reconstruction. The five steps of dehazing not only enable us to implement them step-by-step in MATLAB, but also give us a chance to cast light on the comparison between different methods proposed by different research in each step. This review of image haze removal using dark channel prior will help readers understand the implementations and evaluations of dehazing methods based on DCP.

*Index Terms*—dehazing, image restoration

# INTRODUCTION

D

UE to the absorption and scattering by atmospheric particles in haze, outdoor images have poor quality under hazy weather. Poor quality images with low visibility bring negative effects on photography appreciation as well as computer vision applications for outdoor environments, such as object detection in satellites’ remote sensing. Haze removal, which is also named as dehazing, is considered as a critical process since clear, haze-free images can not only be visually favorable, but also be an important factor that help to improve the performance of computer image processing techniques

selection of the high-quality parametric representation of acoustic data is an important task in the design of any speech recognition system. The usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences. When the amount of reference information in speech signal is significant, compact storage of the information becomes an important practical consideration.

Because the choice of speech segmentation methods prior to speech feature extraction in speech automatic recognition is basic to the decision as to what acoustic information is useful, and the choice of parametric representations significantly affects the recognition results [1], researchers have been attempting to find a parametric representation that can be generalized to differently organized speech recognition system.

A preliminary experiment [2] showed that the cepstrum coefficients were useful for representing consonantal information. One widely used cepstrum coefficients is Mel-Frequency Cepstrum Coefficients (MFCC).

This review paper covers a brief development of MFCC. The next section describes another type of cepstrum coefficients named as Linear Prediction Cepstrum Coefficients (LPCC), and its comparison with MFCC. The following section describes the step-by-step implementation of MFCC. Finally, a new technique aiming at feature extraction called Complex Linear Projection (CLP) are introduced briefly.

# Linear Prediction Cepstrum Coefficients

The calculation of cepstrum of speech signals introduced in [3] is based on the periodogram estimate of the power spectrum. The process can be written mathematically as .

The Linear Prediction Cepstrum Coefficients (LPCC) are computed in the same way with cepstrum except that they are computed from the smoothed auto-regressive power spectrum instead of the periodogram estimate of the power spectrum. As the introduction in [3], Levinson Durbin algorithm can be applied to the calculation of LPCC. In MATLAB, we can utilize the build-in function “lpc” to compute LPCC for a certain frame of speech in following steps:

>> [lp,g] = lpc(SpeechFrame,10);

>>ARPowerSpectrum = g./abs(fft(lp,1024)).^2;

>>LPCC = ifft(log(ARPowerSpectrum),1024);

Another way to calculate LPCC mentioned in [4] is to use a recursive formula computing LPCC directly from Linear Prediction Coefficients (LPC) without doing any DFTs. The recursive formula is:

,

in which are the linear prediction coefficients.

In [5], five parametric representations (MFCC, LFCC, LPCC, LPC, and RC) were used to do the speech recognition test. The results are shown in Table 1 and Fig.1.

TABLE I

Recognition Rates Resulting from Use of Various Acoustic Representations.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Acoustic Representation | Number of coefficients | | Frame Separation (ms) | | Speaker | | Open Test % | |
| Mel-frequency cepstrum  Mel-frequency cepstrum  Linear-frequency cepstrum  Linear-prediction cepstrum  Linear-prediction spectrum  Reflection coefficients | | 10  6  10  10  10  10 | | 6.4  12.8  6.4  12.8  6.4  12.8  6.4  12.8  6.4  6.4  12.8 | | DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL  DZ  LL | | 96.5  95.0  95.6  93.8  96.5  92.0  95.0  90.2  94.7  87.6  93.2  84.9  92.6  87.3  91.7  86.4  85.2  84.3  83.1  77.5  80.5  74.6 | |

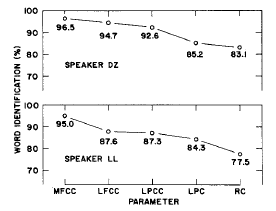


Fig. 1. Performance of parametric representations for recognition.

The results listed in Table 1 and shown in Fig.1 for open tests with 10 coefficients and 6.4 ms frames. Regardless of the frame separation, type of testing or speaker, these data indicate superior performance of the MFCC when compared with the other parametric representations, including LPCC.

The conclusion given in [5] mentioned that parameters derived from the short-term Fourier spectrum, including MFCC, of the acoustic signal preserve information that parameters from the LPC spectrum, including LPCC. The differences lie most frequent in the recognition of consonants and may be due to the inaccurate representation of the consonantal spectra by the linear prediction technique.

The advantage of MFCC was also given in [5], saying that MFCC, being motivated by perceptual factors, has a better ability to capture the perceptually relevant information. In next section, the implementation details of MFCC will be introduced.

# Mel-Frequency Cepstrum Coefficients

In this section, in order to give a full review on mel-frequency cepstrum coefficients (MFCC), the mel-frequency will be introduced first. Then, the steps of calculating MFCC and the reason for doing them will be covered.

## Mel-frequency

As the introduction given in [6], mel-frequency scale is a perceptually motivated scale. It is linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above 1 kHz. In the experiments carried out in [7], the mel scale is based on experiments with simple tones in which subjects were required to divide given frequency ranges into four perceptually equal intervals or to adjust the frequency of a stimulus tone to be half as high as that of a comparison tone. The mel frequency is introduced with hope that it can model the sensitivity of the human ear more closely than a purely linear scale and provides for greater discriminatory capability between speech segments.

One approximation of mel-scale frequency is

in which *f* is the frequency scale, B(*f*) is the mel frequency scale.

The mel frequency scale is plotted in Fig. 2.

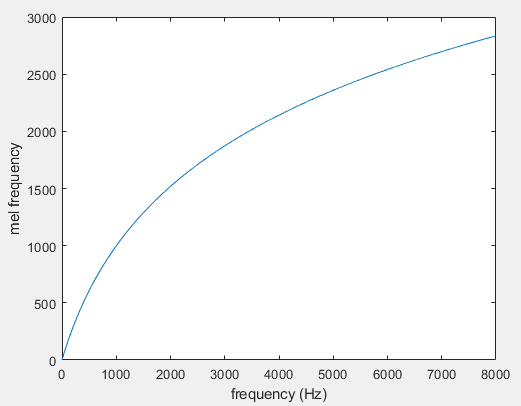


Fig. 2. Mel frequency scale.

## Steps of calculating MFCC

Concluding the contents in [8], the steps of calculating MFCC can be listed as follows:

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimation of the power spectrum.
3. Apply the mel filter bank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filter bank energies.
5. Take the discrete Fourier transform (DCT) of the log filter bank energies.
6. Keep the 2nd to 13th DCT coefficients and discard the rest.

Sometimes, the delta-delta features for tracking dynamic changing trend are also append to the calculated MFCC.

In next parts, I will explain why and how these steps are done.

## Frame the signal

As we have learnt in our “Speech Signal Processing” course, the speech signal is constantly changing, but concerning short time scales, the speech signal does not change a lot, achieving a statistically stationary. The reason for framing the original speech signal is to divide the signal into short time scales to achieve such statistically station

A typical frame length is about 20 -40 ms. If the frame is too short then there is not enough sample to get a reliable spectral estimate. If the frame is too long, then the signal changes too much throughout the frame. If the speech file does not divide into an even number of frames, pad it with zeros.

## Periodogram estimation of power spectrum

The calculation of power spectrum is motivated by the behavior of human cochlea – it vibrates at different locations depending on the frequency of the incoming sounds. The power spectrum estimation does a similar job for us, identifying the frequencies presenting in a certain frame.

Mathematically, we first take the DFT of the speech frame:

where is a certain speech frame, is an N sample long hamming window, K is the length of DFT.

Then, the periodogram estimation of power spectrum can be calculated as:

However, by taking the power spectrum, we discard the information embedded within the angle of discrete Fourier transform (DFT), which is proved as useful in Complex Linear Projection (CLP) method proposed by Google Inc.7u

## Mel filter bank

The periodogram estimation of power spectrum contains a lot of information that is not required for speech recognition. In particular the cochlea cannot discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. Therefore, we can further compress the information to simplify our parametric representation.

What we need to do is to filter the power spectrum with a series of filters (filter bank) and sum the filtered power spectrum within each filter region to get an idea of how much energy exists in various frequency regions.

Because all the manipulations are on the basis of mel frequency scale, the mel-scale filter bank has a special characteristic: the filter gets wider as the frequency gets higher. This is because as mentioned, our ears cannot distinguish closely placed high frequencies.

The number of filters in the filter bank ranges from 20 to 40 (26 is a standard). All of the filters are triangular filters. As introduced in [8], they can be expressed mathematically by:

in which can be calculated as:

and M is the number of filters. refers to the mel-scale frequency.

After having the mathematical formulas, we can easily construct such a bank of filters in MATLAB.

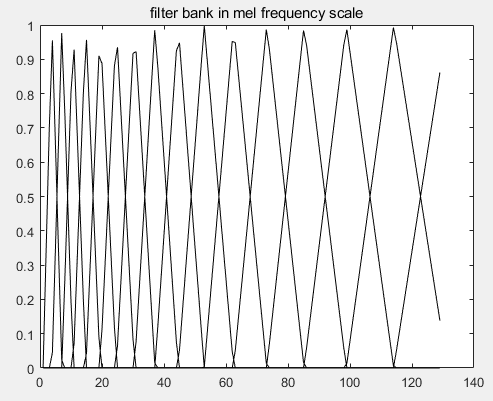


Fig. 3. Mel filter bank.

## Logarithm of all filter bank energies

The logarithm of all filter bank energies is also motivated by human hearing – we hear loudness in a logarithm scale. This manipulation can help our parameter representation match more closely to the actual conditions.

## Discrete cosine transform (DCT)

There are mainly two reasons for taking DCT: one is that DCT can decorrelates the overlapping filter bank energies, which means that diagonal covariance matrices can be used to model the features in a Hidden Markov Model (HMM) classifier. Another one is that because DCT has cosines with half-integer numbers of cycles, it can compactly fit the shape of power spectrum energies with few DCT coefficients, which is better than DFT. For a particular frame in the speech file “s5.wav”, I write a MATLAB program to test how many DCT coefficients are sufficient to represent over 99% of the energy. The result shows that only 84 out of totally 208 DCT coefficients are sufficient to represent over 99% of the energy. If we use these 84 coefficients to reconstruct the speech signal in this frame, comparing with the original frame, the result would be shown in Fig. 4.

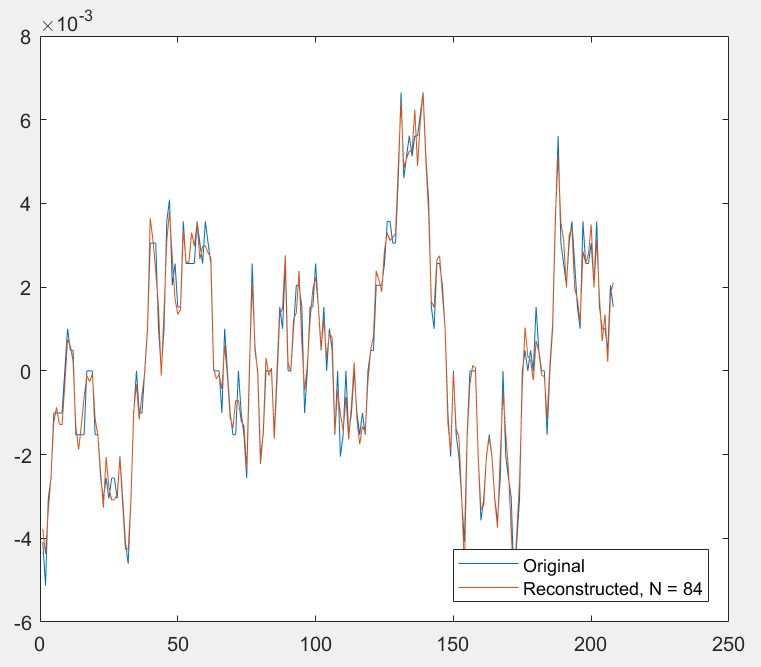


Fig. 4. Original speech frame and reconstructed speech frame from 84 DCT coefficients.

## Coefficient discarding

In fact, 84 coefficients are redundant for MFCC. Only 12 DCT coefficients are kept since the higher DCT coefficients represent fast changes in the filter bank energies and it turns out in [5] that these fast changes are not useful in speech recognition. By discarding coefficients can further simplify our parametric representation.

## Delta coefficient

The 12 MFCC feature coefficients describe only the power spectral envelope of a single frame, but the speech would also have information lying in dynamics, such as the changing trend of the MFCC. Therefore, it would be helpful to append delta coefficients after MFCC.

The delta coefficients are defined as:

where is a delta coefficient, t is the current frame number, is the MFCC in frame n. By introducing MFCC in previous and later frames into delta coefficients, the dynamic changing trend of MFCC has been preserved.

# Complex Linear Projection

In [9] proposed by Google Inc., the authors claimed a disadvantage of MFCC, saying MFCC separates the perceptually motivated filters from the acoustic model, which is not always the best choice in statistical modeling frameworks such as automatic speech recognition (ASR), where the end goal is word error rate minimization.

In order to overcome this disadvantage, the authors introduced a neural-network-based technique called Complex Linear Projection (CLP), which performs both filtering layer and pooling layer in neural network in frequency domain and produces set of ASR features which can be fed to the backend neural network acoustic model. By doing this, the CLP model can be jointly optimized with the acoustic model.

## Comparison between CLP and MFCC

Introduced in [9], the output of the filter in CLP model is:

.

where is a complex matrix called projection matrix in . is the DFT of input signal.

The CLP model is different from MFCC in terms of their pooling operation: MFCC uses pooling while CLP uses a simple summation pooling. The MFCC filter bank operates on features while CLP model operates on . This means that the phase information in is removed in the MFCC while it is preserved in CLP model. There is a long-time debate about the improvement of phase for single microphone speech recognition. However, in multi-channel recognition, it is agreed by [10] that phase information is necessary, because it can preserve the relative delay of the speech signal at each microphone channel.

## Experiment results comparing CLP and MFCC

In one of the experiments in [9], the performance of CLP is evaluated in terms of the effectiveness of CLP model to achieve state-of-the-art performance using MFCC. The results are shown in Table II and Table III.

Table II presents the baseline WERs for MFCC model across three different window sizes. Typically, a window size of 25 msec with a 10 msec shift is used in ASR. As the results show, longer window contains more temporal information as well as localization of multi-channel processing which result in WER improvement over shorter windows.

The baseline CLP model results are shown in Table III. The CLP model performance is in the same level as MFCC model for single channel but yields a gain of about 4% over the 2-channel baseline models in Table II.

TABLE II

Word Error Rate for the Baseline Models of MFCC

|  |  |  |
| --- | --- | --- |
| Model | 1-Channel | 2-Channel |
| 25 msec | 23.4 | 21.8 |
| *32 msec* | 22.8 | 21.3 |
| *64 msec* | 21.8 | 20.7 |

TABLE III

Word Error Rate for the Baseline Models of CLP

|  |  |  |
| --- | --- | --- |
| Model | 1-Channel | 2-Channel |
| 25 msec | 23.2 | 21.5 |
| *32 msec* | 22.8 | 20.9 |
| *64 msec* | 22.0 | 20.5 |

To sum up, the Complex Linear Prediction (CLP) model achieves superior performance compared to MFCC by preserving all the information in the signal, including the time delay or phase information from multiple channels in a microphone array, which makes it appropriate to automatically learn the optimal feature extraction parameters in the multi-channel setting at the help of neural network.

# Conclusions

This paper cast light on the development stages of mel-frequency cepstrum coefficients (MFCC). The cepstrum-based linear prediction cepstrum coefficients (LPCC) is firstly introduced. Its calculation is based on smoothed auto-regressive power spectrum or recursive formula from LPC. However, LPCC is inferior to MFCC in the comparison between parametric representations for speech recognition regardless of the frame separation, type of testing or speaker. MFCC is introduced secondly. The better performances of MFCC are largely due to its perceptual-based mel-frequency scale, which can better model the human hearing. The computation of MFCC is divided into six steps. The paper reviewed the mathematical formulas as well as MATLAB implementations for these six steps. A technique yielding superior performances over MFCC called complex linear projection (CLP) is included thirdly. By preserving the phase angle information in CLP, it can have a smaller word error rate than MFCC in multi-channel speech recognition.

TABLE I

Units for Magnetic Properties

|  |  |  |
| --- | --- | --- |
| Symbol | Quantity | Conversion from Gaussian and  CGS EMU to SI a |
| Φ | magnetic flux | 1 Mx → 10−8 Wb = 10−8 V·s |
| *B* | magnetic flux density,  magnetic induction | 1 G → 10−4 T = 10−4 Wb/m2 |
| *H* | magnetic field strength | 1 Oe → 103/(4π) A/m |
| *m* | magnetic moment | 1 erg/G = 1 emu  → 10−3 A·m2 = 10−3 J/T |
| *M* | magnetization | 1 erg/(G·cm3) = 1 emu/cm3  → 103 A/m |
| 4π*M* | magnetization | 1 G → 103/(4π) A/m |
| σ | specific magnetization | 1 erg/(G·g) = 1 emu/g → 1 A·m2/kg |
| *j* | magnetic dipole  moment | 1 erg/G = 1 emu  → 4π × 10−10 Wb·m |
| *J* | magnetic polarization | 1 erg/(G·cm3) = 1 emu/cm3  → 4π × 10−4 T |
| χ*,* κ | susceptibility | 1 → 4π |
| χρ | mass susceptibility | 1 cm3/g → 4π × 10−3 m3/kg |
| μ | permeability | 1 → 4π × 10−7 H/m  = 4π × 10−7 Wb/(A·m) |
| μr | relative permeability | μ → μr |
| *w, W* | energy density | 1 erg/cm3 → 10−1 J/m3 |
| *N, D* | demagnetizing factor | 1 → 1/(4π) |

Vertical lines are optional in tables. Statements that serve as captions for the entire table do not need footnote letters.

aGaussian units are the same as cg emu for magnetostatics; Mx = maxwell, G = gauss, Oe = oersted; Wb = weber, V = volt, s = second, T = tesla, m = meter, A = ampere, J = joule, kg = kilogram, H = henry.

References

1. G. M. White and R. B. Neely, “Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming,” *IEEE Trans, Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 183-188, Apr. 1976.
2. P. Mermelstein, “Distance measures for speech recognition, psychological and instrumental,” in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed. New York: Academic, 1976, pp. 374-388.
3. L. R. Rabiner and R. W. Schafer, “The Cepstrum and Homomorphic Speech Processing,” in *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, United States: Prentice Hall, 2011, ch. 8, sec. 4, pp. 429-440
4. X. Huang, A. Acero and H. Hon, “Cepstrum of Speech Signals,” in *Spoken Language Processing*, 1st ed, Upper Saddle River, NJ, United States: Prentice Hall, 2001, ch. 6, sec. 4, pp.312-314
5. S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentecnces,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357-366, 1980
6. X. Huang, A. Acero and H. Hon, “Speech Perception,” in *Spoken Language Processing*, 1st ed, Upper Saddle River, NJ, United States: Prentice Hall, 2001, ch. 2, sec. 1, pp.33-34
7. Steven, S.S. and J. Volkman, “The Relation of Pitch to Frequency”, *Journal of Psychology*, 1940, 53, pp. 329
8. X. Huang, A. Acero and H. Hon, “Mel Frequency Cepstrum,” in *Spoken Language Processing*, 1st ed, Upper Saddle River, NJ, United States: Prentice Hall, 2001, ch. 6, sec. 5, pp.316-318
9. E. Variani, T.N. Sainath, I. Shafran, and M. Bacchiani, “Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling,” in INTERSPEECH, San Francisco, CA, USA, 2016
10. T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, “Speaker Localization and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms,” in *Proc. ASRU*, 2015.

1. Zhuowen Lin is with the Information Engineering major, Southern University of Science and Technology, Shenzhen, Guangdong, China (e-mail: 11510818@mail.sustc.edu.cn). [↑](#footnote-ref-1)