MFCC and Prosodic Feature Extraction Techniques: A Comparative Study

Nilu Singh
RGU, Itanagar
Department of CSE
INDIA

R.A Khan
Associate Professor & H.O.D
Department of I T
BBAU, Lucknow
INDIA

Raj Shree
Assistant Professor
Department of I T
BBAU, Lucknow
INDIA

ABSTRACT

In this paper our main aim to provide the difference between cepstral and non-cepstral feature extraction techniques. Here we try to cover-up most of the comparative features of Mel Frequency Cepstral Coefficient and prosodic features. In speaker recognition, there are two type of techniques are available for feature extraction: Short-term features i.e. Mel Frequency Cepstral Coefficient (MFCC) and long-term features (Prosodic) extraction techniques. In this paper, we explore the usefulness of prosodic features for syllable classification and MFCC for feature extraction of a speech signal followed by comparison between them. The Me1 Frequency Cepstral Coefficients (MFCC) is one of the most important features extraction techniques, which is required among various kinds of speech applications. The MFCC features are extracted from the speaker phonemes in the pre-segmented speech sentences. Now days Prosodic features are currently used in most emotion recognition algorithms Prosodic features are relatively simple in their structures and known for their effectiveness in some speech recognition tasks. There are various ways of generating prosodic syllable contour features that have recently been applied to enhance systems for speaker recognition.

General Terms

Speaker Recognition, Mel Frequency Cepstral Coefficient (MFCC), Prosodic.

Keywords

MFCC, Prosodic, filter bank, speech feature, Filter bandwidth.

1. INTRODUCTION

Several papers given the concept about speaker recognition system that the speaker recognition systems rely on spectral features extracted from very short time segments (frame) of speech formally known as MFCC. This approach, while highly successful in clean or matched acoustic conditions, suffers significant performance degradation in the presence of handset variability. The approach used in [1][2] Modeling long-range features such as lexical, prosodic, and discourse-related habits, in automatic speaker recognition is motivated for at least three reasons- First, such features can increase performance beyond that of cepstral features. Second, unlike frame-based features, longer-range features reflect voluntary behavior, and as such could potentially be useful not only for recognizing speakers, but also for recognizing characteristics of the speech, such as the speaking style. Finally, regardless of the applied task, research on long-range features should be of fundamental scientific interest to researchers interested in understanding speaking behavior. As the concept known that the physiological structure of a vocal tract is different for every person. Due to this property, we can differentiate one person's voice from others. This difference in vocal tract structure is reflected in the frequency spectrum of speech signal. This speech spectrum is used for speaker recognition [2]. Automatic speaker recognition system can be generally viewed as two main stages: feature extraction and speaker classification. Feature extraction process transforms the raw speech samples into a compact and effective representation which is more stable and discriminative than the original signal. The recognition rate of the classifier strongly depends on the robustness and cue preserving speaker specific characteristics of the features. As we know Mel-frequency cepstral coefficients (MFCC) feature was first proposed for speech recognition. MFCC is a filter bank based approach, the design of filters in such a way that they be similar to the human auditory frequency perception. Researchers have suggested that directly computed filter bank features are more robust for recognition of speech in noisy condition [2]. As the human ear is also a good speaker recognizer, people tried MFCC feature for speaker recognition. Presently MFCC is the most widely used feature for speaker recognition. Prosodic features are the rhythmic and intonational properties in speech, as examples are voice fundamental frequency (F0), F0 gradient (pitch), intensity (energy) and duration. They are relatively simple in structures, and are believed to be effective in some speech recognition tasks [3]. Evaluation for speaker Recognition systems has shown that the use of prosodic information, to enhance acoustic state-of-the-art systems has become very popular. While most participants use classical prosodic features like duration, energy and pitch in a long temporal context [4].

2. MEL FREQUENCY CEPSTRAL COEFFICIENT

For Speaker Recognition a feature extraction technique that extracts both linear and non-linear features is required and here we implement the Mel-frequency Cepstral Coefficients (MFCC). The MFCC is a type of wavelet in which frequency scales are placed on a linear scale for frequencies less than 1 kHz and on a log scale for frequencies above 1 kHz. MFCC is capable to capturing the important characteristic of audio signals [1] [5] [7].The complex cepstral coefficients are called the MFCC. The MFCC contain both time and frequency information of the signal and this makes them more useful for feature extraction. MFCC have widely been used in the field of speech recognition and have managed to handle the dynamic features as they extract both linear and non-linear properties of the signal. MFCC can be a useful tool of feature extraction in vibration signals as vibrations contain both linear and non-linear features [6]. As many studies show that the most common features that are used in state-of-the-art speaker verification/Identification systems are MFCC. MFCC is widely used in Automatic Speaker Recognition systems because of:

- The cepstral features are roughly orthogonal because of the DCT.
- Cepstral mean subtraction eliminates static channel noise.
- MFCC is less sensitive to additive noise than some other feature extraction technique such as linear prediction cepstral coefficients (LPCC).

For feature extraction MFCC have the following steps: Firstly a signal preprocessing is applied on a speech signal. It consists on a pre-emphasis filter to equalize the accurate size. A Hamming Window is applied on each block in order to decrease the edge effects due to the windows cutting. A Fast Fourier Transform is applied on the treated signal and smoothed by a series of triangular filters distributed on a Me1 Scale. The MFCC are then calculated. The scale Me1 calculated using formula –

$$M = \frac{1000}{\log 2} \log\left(1+\frac{f}{1000}\right) \quad \ldots\ldots (1)$$

Where f is the frequency.

MFCC have the following steps when feature extracted from a speech signal-

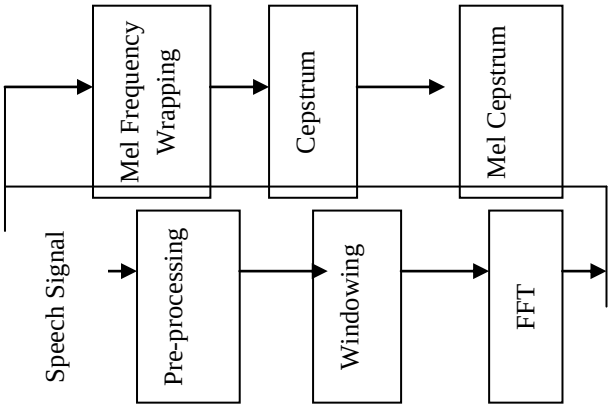

Fig. 1: MFCC steps for feature extraction

The main component of MFCC which is responsible for noise robustness is the filter bank. The filters smooth the spectrum, reducing variation due to additive noise across the bandwidth of each filter [8].

3. PROSODIC FEATURE EXTRACTION

In prosodic there are three main characteristic i.e., pitch, duration, and intensity is taken from any Speech signal for Speaker Recognition. As discussed in many papers [1][9][10] Prosodic speech features, are well known to provide useful information about the speaking style of a person, and thus, are widely-used in speaker recognition applications. A prosodic feature extracted at the syllable level was used for different tasks, like: automatic stress detection, speaker recognition and even language modeling. The most important pitch features are those that capturing pitch level, whereas the most important energy features reflect patterns of rising and falling the energy level. For duration features, nucleus duration is more important for speaker recognition than are durations from the onset of a syllable. Prosodic features contained longer term characteristics because they provide a description of the habitual attributes of the speaker. The main feature Pitch and energy have a robust performance in speaker recognition specially when data noisy and mismatched channels. In addition prosodic feature have speaker specific information, due to vocal folds physical differences between speakers. The unpractical aspect of prosodic features is the high amount of data needed for a successful recognition, also the procedure required to obtain them is complicated and computationally expensive because prosodic features are believed to be carried by syllables in speech segmentation is first done to obtain syllable-like units called pseudo syllables [8][9].

As based on the study of [1][9] Short-term cepstral features are generally referred to as low level features reflecting the voice parameters of the speaker as opposed to higher-level features that capture phonetic, prosodic, and lexical information. Unfortunately, some prosodic features are very hard to compute, while others are inherently difficult to infer solely from acoustics such as lip-roundness. Therefore, higher-level features have increasingly come in use only in the last decade. The relevant subset of studies collected by many research paper [6][9][10] Prosodic features capture variations in intonation, timing, and loudness that are specific to the speaker. As we know such features are supra segmental, i.e., extend beyond one segment, they can be considered a subset of long-term features. Here, mainly pitch and energy dynamics are investigated. There are some other types of prosodic features such as syllable-based prosody sequences, inter pause conversation level statistics, and durational features.

There are many challenges to computing the prosodic features such as the given speech signal-

- Which portion of speech signal useful to find the information.
- What computational model suitable to give better performance of prosodic.
- How much robust and efficient(when use single and when combined with other like cepstral features). Etc

The above discussed points are the major challenge of computation prosodic features. Those prosodic features which based on pitch should be less susceptible to handset and channel effect. Prosodic have the following steps when feature extracted from speech signal-

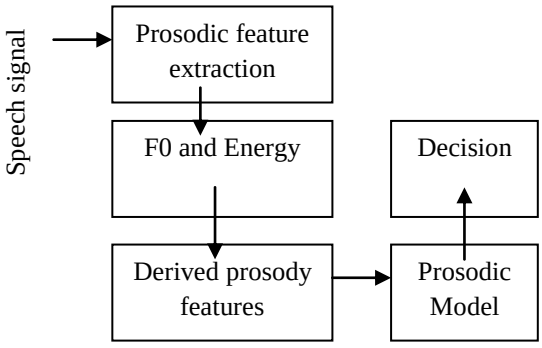

Fig 2: Steps of prosodic feature extraction

The feature extraction model are probabilistic that is use decision trees for recognition rate. After feature extraction once recognition output with detailed time alignments is

available, we can start to model features beyond phones and words. An important aspect of prosodic variation is the duration of speech units.

4. A COMPARATIVE STUDY ON MFCC AND PROSODIC

There are various speech features that were used in the speaker recognition system during the years. Both spectrum-based speech features, related to the shape of the vocal tract, and prosodic features, related to the excitation of the vocal tract and the speaking style of a person. In Speaker Recognition the extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it drastically affects the recognition performance [7]. There are many studies have been done on MFCC and Prosodic which tell us that the MFCC use A small set of standard features while Prosodic uses long term features. In the field of Speaker Recognition System Different studies have been done to use dynamic information contained in speech. The 'high level' also called longer term speaker dependent features, as prosodic, phonetic and linguistic, But these require a lot of speech samples and are also time consuming and computationally complex. As discussed in [9] as compared to short-term cepstral features in speaker recognition, a number of long-term features can provide more significant information for speaker discrimination. As already suggested by, looking at patterns derived from a larger segment of speech can reveal individual characteristics of the speakers' voices as well as their speaking behavior, which cannot be captured by exclusively using frame-based short-term cepstral analysis .

If we talk about limitation of MFCC, a serious limitation of the original MFCC feature extraction technique is that the filter bandwidth is not an independent design parameter but instead is determined by the frequency range of the filter bank and the number of filters used in the recognition. Therefore the number of filters can be ads or subtracts to provide the sampling frequency of experiments. The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by the number of filters and type of window [2] [7] i.e. accuracy rate fluctuate to increase the number of MFCC coefficient and also to decrease the MFCC coefficient. In addition too few or too many filters do not result in better accuracy. Also is described in [7] that efficiency is maximum while using hanning window.

Based on the studies the conclusion is that the prosodic parameters, especially in the manner in which they were modeled in the past, are also much easier to mimic than the parameters that describe the characteristics of the vocal tract's filter function. Thus, in real-world deployment of the speaker verification technology, that's why the use of prosodic parameters has to be cautious, especially in applications where the risk of fraud attempts is significant, or where the consequences of false acceptance of impostors as clients are costly.

On the basis of comparative study of many papers [2][9][12] It was shown that systems using a combination of cepstral and higher-level features outperformed standard systems, especially when the amount of available training data was increased. This confirms the assumption that short-term cepstral systems generally perform well because they reflect information about the speaker's physiology and do not rely on the phonetic content. However, long-range information that also resides in the signal is only exploited in the combined systems. As pointed out by [3] that the higher-level features also have the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions.

As discussed in [9] long-term features have been investigated for several years and indications have been provided that they can be useful for speaker recognition. Now days, speaker verification systems using different kinds of prosodic features have been proposed. Although it has been shown that most of these speaker verification systems can improve system performance using score-level fusion with state of- the-art cepstral based systems, a systematic comparison of the prosodic modeling algorithms used in these prosodic systems has not yet been performed [8][9][11].

Many research paper [9][10] which describe that It is believed that prosodic features are less vulnerable to the channel distortion than cepstral features (MFCC). On the other hand prosodic features alone cannot perform as well as cepstral features, the fusion of these two types of features has been proposed to further improve the performance of conventional cepstral-based speaker verification systems.

Cepstral based features, which typically represent the magnitude properties of speech spectrum, are widely used in speech processing as well as in speaker recognition. Choosing effective features is important to achieve a high performance, and the most popular cepstral features technique is MFCC. There are some research papers which is shows that system achieves a slightly better performance with MFCC than other feature extraction technique.

Prosodic features i.e. Pitch and energy contours of speech, are known to give information about the Speaker Identification/Verification. As discussed in [4][9][17] reported on the use of pitch parameters in speaker recognition in the 1970's and early 1980's. However interest in research in the use of prosodic features appears to have diminished in recent years because these features alone could not give the level of performance required for speaker identification and verification in text dependent systems and it was difficult to see how they could be incorporated in a text independent system, Pitch extraction was also error prone and computationally expensive. The channel distortions and noise is the serious problem in Speaker Recognition. Prosodic features are known to be less effected by these problems than spectral features such as the low order cepstral coefficients. Prosodic features are therefore worth re-examining for speaker identification particularly when used to improve the performance of algorithms using Hidden Markov Model techniques.

As disused in [2][7][13] The average accuracy of MFCC features set was above than 95%, And using Prosodic features set of speech signal the average accuracy rate was 69%. There are many studies shows that on the basis of acoustics analysis based on the MFCC, which represent the ear model, have proved good results in speaker recognition mainly when a high number of coefficient is used**.** In addition, it is considered the most successful speaker recognition system when present to different variations such as: prosody, intonation, noise etc. It executes also the task of filtering, modeling, and processing, decoding, phonemes or words and languages distinction. The table shows that some parameters which differentiate MFCC and Prosodic in such a way that is some where cepstral feature (short term feature) is better and some where long term feature or prosodic gives the better

result. We have discussed some parameters of MFCC and Prosodic in the below table.

Table 1. MFCC & Prosodic: A comparative Chart

Parameter	MFCC	Prosodic
Vocal Tract	Depends on shape of the vocal tract	Excitation of the vocal tract and the speaking style
Features set	Uses a small set of standard features	Uses long term features
Features type	Uses Cepstral features	Uses Non cepstral features
Secure	Not easy to mimic	easier to mimic
Filters	Uses filter bank	Does not use filters
Speaker Verification/Identification	It give better results for both i.e. SI/ SV	It give better result for SV
Channel effect	Cepstral features affected by the channel distortion.	It is believed that prosodic features are less vulnerable to the channel distortion.
Performance	MFCC lonely able to perform well.	Prosodic features alone cannot perform well.
Speech Sample	It requires less Speech sample less time and not so computationally complex.	It require a lot of speech samples and are also time consuming and computationally complex

The above table have mentioned some basic parameters, from which we able to distinguish between Mel frequency Cepstral Coefficient and prosodic.

5. SUMMARY

In research of Speaker Recognition this is found that there are many factors affecting the speaker recognition efficiency such as voice variation, channel mismatch, different handset/microphone etc. In these factors the noise has the strongly affecting factor in the performance of Speaker Recognition efficiency [15][16]. The Investigation also shows that there are some other parameters which affecting performance of speaker recognition are given below-

- Spoken language used at the time of training and testing data.
- Quality of voice sample data in training and testing.
- Quality of microphone and distance of microphone from speaker.
- Noise at the time of recording voice sample and testing voice sample.
- Length of the voice sample that is used in training and testing.
- Cover up on microphone or speaker at the time of training and testing.
- Text – dependency i.e. training and testing data is same.
- Variation in speaker voice.

The performance of speaker recognitions is affected by the above mentioned factor and some other factors also. Variation in speaker voice is also the major factor affected by speaker recognition.

6. CONCLUSION

In this paper we try to cover up the basic differences of MFCC and Prosodic. This work reflects the results obtained in the evaluation of a prosodic and MFCC features, in this study we present the feature extraction techniques for speaker recognition were discussed MFCC and Prosodic. The conclusion is that on the basis of researcher's point of view and practical implementation, MFCC is better than prosodic and well known techniques used in speaker recognition to describe the signal characteristics, relative to the speaker discriminative vocal tract properties.

And also the concept given that, those systems using a combination of cepstral and higher level features outperformed standard systems, especially when the amount of available training data was increased. This confirms the assumption that short-term cepstral systems generally perform well because they reflect information about the speaker's physiology and do not rely on the phonetic content. However, long-range information that also resides in the signal is only exploited in the combined systems. As the study shows that MFCC is better but, higher-level features also have the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions.

7. REFERENCES

[1] Shriberg, E, L Ferrer, S Kajarekar, A Venkataraman, and A Stolcke. "Modeling prosodic feature sequences for speaker recognition." 46 (2005): 455–472. Print.

[2] Sen, Nirmalya, T.K Basu, and Hemant.A. Patil. "New Features Extracted from Nyquist Filter Bank for Text-Independent Speaker Identification." *Annual IEEE India Conference (INDICON)*. 978-1-4244-9074-5/10. (2010): 1-5. Print.

[3] W. M. Ng, Raymond, tan Lee, Cheung Chi Leung, Bin Ma, and Haizhou Li." *Analysis and Selection of Prosodic Features for Language Identification*. 978-0-7695-3904-1/09. (2009): 123-128. Print.

[4] Kockmann, Marcel, Lukas BurgetLast, and Jan Honza Cernocky. "INVESTIGATIONS INTO PROSODIC SYLLABLE CONTOUR FEATURES FOR SPEAKER RECOGNITION." *ICASSP 2010*. 978-1-4244-4296-6/10.2010 (2010): 4418-4421. Print.

[5] Abdulaziz, Yousra, and Sharrifah mumtazah Syed Ahamad. "Infant cry recognition System:A comparision of System Performance based on Mel Frequencyand Linear Prediction cepstral coefficient." *IEEE*. 978-1-4244-5651-2/10. (2010): 260-263. Print.

[6] Nelwamondo, Fulufhelo V., and Tshilidzi Marwala. "Faults Detection Using Gaussian Mixture Models, Mel-Frequency Cepstral Coefflcients and Kurtosis." *2006 IEEE International Conference on Systems, MAan, and Cybernetics October 8-11, 2006, Taipei, Taiwan*. 1-4244-0100-3/06. (2006): 290-295. Print.

[7] SEDDIK, HASSEN, AMEL RAHMOUNI, and MOUNIR SAYADI."TEXT INDEPENDENT SPEAKER RECOGNITION USING THE MEL FREQUENCY CEPSTRAL COEFFICIENTS AND A NEURAL NETWORK CLASSIFIER." *ieee*. 0-7803-8379-6/04. (2004): 631-634. Print.

[8] Geravanchizadeh, Masoud, and Amir Karirnpour. "Improving the Noise-Robustness of Mel-Frequency Cepstral Coefficients for Speaker Verification." *Proceedings of the 4th International Symposium on Communications, Control and Signal Processing, ISCCSP 2010, Limassol, Cyprus , 3-5 March 2010*. 978-1-4244-6287- 2 /10. (2010): 1-4 . Print.

[9] Friedland, Gerald, Oriol Vinyals, Yan Huang, and Christian Miiller. "Prosodic and other Long-Term Features for Speaker Diarization." *IEEE Transaction on Audio, Speech and Languages Processing*. Vol. 17, NO. 5. (july 2009): n. page. Print.

[10] Chi Leung, Cheung, Marc Ferras, Claude Barras, and Jean Luc Gauvain. "Comparing Prosodic Models for Speaker Recognition." *ISCA*. September 22-26. (2008): 1945-1948. Print.

[11] Ananthakrishnan, Sankaranarayanan, and Shrikanth S. Narayanan. "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence." *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. VOL. 16, NO. 1.2008 (2008): 216-228. Print.

[12] Huang, Zhongqiang, Lei Chen, and Mary P. Harper. "Purdue Prosodic Feature Extraction Tool on Praat." *Spoken Language Processing Lab School of Electrical and Computer Engineering Purdue University, West Lafayette*. (2006): 1-35. Print.

[13] Jayanna, HS , and SR Mahadeva Prasanna. "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition." *Academic Journals in U.S.*. 26.3 (2009): 181-190. Print.

[14] Ferrer, Luciana, Nicolas Scheffer, and Elizabeth Shriberg. "A COMPARISON OF APPROACHES FOR MODELING PROSODIC FEATURES IN SPEAKER RECOGNITION." *ICASSP 2010*. 978-1-4244-4296-6/10. (2010): 4414-4417. Print.

[15] Huang, Zhongqiang, Lei Chen, and Mary Harper. "An Open Source Prosodic Feature Extraction Tool." *School of Electrical and Computer Engineering Purdue University West Lafayette, IN 47907*. (2006): 1-6. Print. <ftp://ftp.ecn.purdue.edu/ harper>.

[16] Singh, Satyanand , and Dr. E.G Rajan. "MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors." *International Journal of Computer Applications (0975 – 8887)*. 21.6 (2011): 1-6. Print.

[17] Ezzaidi, Hassan, Jean Rouat, and Douglas O' Shaughnessy. "Combining pitch and MFCC for speaker recognition systems." *NSERC, Communications Security Establishment and the FUQAC*. (2001): 1-6. Print.