Mel-frequency cepstral coefficient analysis in speech recognition

Conference Paper · July 2006

DOI: 10.1109/ICOCI.2006.5276486 · Source: IEEE Xplore

Chin Kim On
Universiti Malaysia Sabah (UMS)
93 PUBLICATIONS **675** CITATIONS
SEE PROFILE

Paulraj M P
Universiti Malaysia Perlis
157 PUBLICATIONS **1,091** CITATIONS
SEE PROFILE

Sazali Yaacob
University of Kuala Lumpur
366 PUBLICATIONS **5,090** CITATIONS
SEE PROFILE

Azali Saudi
Universiti Malaysia Sabah (UMS)
150 PUBLICATIONS **568** CITATIONS
SEE PROFILE

Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition

Chin Kim On Paulraj M. Pandiyan Sazali Yaacob Azali Saudi
School of Engineering and Information Technology
Universiti Malaysia Sabah
Log Beg No 2073, 88999 Kota Kinabalu, Sabah, Malaysia.
E-mail: c_kim_on@yahoo.com

***Abstract*-Speech recognition is a major topic in speech signal processing. Speech recognition is considered as one of the most popular and reliable biometric technologies used in automatic personal identification systems. Speech recognition systems are used for variety of applications such as multimedia browsing tool, access centre, security and finance. It allows people work in active environment to use computer. For a reliable and high accuracy of speech recognition, simple and efficient representation methods are required. In this paper, the zero crossing extraction and the energy level detection are applied to the recorded speech signal for voiced/unvoiced area detection. The detected voiced signals are applied for segmentation. Further, the MFCC method is applied to all of the segmented windows. The extracted MFCC data are further used as inputs for neural network training.**

I. Introduction

Speech has evolved over a period of ten thousand years as the primary means of communication between human beings. The speech signal has been studied for various reasons and applications by many researches for many years. Speech has evolved as a primary form of communication between humans, nevertheless, there often occur conditions under which we measure and then transform the speech signal to another form in order to enhance human ability to communicate [1, 2, 3]. Some of the researchers broke down the speech signal into its smallest portions, called phonemes. With these phonemes information, many of the methods are applied in order to estimate all of the information in speech signal [2, 3, 4]. Now a day, the speech recognition technologies is very important because it can provided the services for access centre, security, finance and the replacement of handwriting in PC system. In this paper, the MFCC extraction is proposed for the speech recognition system. The speech recognition system is divided into three parts, which is pre-processing of speech signal that involved filtering and normalization, the energy level and zero crossing detection for voiced and unvoiced signal detection and the MFCC feature extraction. Finally, the extracted MFCC data are further used as the input for neural network training.

II. Methodology

The original speech signal is recorded using a multi-channel microphone with the input speech of digit from 0 to 9 as password. The recorded speech signal consists of background noise. The background noise is removed using a filtering method. Further, the filtered speech signal is applied for the normalization process. The normalization technique is used to eliminate data redundancy in filtered speech signal. The normalized speech signal is then applied for the energy level and zero crossing extraction. The energy level and zero crossing methods are applied to the speech signal for voiced and unvoiced speech area detection. Each of the detected voiced speech signals are then applied for segmentation. Further, the detected windows are applied for the MFCC extraction and finally the extracted data are used as input for neural network training.

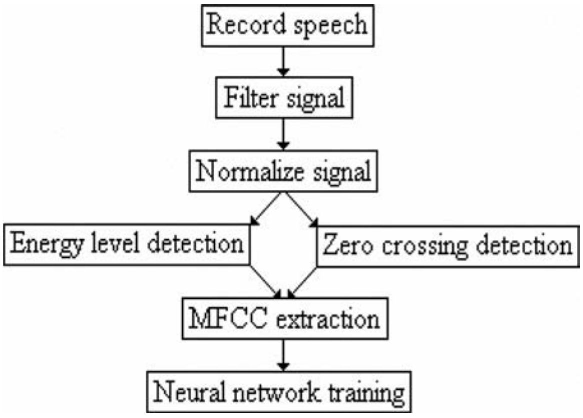

Fig. 1. Methodology overview.

III. Per-Processing of Speech Signal

The pre-processing of speech signal involved the stages of speech recording, filtering, normalization, energy level detection and zero crossing extraction. The recorded speech signal is saved in wav file format. The recorded speech signal consists of some background noise. Noise is defined as any unwanted and unmodulated energy that is always presented to some extent within any signal [2, 3, 4, 5]. An effective and efficient filtering process is needed in order to remove and reduce the background noise. In this paper, the recorded speech signal is filtered in both the forward and reverse direction. The filtering method is used to minimize the start-up and ending transients by matching the initial conditions and it is works for both real and complex inputs. This filtering method is implemented as a direct form II transposed structure where it is also operated as a Z-transform domain and the formula is shows in (1).

$$Y(z) = \frac{b(1) + b(2)z^{-1} + \ldots + b(nb+1)z^{-nb}}{1 + a(2)z^{-1} + \ldots + a(na+1)z^{-na}} X(z) \qquad (1)$$

The filtered speech signal is further applied for logarithmic normalization process. The logarithmic normalization process is defined as the recalculation of raw count data by a carefully chosen denominator. The normalization process gives the user a clearer statistical picture of the normalized data. The normalization technique is needed in order to eliminate data redundancy in filtered speech signal [6, 7]. A normalized data structure is also minimizes data element coupling and maximizes data element cohesion. Figs. 2, 3 and 4 below show the recorded speech signal, filtered speech signal and normalized speech signal.

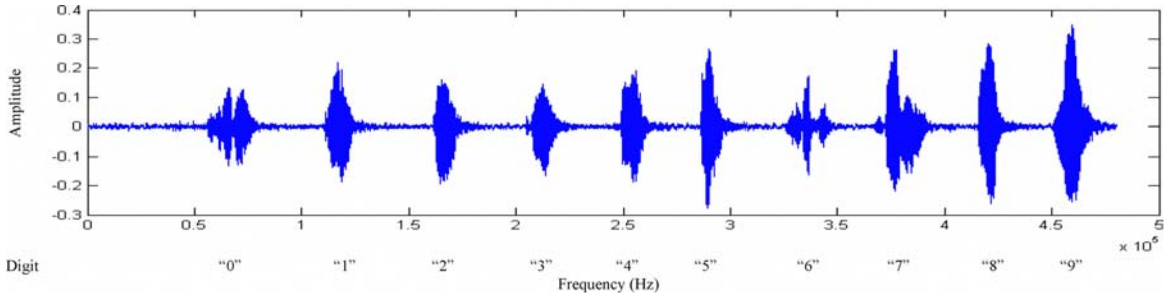

Fig. 2. Recorded original Speech Signal with digits of 0, 1, 2,…,9.

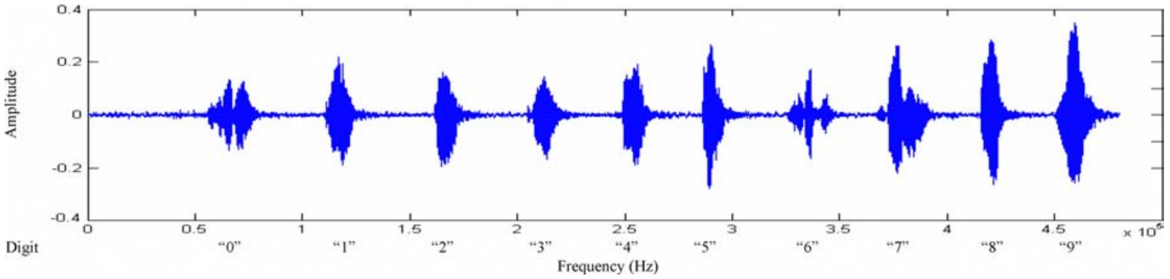

Fig. 3. Filtered speech signal.

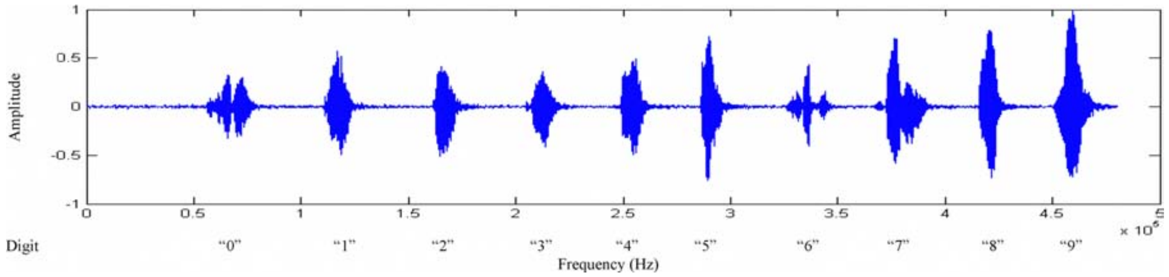

Fig. 4. Normalized speech signal.

From the figures shown above, it is stated clear that the recorded speech signal is a continuous speech signal. The continuous speech signal is the combination of voiced and unvoiced speech signals. An effective skill is needed to detect the voiced and unvoiced area in the normalized speech signal. In this paper, the energy level and zero crossing detection are used for the voiced and unvoiced signal detection. The algorithms of energy level and zero crossing detection are listed as below with respectively.

```
Algorithm of energy level detection
Start
1.0 Read normalized speech signal
2.0 Apply the segmentation for the normalized
    speech signal.
3.0 Compute the energy level for each segmented
    window.
4.0 Compute the average energy level of speech
    signal.
    4.1 If the energy level in each segmented
        window is larger than the average energy
        value
        4.1.1 Assign the value in the segmented
              window as 1.
    4.2 Else
        4.2.1 Assign the value in the segmented
              window as 0.
```

```
End
```

Algorithm of zero crossing detection

```
Start
1.0 Read normalized speech signal.
2.0 Find out the positive and negative value in speech signal.
    2.1 If the detected value is positive,
        2.1.1 Assign the value of speech signal as 1.
    2.2 Else
        2.2.1 Assign the value of speech signal as -1.
3.0 Apply the segmentation for the speech signal.
4.0 Find out the changing value in each of the segmented window.
    4.1 If the changing value found.
        4.1.1 Zero crossing = Zero crossing + 1 (Assign the zero crossing value for each of the segmented window).
5.0 Compute the average zero crossing value for all of the segmented windows.
6.0 Set the average zero crossing value as zero crossing threshold.
7.0 For each of the segmented windows.
    7.1 If zero crossing value > zero crossing threshold
        7.1.1 Assign the segmented window value as 1
    7.2 Else
        7.2.1 Assign the segmented window value as 0
End
```

Further, the energy level and zero crossing signals are used as references for the MFCC feature extraction and it is further discusses in section E below. The results after the energy level and zero crossing algorithms applied to the normalized speech signal are shown as the Figs. 5 and 6 below.

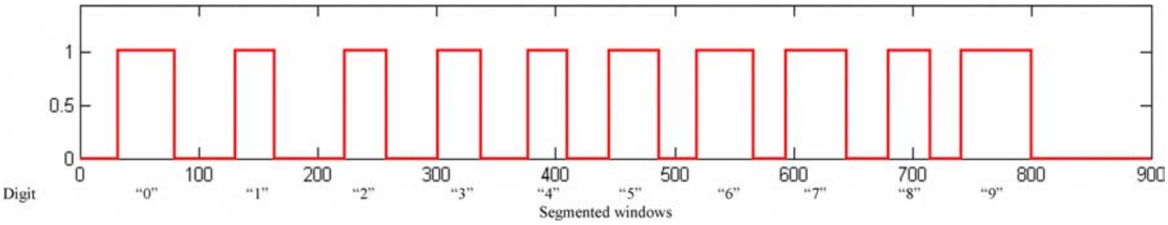

Fig. 5. Energy level detection (energy level signal).

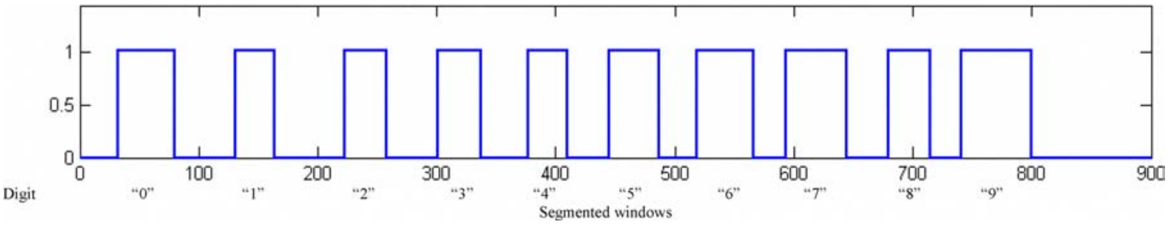

Fig. 6. Zero crossing extraction (zero crossing signal).

IV. MFCC Feature Extraction

MFCC or MSCC is the method that is applied for speech parameterization. Human ear has been proven to resolve frequencies non-linearly across the audio spectrum, thus filter bank analysis is more desirable than Linear Predictive Coding (LPC) analysis since it is spectrally based method. Mel-Scale is a frequency-binning method based on the human ear's frequency resolution. With the use of frequency bins on the mel-scale, MFCC is computed and used to parameterize speech data. The mel-scale also attempts to mimic the human ear in terms of the manner with which frequencies are sensed and resolved. The mel-scale is a unit of measurement of perceived frequency (pitch) of a tone [6, 7, 8].

The MFCC extraction method can be achieved by two ways, either FFT based (Fast Fourier Transform) or LPC based (Linear Predictive Coding). In this paper, MFCC extraction based FFT is used. Generally, MFCC extraction involves several stages, which are pre-emphasis, framing/segmentation, windowing, FFT spectrum, mel-spectrum extraction and mel-cepstrum extraction.

The general procedure of mel-cepstrum extraction actually involve, dividing the signal into frames, to obtain the power of spectrum, to convert the mel-spectrum and lastly uses the Discrete Cosines Transform (DCT) to get the cepstrum coefficient. Fig.

7 shows the general MFCC extraction methodology and the algorithm of the MFCC extraction is listed below the figure.

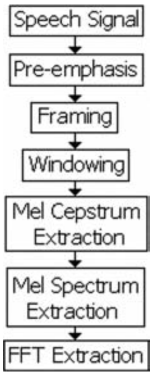

Fig. 7. General MFCC extraction methodology overview

The algorithm of MFCC extraction:

```
Start
  1.0 Read normalized speech signal.
  2.0 Read energy level and zero crossing signals.
  3.0 If energy level and zero crossing signals value are equal to 1.
      3.1 Extract the signal value.
  4.0 For each of the detected isolated word (applied for 15 times).
      4.1 Apply for pre-emphasis of normalized speech signal using formula.
          $H(z) = 1 + (-0.95 * z^{-1})$   (2)
      4.2 Apply for framing/segmentation (10-20 frames are selected).
      4.3 Apply for signal windowing with formula.
          $0.54 - 0.46 \cos\left(\frac{2\pi n}{300 - 1}\right)$   (3)
      4.4 Apply for FFT extraction with formula.
          $\left|\alpha_{mn} = \frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} gij\, e^{-2\pi j\left[\frac{im+jn}{N}\right]}\right|^2$   (4)
          where N = 300, m and n = 1,2,…,$n$.
      4.5 Apply for mel spectrum extraction formula.
          $\tilde{S}(i) = \sum_{k=0}^{N/2} S(k) M_i(k)$   (5)
          Where $i = 0,1,\dots,L-1$
      4.6 Apply for mel cepstrum extraction with formula.
          $c(i) = \sqrt{\frac{2}{L}}\sum_{m=1}^{L} \log(\tilde{S}(m)) \cos\left[\frac{\pi i}{L}(m - 0.5)\right]$   (6)
          Where I = 1,2,…,C – 1; and C is cepstral coefficient desired.
End
```

A set of data is collected after the normalized, the energy level and the zero crossing signals applied for the MFCC extraction. Each of the detected isolated word provided 15 mel-scaled values, that means each complete signal consists of 150 mel-scaled values. The collected mel-scaled data are further used as inputs for neural network training

V. Neural Network Architecture

Backpropagation neural networks are widely used for classification, approximation, prediction, and control problems. Based on biological analogy, neural networks try to emulate the human brain's ability to learn from examples, learn from incomplete data and especially to generalize concepts [9]. The aim of the neural network is to train the net to achieve a balance between the net's ability to respond and the net's ability to give reasonable responses to the input that is similar, but not identical to the one used in the training. A neural network is composed of a set of nodes connected by weights. The nodes are partitioned into three layers namely input layer, hidden layer and output layer. The backpropagation algorithm is most commonly used to derive the weights of the network [9]. The weight adjustment is based on the generalized delta rule. Fig. 8 below shows the architecture of the neural network.

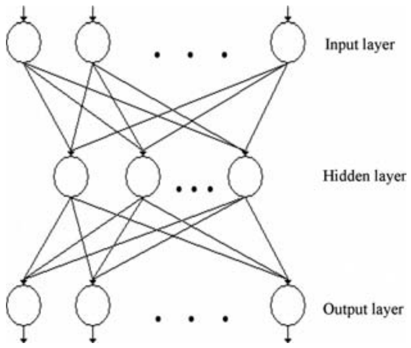

Fig. 8. Neural network architecture

VI. Neural Network Training

It is 150 extreme values are successfully extracted from each speech signal. The input neurons are 150 and output neuron is 1. The input and hidden neurons has a bias value of 1.0 and are activated by binary sigmoidal activation function. A number of 30 hidden neurons are used for the neural network training and the momentum factor value is fixed as 0.80 and the learning rate is fixed as 0.20. The results for training the neural network is tabulated in Table 1 below and the graph of cumulative error plot epoch is shows in Fig 9 below.

TABLE 1. Extracted Mfcc Data Classification

Input Neurons:	150
Hidden Neurons:	30
Output Neuron:	1
Learning rate:	0.2
Momentum factor:	0.8
Training Tolerance:	0.005
Testing Tolerance:	0.1
Function:	*$1/(1 + e^{-x})$*
Training Samples:	100
Testing Samples:	120
Number of Epoch:	50000
Bias:	1.0,1.0.
Failure	0
MinEpoch	48999
Max Epoch	135624
Mean Epoch	96254
Standard Deviation	78957.12
Min Time (s)	480
Max Time (s)	1698
Mean Time (s)	1086
Classification (%)	98.9

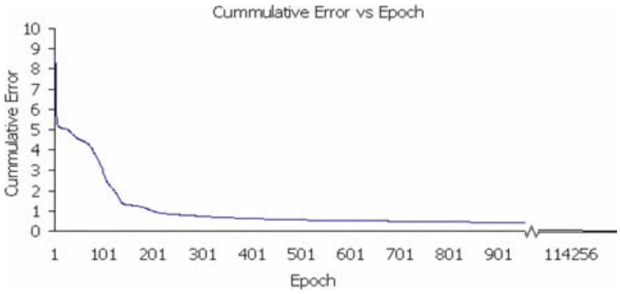

Fig. 9. Cummulative Error versus Epoch (MFCC Data).

VI. Conclusion

In this paper, the combination of energy level and the zero crossing methods are proposed for voiced and unvoiced area detection. Further, the MFCC extraction method is applied to the voiced speech signal. The extracted MFCC data are used as input for the backpropagation neural network training. The training result has shows that the personal identification for speech recognition using backpropagation neural network achieved 98.9% of classification.

References

[1] Mazin, G. Rahim. *Artificial Neural Network for Speech Analysis/Synthesis*. Chapman & Hall, London. 1-14. 1999.Syrdal, A., Bennett, R., & Greenspan, S. *Applied Speech Technology*. CRC Press. 50-51. 1997.

[2] John, G. Ackenhusen. Real-time Signal Processing: design and implementation of signal processing systems. Prentice Hall PTR, Upper Saddle River, New Jersey. 261-289. 2001.

[3] Thomas, E. Quatieri. Discrete-time Speech Signal Processing: Priciples and Practice. Prentice Hall Signal processing Series. 1-3. 2002.

[4] Ben Gold & Nelson Morgan. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, Inc, New York. 280-291. 2000.

[5] Skowronski, M.D. & Harris, J.G. "Acoustics, Speech, and Signal Processing", 2002. *Proceedings. (ICASSP '02), IEEE International Conference* **1**:I-801 - I-804. 2002.

[6] Viikki, O., Bye, D., & Laurila, K. "Acoustics, Speech, and Signal Processing". *ICASSP '98, Proceedings of the 1998 IEEE International Conference* **2**: 733-736. 1998.

[7] Zhu Xuan, Chen Yining, Liu Jia, & Liu Runsheng. "Feature Selection in mandarin large Vocabulary Continuous Speech Recognition". *Signal Processing, 2002 6th International Conference* (ICSP'02). **1**:508 – 511. 2002.

[8] Mark Greenwood and Andrew Kinghorn. "SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech". The University of Sheffield, UK, 1999.

[9] Sivanandam, S.N., Paulraj, *M.*, *Introduction to Artificial Neural Network*, Vikas Publication, New Delhi, (2003)