ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Computer Control with Voice Command (MFCC) using Ad-hoc Network

Pooja, Mukesh Kumar

M.Tech (pursuing), Dept. of Electronics & Communication Engineering, Shri. Ram College of Engineering and Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

Assistant Professor, Dept. of Electronics & Communication Engineering, Shri Ram College of Engineering and Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

ABSTRACT: Despite many years of research, Speech Recognition remains an active area of research in Artificial Intelligence. Currently, the most common commercial mechanism of this technology on mobile devices uses a wireless client – server approach to meet the computational and memory demands of the speech recognition process. Unfortunately, such an approach is unlikely to remain viable when fully applied over the approximately 7.22 Billion mobile phones currently in circulation. In this scheme we present an On – Device Mobile Speech recognition system. Such a system has the potential to completely eliminate the wireless client-server bottleneck vide you can control your machine or laptops from mobile phones using voice commands. For the Voice Activity Detection part of this work, this thesis presents novel algorithms used to detect speech activity within an audio signal. The algorithm is based under the MFCC is augmented in scheme. This algorithm uses the frames within the speech signal with the minimum and maximum standard deviation, as candidates for a linear cross correlation against the rest of the frames within the audio signal. This novel mechanism of the linear cross correlation technique to cepstral coefficients feature vectors provides a fast computation method for use on the mobile platform.

KEYWORDS: , Mel Frequency Cepstral Coefficients (MFCC), Hidden Markov Model (HMM), Dynamic Time Wrapping (DTW), Differential pulse code modulation (DPCM), Linear Predictive Codes (LPC).

I. INTRODUCTION

The Mel Frequency Cepstral Coefficients (mfcc) features are mainly being used as features in speech recognition systems, in some pitch detection algorithms and for identification of the timbre of human voice or musical instruments. In order to compute the mfccs, in a first stage the signal is segmented into small chunks called frames and in every frame a function (often cosine based function such as hamming or Hanning) is multiplied to the signal to avoid discontinuities in the beginning and end of the signal. Such discontinuities might cause extra frequencies in the Fourier transformation that are not present in the original signal.

The two functions that are most commonly used are the Hamming (1) and Hanning (2) and are given from the following two mathematical functions:

$$\omega(n) = a - \beta \cos\left(\frac{2\pi n}{N} - 1\right) \qquad (1)$$

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N} - 1\right)\right) \qquad (2)$$

where, ω is n is the number of bins (time) N is the sum of the bins of the window α = 0.54 β = 1-α = 0.46

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

In a following step, the spectrum is calculated with the use of the fast fourier transform (3) and the spectrum is converted in to Mel bands. The magnitude and the phase of the frequency are shown in equation (5). The transformation from the linear spectrum to the Mel bands is important since the Mel bands describe in a more realistic manner the frequency response of the human auditory system (equation 6).

$$X_k = \sum_{n=0}^{N-1} X_n e^{-i2\pi k/Nn} \qquad (3)$$

$$|X_k| / N = \sqrt{\left(\mathrm{Re}\left(X_k\right)^2\right) + \mathrm{Im}\left(X_k\right)^2} \,/\, N \qquad (4)$$

$$\arg\left(X_k\right) = a\tan 2\left(\mathrm{Im}\left(X_k\right), \mathrm{Re}\left(X_k\right)\right) \qquad (5)$$

Where: n: the number of bins (time) k: frequency index N: the sum of the bins of the window
Re: the real part of the transformation Im: the imaginary part of the transformation

$$m=2595\log_{10}(1+ f / 700) \qquad (6)$$

Where: f = the linear spectrum of the frequency

The next step is to multiply every Mel band with a sum of triangle filters (7) and for every band a cosine transformation is applied (8) and the value that has been extracted from the cosine transformation is the mfcc coefficient.

$$w(n) = 1 - \left|\frac{n-(N-1)/2}{(N+1)/2}\right| \qquad (7)$$

$$X_k = \frac{1}{2}\left(X_0 + (-1)^k X_{N-1}\right) + \sum_{n=1}^{N-2} X_n \cos\left[\pi nk / N - 1\right] \qquad (8)$$

Where: n: the number of bins (time) N: the sum of the bins of the window k = 0, 1,…, N-1

The overview of the feature extraction is presented in the diagram below. The input files are either digitized first if not pre-recorded commands, and split into short-time frames. The 'Cepstral coefficient extraction' returns MFCC coefficients and the Frame Energy. These coefficients and energy is further processed to derive the Delta-Cepstral and Delta-delta Cepstral coefficients.

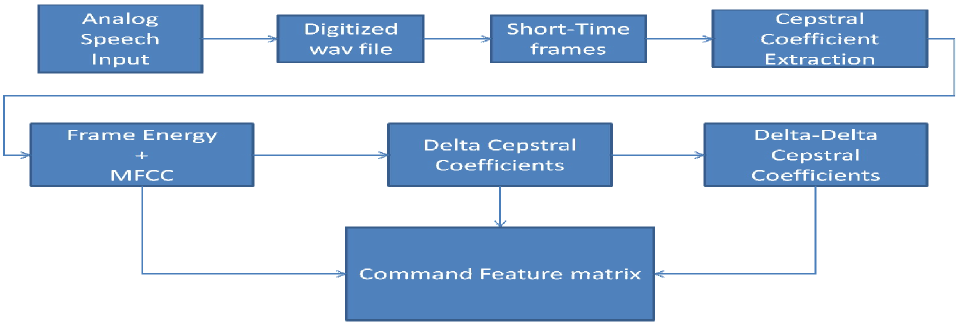

Figure 1.1 - Feature Extraction process

The Cepstral coefficient extraction can be shown in plots in the following section. The block diagram of the cepstral coefficient extraction block is as shown below:

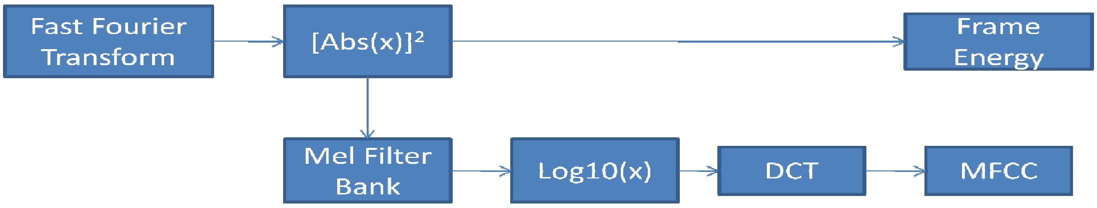

Figure 1.2 - Cepstral coefficient calculation from frames using Fast Fourier Transform

The MFCC dataset is an error correction mechanism which corrects the error produce by the ASR system. The proposed method works on a post editing approach wherein spell checking of the output obtained from the ASR system is done after the input speech is converted into text. The MFCC dataset using Fast Fourier Transform is being used since it is a huge repository of data with data being collected from world web pages and internet documents, which have data ranging from proper nouns, domain specific terms, special expressions, technical words, acronyms and terminologies; covering an ample number of words of the language.

II. RELATED WORK

Fatih AYDINOĞLU [1] depicted that, speech is the most important way of communication for people. Using the speech as interface for processes became more important with the improvements of artificial intelligent. In this scheme it is implemented to control a robot with speech comment. Speech commends were taken to the computer by microphone, the features were extracted with The Mel Frequency Cepstral Coefficients algorithms and they were recognized by the help of Artificial Neural Networks. Finally the comments were converted the form in which the robot can recognize and move accordingly.

P. M. Grant [2] depicted that, transformation of a segment of acoustic signal, by processing into a vectorial representation such as the spectrum, can permit the identification of the constituent phonemes within spoken speech. Subsequent comparison against a previously stored representation using techniques such as dynamic time warping or hidden Markov modeling then permits a speech recognition operation to be accomplished. These signal-processor-intensive transform and graph-search-based pattern-matching techniques are reviewed and currently achievable recognition accuracies are reported.

Sahar E. Bou-Ghazale [3] depicted that, it is well known that the performance of speech recognition algorithms degrade in the presence of adverse environments where a speaker is under stress, emotion, or Lombard effect. This study evaluates the effectiveness of traditional features in recognition of speech under stress and formulates new features which are shown to improve stressed speech recognition. The focus is on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. The stressed speaking styles considered are simulated angry and loud, Lombard effect speech, and noisy actual stressed speech from the SUSAS database which is available on CD-ROM through the NATO IST/TG-01 research group and LDC1 . In addition, this study investigates the immunity of linear prediction power spectrum and fast Fourier transform power spectrum to the presence of stress. Our results show that unlike fast Fourier transform's (FFT) immunity to noise, the linear prediction power spectrum is more immune than FFT to stress as well as to a combination of a noisy and stressful environment. Finally, the effect of various parameter processing such as fixed versus variable pre-emphasis, filtering, and fixed versus cepstral mean normalization are studied. Two alternative frequency partitioning methods are proposed and compared with traditional mel-frequency cepstral coefficients (MFCC) features for stressed speech recognition. It is shown that the alternate filter bank frequency partitions are more effective for recognition of speech under both simulated and actual stressed conditions.

M. Chetouani, B. Gas, J.L. Zarader, C. Chavy [4] depicted that, we present a predictive neural network called Neural Predictive Coding (NPC). This model is used for non linear discriminated features extraction (DFE) applied to phoneme recognition. We also, present a new extension of the NPC model : DFE-NPC. In order to evaluate the performances of the DFE-NPC model, we carried out a study of Darpa-Timit phonemes (in particular /b/, /d/, /g/ and /p/, /t/, /q/ phonemes) recognition. Comparisons with coding methods (LPC, MFCC, PLP, and RASTA-PLP) are presented: they put in obviousness an improvement of the classification.

Raymond Low and Roberto Togneri [5] depicted that, Most commonly used score normalization methods can improve the performance of speaker verification systems, but need extra speech data or cohort models, more memory and computation MIPS. In this paper we present a low-cost adaptive online score normalization (LAOSN) method to improve the performance of speaker verification without any extra data. The computation and memory cost of LAOSN is very small. The procedure begins with initialization of the normalization parameters with existing scores of enrolment utterances from a given enrolment speaker model, and the normalization parameters will be online updated with the scores of subsequent test utterances. By this means, an accurate estimation of the unknown score distribution is archived to normalize current test score. Experiments on the Polycost corpus suggest that the LAOSN can achieve much better performance comparing to the well-known Z-norm method without any extra memory and computation cost.

III. PROPOSED ALGORITHM

The algorithm initially determines the threshold T, which is a percentage of the maximum value of the frame with the maximum cluster measure. This is determined by first looping through the entire frames and selecting the frame with the maximum cluster value. Then a chosen percentage measure ST is used to determine the threshold T for that audio signal. The chosen percentage measure could be determined by the user. Experiments were conducted to determine the optimum generalised threshold to be used for any input signal. After the threshold T is selected, another variable called frame distance is used to separate the frames in the voice cluster (VC) in order to group them into their respective digits. The second determinant variable called the frame distance is used only after determining the threshold value and effectively clustering the different frames into voiced frames and unvoiced frames.

```
For  every frame Fn
        If Fn > Max
          Max = Fn
        End
T = Max x ST
        For every Frame Fn
                If Fn > T
                        Insert Fn into VC
                 Else if
                         Fn < T
                Insert Fn into UC
End Frames
```

- Frames with speech VC
- Frames without speech UC
- Fn = Raw Short time energy of the nth frame
- Max = Value of frame with the highest Cluster value
- T = Threshold
- ST = Selected Threshold (chosen percentage of Max)
- VC = Voice Cluster
- UC = Unvoiced Cluster

IV. PSEUDO CODE

```
Step1: $\eta > 0$, $E_{max.}$ chosen.
Weights W and V are initialized at small random values; W is (K×J), V is (J×I). K represents selected threshold
F represents frame,
Step 2: Training step starts here, input is presented and the layer's output computed [f(net)]
    $f(net) = \frac{2}{1+\exp(-\lambda net)} - 1$    (1)
    $y_j \leftarrow f(v_j^t x), \text{for } j = 1,2,3,...,J$    (2)
Where $v_{j,}$ a column voice cluster, is the j'th row of V, and
    $o_k \leftarrow f(w_k^t y), \quad \text{for } k = 1,2,3,...,K$    (3)
Where $w_{k,}$ a column voice cluster, is the k'th row of W.
Step3: Error value is computed:
    $E \leftarrow \frac{1}{2}(d_k - o_k)^2 + E, \quad \text{for } k = 1,2,3,...,K$    (4)
Step 4: Error signal voice cluster $\delta_o$ and $\delta_{y\ of}$ both layers are computed. voice cluster $\delta_{o\ is}$ (K×1), $\delta_y$ is (J×1). The errors signal terms of the output layer in this step is:
    $\delta_{ok} = \frac{1}{2}(d_k - o_k)(1 - o_k^2), \quad \text{for } k = 1,2,3,...,K$    (5)
The error signal term of the hidden layer in this step is
    $\delta_{yj} = \frac{1}{2}(1 - y_j^2)\sum_{k=1}^{K}\delta_{ok}w_{kj}, \quad \text{for } j = 1,2,3,...,J$    (6)
Step 5: output layer weights are adjusted:
    $w_{kj} \leftarrow w_{kj} + \eta\delta_{ok}y_j, \quad \text{for } k = 1,2,3,...,K \text{ and } j = 1,2,3,...,J$    (7)
Step 6: Hidden Layer weights are adjusted:
    $v_{ji} \leftarrow v_{ji} + \eta\delta_{yj}x_i, \quad \text{for } j = 1,2,3,...,J \text{ and } i = 1,2,3,...,I$    (8)
Step 7: If more patterns are presented repeated by go to step 2 otherwise go to step 8.
Step 8: The training cycle is completed
        For $E<E_{max}$ terminate the training session.
        If $E>E_{max}$ then    $E \longleftarrow 0$, and initiate the new training cycle by going to step 2.
```

V. SIMULATION RESULTS

The next step is to calculate the frame energy. The Fourier transform gives complex values out in its result. In order to make use of those values, they must first be converted to real values. The Absolute value of a complex number returns the magnitude of the complex numbers in the array in real numbers, and the real numbers are squared to calculate the energy. The magnitudes of the Fast Fourier transform are plotted in the figure below:-

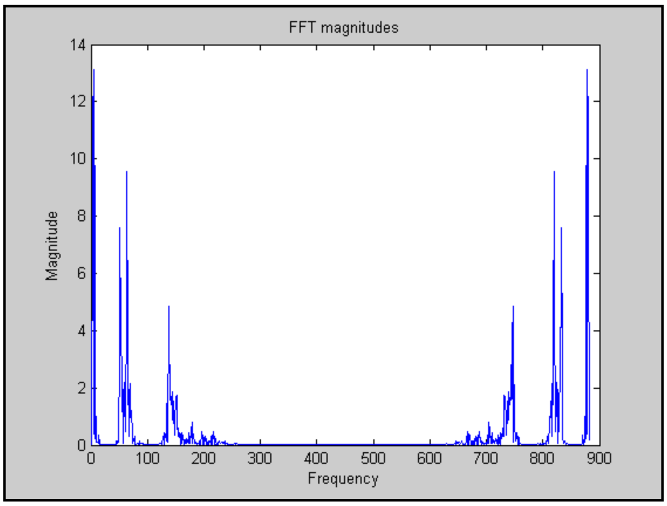

Figure 1.3 - FFT Magnitude spectrum for a sample frame depicting frequency measuring for human sound

These magnitude squares are summed up to form frame energy as one of the parameters. The magnitude squares are also passed downstream to the Mel-Filter banks for further processing. The Mel Filter bank are filters designed based on the Mel Frequency scale. The Mel frequency scale is designed so that it represents the way human percept the sound. The frequency mapping in the Mel scale is shown with respect to the linear frequency below:

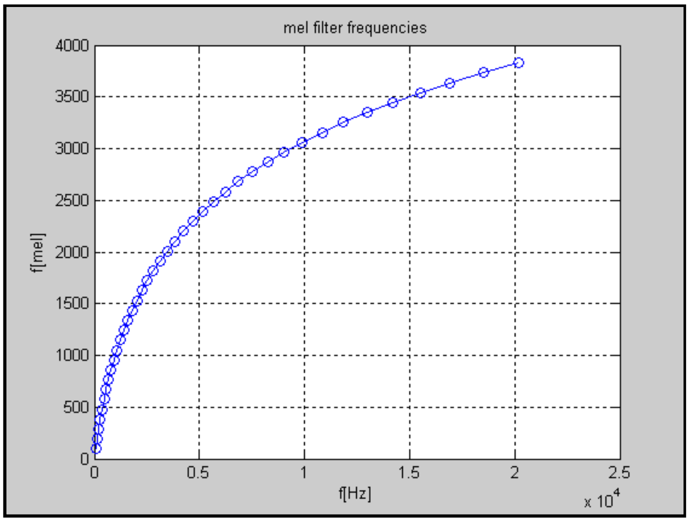

Figure 1.4 - MEL frequency depicting the linear frequency for accurate scale map of speech and vocabulary.

VI. CONCLUSION AND FUTURE WORK

The proposed scheme is implemented using MFCC vide whereas using two different components as server component and client component vide analogy. The client is basically the mobile mechanism that has been created using Java & Android is installed in the mobile phone and the server is in a desktop or laptop computer. The system communicate themselves through the ad-hoc connection and allow the remote controller of the computerize the mechanism to control the machine. It has developed a graphical user interface which will very user friendly and very easy to learn and understand for the end users while the algorithm design for voice integration is done using MFCC context enabling to produce the below mentioned scenarios

.

- High speed and performance.
- Anywhere, anytime access of remote machine by the user.
- One click on most modules allows the administrator to access the remote computer without doing anything on the remote side.
- User friendly and intuitive GUI.
- Intrusion detection mechanism to provide improved security to the user to restrict access to unauthorized users.

For the future the above scheme can be integrated as firmware in the mobiles phone and can shipped to customer ready to use and can also inculcated with security measures to adhere the exclusiveness and classification of user resource.

REFERENCES

1. A Fatih AYDINOĞLU, Robot Control System with Voice Command Recognition, Yıldız Technical University, Department of Computer Engineering
2. P. M. Grant, Speech recognition techniques
3. Sahar E. Bou-Ghazale, Member, IEEE, and John H. L. Hansen, Senior Member, IEEE, A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress
4. M. Chetouani, B. Gas, J.L. Zarader, C. Chavy, Neural Predictive Coding for Speech Discriminant Feature Extraction : The DFE-NPC, Laboratoire des Instruments et Systèmes d’Ile de France
5. Raymond Low and Roberto Togneri, Speech Recognition Using the Probabilistic Neural Network, The University of Western Australia, Department of Electrical and Electronic Engineering
6. M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf and M. A. AlAlaoui, Arabic Speech Recognition Using Recurrent Neural Networks, Faculty of Engineering and Architecture – American University of Beirut
7. Dongsuk YUK, Robust Speech Recognition Using Neural Networks and Hidden Markov Models, The State University of New Jersey
8. Bülent Bolat, Ünal Küçük, Speech Music Classification By Using Statistical Neural Networks, Yıldız Technical University, Department of Electric and Electronic Engineering
9. Paolo Marroe, A Complate Guide All You Need to Know Aboat Joone, 17.1.2007
10. Harshavardhana , Varun Ramesh , Sanjana Sundaresh, Vyshak B N, Speaker Recognition System Using MFCC, A Project Work Of 6th Semester Electronics &Communication Engineering, Visvesvaraya Technological University
11. Audio: A Feature Extraction Library, Daniel McEnnis, Cory McKay, Ichiro Fujinaga, Philippe Depalle, Faculty of Music, McGill University
12. ASR Context- Sensitive Error Correction Based on Microsoft MFCC Dataset, Youssef Bassil and Paul Semmaan, Journal of Computing, Volume 4, Issuel, January 20l2, ISSN 2l52-96l7
13. Discussion and a new attack of the optical asymmetric cryptosystem based on phase-truncated Fourier transform by Xiaogang Wang, Yixiang Chen, Chaoqing Dai, and Daomu Zhao in Sept 20l4
14. An Effective & Coherent Speech Recognition System-Review, Copyright (c) 20l5 CH. Suneetha. Kodell, Vijaya Kumar. Vadladi