Speech Analytics by Generating Mel Frequency Cepstral Coefficient

Engr. Ranil M. Montaril, MS ECE**, Reynald R. Alolor*, Jan Andrew S. Camero*, Jairus Roben T. Catacutan*, John Edwin M. Ibe*, Redentor A. Periabras*, Clint Santos*, and Shekiera Anne O. Soria*

*Undergraduate Students, Bachelor of Science in Computer Science, Polytechnic University of the Philippines
**Faculty Member, Department of Computer Science, Polytechnic University of the Philippines

Abstract— This paper describes an approach of speech analysis by implementing the Mel-Scale Frequency Cepstral Coefficients (MFCC) given a speech signal of spoken words as an input. The tool will be designed to observe the collective use of different disciplines under Signal Processing such as Windowing, Frameshift, FFT, and DCT.

Index Terms— Speech Analytics, Mel-Frequencies, DFT, DCT, MFCC

1 Background of the Study

In recent years, speech analysis has always been an open research area in the field of signal processing. Many research have shown that with this research problem, countless of algorithms and methods could be applied. Mel-Frequency Cepstral Coefficient or more colloquially known as MFCC is one of the most commonly used among others. MFCC can be applied in speech analysis, speech feature extraction, music information retrieval, and genre classification. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound. This feature extraction method that was first mentioned by Bridle and Brown in 1974 is based on experiments of the human misconception of words.

The ideal objective of implementing MFCC is to extract a feature vector containing all information about the linguistic message, in this case, the MFCC is expected to mimic some parts of the human speech production and speech perception. Moreover, the MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. [1]

In this paper, Mel-Frequency Cepstral Coefficient will be implemented in speech analysis. The tool is expected to generate a 26-cepstral coefficient together with its Mel-Filterbank given a sample signal as an input. The tool will be outlined to perform different fundamental disciplines under signal processing, such as Windowing, Fast Fourier Transform, and Discrete Cosine Transform. The tool will take a voice recording as an input and convert these record into its numerical value in a time domain samples. The signal will then undergo signal processing methods in which it will lead to an overall 26-MFCCs.

The 26-mfcc output of the tool may then be applied along with different learning algorithms such as HMM, ANN, or SVM, for the purpose of speech recognition. MFCC alone will fail to recognize a speech signal, but it is observable that it can be effective to aid learning algorithms to analyze speech signals and uncover its linguistic equivalence.

2 Problem Statement

The study aims to perform speech analysis by implementing Mel Frequency Cepstral Coefficients.

Specifically, the paper aims to answer the following questions:

1. What is the effect of change in the value of the frameshift in relationship with the 26-MFCC results?
2. What is the effect of using a windowing function with and without overlapping?

3 Related Studies

In a paper by Huang et al [2], they made an experiment in which they will recognize the voice using Mel-Frequency Cepstral Coefficients (MFCCs) and Dynamic Time Warping; in such a way, the signal analysis by using MFCC provide spectrum factors which represents the exact vocal system for stored words. MFCC provide a high level of perception of the human voice, where they work to remove all unimportant information, then give a better representation of the signal, which leads to a higher resolution in the performance of recognition.

There are many feature extraction techniques like LPC (Linear Predictive Coefficients), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Predictive

Coefficients) and many more are used in Speaker Identification. MFCC gives efficient identification results. Speaker Identifications is affected in terms of noise, sampling rate, and number of frames; wherein noise is the most critical factor. MFCC is not effective in a noisy environment, selectively when the noise condition mismatch; wherein the results of the identification becomes low because the noise level increases. They proposed a technique to improve the performance of Speaker Identification; it includes wiener filter which is good for handling the noise in speech. [4]

According to (Meseguer, 2009), during the MFCC extraction process, much relevant information was lost due to reduction of the spectral resolution in the filterbank analysis and the next truncation into the MFCC components. However, that allowed recovering a smoothed spectral representation in which phonetically irrelevant detail had been removed. [5]

The Mel Frequency Cepstral Coefficients (MFCCs) are widely used in extracting essential information from a voice signal; later became a popular feature extractor used in audio processing. However, MFCC features are usually calculated from a single window (taper) characterized by large variance. [6]

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. [7]

The MFCC feature vector is computed by integration of the spectrum within triangular bins arranged on a mel frequency axis to form a mel frequency binned spectrum, followed by a log and DCT operations. The MFCC vectors and pitch if used, contain sufficient information for continuous speech recognition (at various rates of error). However, it is not obvious that good quality speech can be regenerated from them, since during the feature extraction, a significant amount of information is lost. This includes the phase information, discarded during the spectrum calculation (absolute value of the windowed DFT) and the fine details of the spectrum discarded during the integration. [8]

The feature extraction of speech is one of the most important issues in the field of speech recognition. There are two dominant acoustic measurements of speech signal. One is the parametric modeling approach, which is developed to match closely the resonant structure of the human vocal tract that produces the corresponding speech sound. It is mainly derived from Linear Predictive analysis, such as LPC-based cepstrum (LPCC). The other approach is the nonparametric modeling method that is basically originated from the human auditory perception system. Mel-Frequency Cepstral Coefficients (MFCCs) are utilized for this purpose. In recent studies of speech recognition system, the MFCC parameters perform better than others in the recognition accuracy. [9]

Their paper demonstrated the MFCC speech features extraction method as one of the most commonly used in Automatic Speech Recognition (ASR) systems. Compared to other speech features extraction methods, MFCC is the standard choice for front-end features in state-of-the-art Automatic Speech Recognition (ASR) systems. According to their best knowledge and the review that they performed on the previous studies of Arabic Automatic Speech Recognition (ASR), we found that MFCC dominates the works in this field. [10]

In the paper of (Logan,2000), they examined some of the assumptions of Mel Frequency Cepstral Coefficients (MFCC) – the dominant features used for speech recognition – and examined whether these assumptions are valid for modeling music. MFCCs are short-term spectral-based features. It is clear that the spectral composition of a signal contains much information. This information could certainly be augmented by additional features if required or accumulated over longer time windows. [11]

MFCC is perhaps the best known and most popular, and this feature has been used in their paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves. MFFCs are less susceptible to the said variations.

Several parametric representations of the acoustic signal were compared with regard to word recognition performance in a syllable-oriented continuous speech recognition system. The vocabulary included many phonetically similar monosyllabic words, therefore the emphasis was on the ability to retain phonetically significant acoustic information in the face of syntactic and duration variations. For each parameter set (based on a mel-frequency cepstrum, a linear frequency cepstrum, a linear prediction cepstrum, a linear prediction spectrum, or a set of reflection coefficients), word templates were generated using an efficient dynamic warping method, and test data were time registered with the templates. A set of ten mel-

frequency cepstrum coefficients computed every 6.4 ms resulted in the best performance, namely 96.5 percent and 95.0 percent recognition with each of two speakers. The superior performance of the mel-frequency cepstrum coefficients may be attributed to the fact that they better represent the perceptually relevant aspects of the short-term speech spectrum. [12]

4 Software Design Architecture

The following shows the model designs and the architecture for the Speech Analytics using MFCC.

4.1 Block Diagram

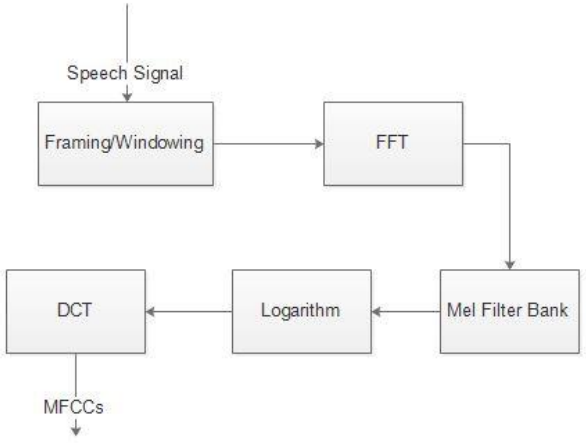


Figure 4.1.1. Block Diagram for Mel Frequency Cepstral Coefficient

Figure 4.1.1 above shows the block diagram for the processes involve in generating MFCCs.

The following steps will show the overview of the processes in generating MFCC; Given a speech signal as an input, frame the signal into short frames. For each frame, calculate the periodogram estimate of the power spectrum by doing Discrete Fourier Transform (DFT) using Fast Fourier Transform (FFT) as an algorithm. Then, apply the mel filterbank to the power spectra and sum the energy in each filter. Take the logarithm of all filterbank energies. Take the Discrete Cosine Transform (DCT) of the log fileterbank energies, the result of these should be the n-MFCCs, having n as the expected number of MFCCs.

4.2 Flowchart

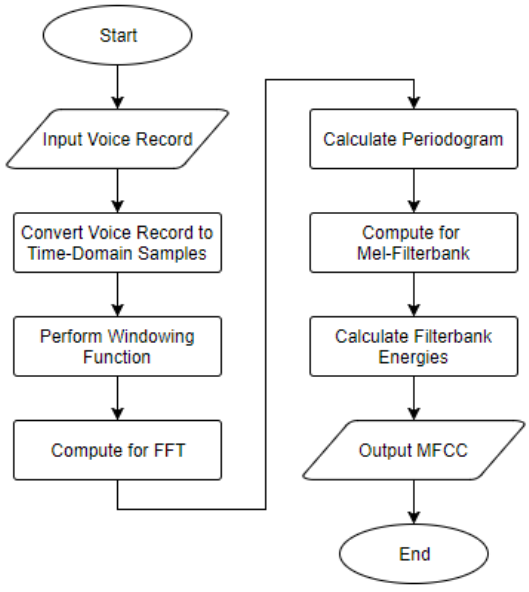


Figure 4.2.1. Flowchart for Mel Frequency Cepstral Coefficient

Figure 4.2.1 above shows the flowchart for the tool implementing MFCC generation.

As an initial process, the tool will take a speech recording as an input. The speech recording will be converted to its corresponding numerical value in a form of a time-domain samples. These samples will undergo windowing, and will be framed accordingly with some frameshift for every window. Each of the samples per frame will be converted to its equivalent Discrete Fourier Transform using FFT. For each of the extracted amplitude of frequency components, calculate the periodogram equivalence. Set a lower and upper limit for the frequency that is going to be used in designing the Mel-Filterbank. Calculate the Mel-Filterbank using the frequencies as an input. Calculate the Filterbank Energies by performing Discrete Cosine Transform in each coefficient in the filterbank. The resulting value of DCT of Filterbank Energies will be the n-MFCCs, having n as the expected number of MFCCs.

5 Simulation Test Results

The following are the simulation test results for Speech Analytics by MFCC. The sample results are in positioned in order of its implementation.

Table 5.1. 26 MFCCs for the word "Power", recorded from three (3) different individuals.

Test Result A	Test Result B	Test Result C
24.67166	25.06929	24.77107
0.521903	0.50182	0.516883
0.752066	0.752066	0.752066
0.741542	0.761626	0.746563
0.443732	0.443732	0.443732

0.954885	0.974968	0.959905
0.335446	0.335446	0.335446
1.051096	1.071179	1.056117
0.256055	0.256055	0.256055
1.077791	1.097874	1.082811
0.271047	0.271047	0.271047
1.086038	1.106121	1.091059
0.262602	0.262602	0.262602
1.095213	1.115297	1.100234
0.254682	0.254682	0.254682
1.102476	1.12256	1.107497
0.231902	0.231902	0.231902
1.133904	1.153987	1.138925
0.229373	0.229373	0.229373
1.128735	1.148818	1.133756
0.231351	0.231351	0.231351
1.125577	1.14566	1.130598
0.234587	0.234587	0.234587
1.118171	1.138254	1.123192
0.242665	0.242665	0.242665
1.133324	1.153407	1.138344
0.224757	0.224757	0.224757
1.136782	1.156865	1.141802

Table 5.1 above shows the test results (in absolute value) for the 26 MFCC output of the word "power" recorded from three (3) different individuals. It is observable that the values analyzed in each column per row are almost equal.

Table 5.2. 26 MFCCs for the word "Digital", recorded from three (3) different individuals.

Test Result A	Test Result B	Test Result C
11.16974	11.35824	11.38567
2.332167	2.341688	2.343074
0.752066	0.752066	0.752066
1.068721	1.078242	1.079628
0.443732	0.443732	0.443732
0.855379	0.8649	0.866286
0.335446	0.335446	0.335446
0.759168	0.768688	0.770074
0.256055	0.256055	0.256055
0.732473	0.741994	0.74338
0.271047	0.271047	0.271047
0.724226	0.733747	0.735132
0.262602	0.262602	0.262602
0.71505	0.724571	0.725957
0.254682	0.254682	0.254682
0.707787	0.717308	0.718694
0.231902	0.231902	0.231902
0.67636	0.685881	0.687266
0.229373	0.229373	0.229373
0.681529	0.69105	0.692436
0.231351	0.231351	0.231351
0.684687	0.694207	0.695593
0.234587	0.234587	0.234587
11.16974	0.701613	0.702999
2.332167	0.242665	0.242665
0.752066	0.686461	0.687847
1.068721	0.224757	0.224757
0.443732	0.683003	0.684389

Table 5.2 above shows the test results (in absolute value) for the 26 MFCC output of the word "digital" recorded from three (3) different individuals. A similar observation from Table 5.1 could be made. The three set of results are all approximation of one another.

Table 5.3. 26 MFCCs for the word "Digital", recorded from one (1) individual using three (3) different Frame Shift as a parameter

Frame Shift 10 ms	Frame Shift 20 ms	Frame Shift 30 ms
-11.2	-12.7	-13.6
-2.33	-2.41	-2.45
-0.75	-0.75	-0.75
-1.07	-1.15	-1.19
-0.44	-0.44	-0.44
-0.86	-0.93	-0.98
-0.34	-0.34	-0.34
-0.76	-0.84	-0.88
-0.26	-0.26	-0.26
-0.73	-0.81	-0.86
-0.27	-0.27	-0.27
-0.72	-0.8	-0.85
-0.26	-0.26	-0.26
-0.72	-0.79	-0.84
-0.25	-0.25	-0.25
-0.71	-0.79	-0.83
-0.23	-0.23	-0.23
-0.68	-0.75	-0.8
-0.23	-0.23	-0.23
-0.68	-0.76	-0.8
-0.23	-0.23	-0.23
-0.68	-0.76	-0.81
-0.23	-0.23	-0.23
-0.69	-0.77	-0.81
-0.24	-0.24	-0.24
-0.68	-0.75	-0.8
-0.22	-0.22	-0.22
-0.67	-0.75	-0.8

Table 5.3 above shows the test result for the different values of Frame Shift applied in the same voice signal. Based on the table above, an increase in the value of the frame shift will produce lesser set of values for the MFCC, and vice versa.

6 Conclusion

Given by the test result (see table 5.1 and table 5.2) of the simulation of the tool, the MFCC result of a particular word will show the distinct numerical representation of the short-term power spectrum of a sound for every single word. Which means, regardless of the speaker's voice parameters, almost same set of numerical values will be generated for each word.

Based on the test result, the 26 MFCC output of the tool is greatly affected by the different parameters (Window shape, frameshift value, sampling frequency, frame size) that is used. For example (see table 5.3), a change in the value of the frameshift shows an inversely proportional relationship between frameshift and MFCC results, a decrease in the value of the frameshift will result into an increase of the value of the MFCC, and vice versa.

Similarly, the implementation of the overlap in the windowing function can result into some changes in the value of the MFCC, this is due to the reason that a windowing function without the implementation of an overlap may encounter spectral leakage, which will lead to a portion of a signal to be loss. To avoid spectral leakage, it is observed that overlapping should be applied.

7 Recommendations

It is recommended for the future researchers to take into account the different parameters that greatly affect the resulting value of the generation of Mel Frequency Cepstral Coefficient. In addition, verifying the result of the MFCC can be very hard unless it is implemented in speech recognition. In order to fully understand the result of the MFCC, an aid of a learning algorithm such as HMM, ANN, and others should be applied.

References

[1] H. Niemann, Klassifikation von Mustern, 2nd ed. Berlin, New York, Tokyo: Springer, 2003.

[2] X. Huang, A. Acero, & H. Hon, "Spoken Language Processing - A Guide to Theory, Algorithm, and System Development", 2001. [Online]. Available: https://www.researchgate.net/file.PostFileLoader.html?id=575be5bb5b4952c0f73e5177&assetKey=AS%3A371736848683010%401465640379360. [Accessed: 14-October-2017].

[3] V.Z. Këpuska, M.M. Eljhani, & B.H. Hight, "Front-end of Wake-Up-Word Speech Recognition System Design on FPGA", 2012. [Online]. Available: https://www.omicsonline.org/open-access/front-end-of-wake-up-word-speech-recognition-system-design-on-fpga-2167-0919.1000108.php?aid=1619. [Accessed: 14-October-2017].

[4] P.M. Chauhan & N.P. Desai, "Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using wiener filter", 16 October 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6921394/. [Accessed: 14-October-2017].

[5] N.A. Meseguer, "Speech Analysis for Automatic Speech Recognition", July 2009. [Online]. Available: https://www.researchgate.net/file.PostFileLoader.html?id=575be5bb5b4952c0f73e5177&assetKey=AS%3A371736848683010%401465640379360. [Accessed: 14-October-2017].

[6] O. Eskidere & A. Gürhanli, "Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features", 18 June 2015. [Online]. Available: https://www.hindawi.com/journals/cmmm/2015/956249/. [Accessed: 14-October-2017].

[7] L. Muda, M. Begam, & I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", March 2010. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1003/1003.4083.pdf. [Accessed: 14-October-2017].

[8] D. Chazan, R. Hoory, G. Cohen, & M. Zibulski, "Speech Recognition from Mel Frequency Cepstral Coefficients and Pitch Frequency", [Online]. Available: https://www.research.ibm.com/haifa/projects/imt/ectts/papers/recon_icassp2000.pdf. [Accessed: 14-October-2017].

[9] A.H.H. Mansour, G.Z.A. Salh, & K.A. Mohammed, "Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithm", April 2015. [Online]. Available: http://research.ijcaonline.org/volume116/number2/pxc3902362.pdf. [Accessed: 14-October-2017].

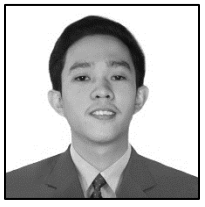
[10] F.S. Al-Anzi & D. AbuZeina, "The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition", 2017. [Online]. Available: http://waset.org/publications/10008047/the-capacity-of-mel-frequency-cepstral-coefficients-for-speech-recognition. [Accessed: 14-October-2017].

[11] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling", October 2000. [Online]. Available: https://pdfs.semanticscholar.org/afe2/38f9ac0678e840ff1521f49c6fe749856109.pdf. [Accessed: 19-October-2017].

[12] M.R Hasan, M.G Rabbani, M.S Rahman, "Speaker identification using mel frequency cepstral coefficients.", 2004, [Online]. Available: https://www.researchgate.net/profile/Golam_Rabbani4/publication/255574793_Speaker_Identification_Using_Mel_Frequency_Cepstral_Coefficients/links/55f05d5908ae0af8ee1d1894.pdf. [Accessed: 19-October-2017].

CURRICULUM VITAE

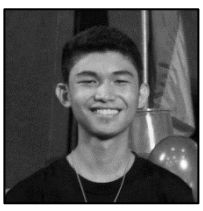
Reynald R. Alolor a graduate of New Era University and is an undergraduate currently taking up Bachelor of Science in Computer Science at Polytechnic University of the Philippines. His research interests include Database Management, Language Modeling, and Operating Systems. He is also an active student that engages in events management within the college.



Jan Andrew S. Camero is an undergraduate student of Polytechnic University of the Philippines, currently taking a Bachelor's degree in Computer Science. An individual that has a stable background with the field of Computing and General Science and equipped with a sufficient skills and knowledge with different programming and scripting languages, specifically with Java, C#, and Matlab. He also has an extensive analysis skills with complicated systems as well as researches.

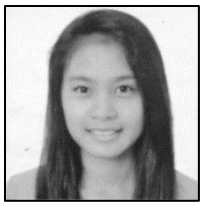
Jairus Roben T. Catacutan lives in Caloocan City, Philippines. He is currently studying for a degree in Bachelor of Science in Computer Science at the Polytechnic University of the Philippines (2014 – 2018). He also graduated from Siena College Quezon City (2010 – 2014). His interests include computational intelligence, modeling and simulation, and computer graphics and visualization.

John Edwin M. Ibe is an undergraduate student of Polytechnic University of the Philippines, pursuing a Bachelor's degree in Computer Science. His field of interest includes Artificial Intelligence, Natural Language Processing, and Software Development. Moreover, he is fond of exercising different programming languages such Python, Java, C#.



Redentor A. Periabras is currently an undergraduate student taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He is pursuing the field of artificial intelligence and data science that satisfies his interests and passion in the field of computing. He is knowledgeable with different programming and scripting languages such as Java, C, Matlab, and other WebDev Scripts

Clint Lennard Santos a student at Polytechnic University of the Philippines is a straight top achiever in elementary and achieved a NCII level at TESDA in Computer Hardware Servicing in high school. He is interested in Web Development and Software Testing. He plays a lot of video games and is also a runner.



Shekiera Anne O. Soria is currently pursuing BSc in Computer Science at the Polytechnic University of the Philippines, Manila. She is a dedicated and enthusiastic learner with enhanced collaboration skills motivated by challenge and has experienced skills in webpage programming, matlab scripting and other languages such as Java and C++.