International Journal of Advanced Science, Engineering and Technology.
ISSN 2319-5924
Vol 1, Issue 1, 2012, pp 26-28
http://www.bipublication.com

FRACTIONAL FOURIER TRANSFORM COMBINATION WITH MFCC BASED SPEAKER IDENTIFICATION IN CLEAN ENVIRONMENT

[1]Upendra Kumar Agrawal, [2]Mahesh Chandra, [1]Chandrkant Badgaiyan

[1]Dept. of Mechatronics, CSIT Durg, CG. upeagrawal@gmail.com
[2]Dept. of Electronics & Communication Engg. BIT Mesra,
Ranchi, Jharkhand. shrotriya@bitmesra.ac.in

[Received-21/09/2012, Accepted-09/10/2012]

ABSTRACT

The Fractional Fourier Transform FrFT transform order for the proper analysis of multi-component signals like speech is still debated. In this paper we have compared as well as verified the technique of speaker identification using FrFT combined with Mel-frequency cepstral coefficients (MFCC). The FrFT with MFCC shows poor result as compare to MFCC with FFT. In text- independent case the difference in result is 3.41%, while in the text-dependent case difference becomes 13.62%. This is because FrFT emphasis only harmonics of speech signal.

Keywords- *DCT; FFT ; FrFT; LFM; GMM; MFCC*

I. INTRODUCTION

Speaker Recognition has two parts speaker identification and speaker verification [1][2] . The process of determining to which of the registered speakers a given utterance belongs is speaker identification. FrFT [3] combined with MFCC [4] and FFT combined with MFCC are used to identify the speakers. Fast Fourier transform is not suitable for signals whose frequencies are changing with time. The drawback in FFT is that it assumes that signal is stationary in nature. From the last few years, properties of FrFT have been exploited in the field of optics and sonar signal processing. It is mainly useful for chirp signals because FrFT signal can be taken as a decomposition of the signal in terms of chirps. The voiced speech is almost harmonic in nature adjusting the order of the FrFT to be some value associated with harmonic or formants might improve the representation of the time-varying properties of speech. It will also useful in tracking the dynamic properties of speech harmonics.

II. AUDIO FEATURE EXTRACTION

The non-relevant information such as background noise is removed from the speech signal in the preprocessing stage.

This involves pre-emphasis, framing and windowing. After this FrFT is applied to the windowed signal and features are extracted using MFCC. Log makes the shape of speech spectral insensitive to different loudness levels. Mel scaled log energy features are highly correlated. So, DCT is applied to make the features independent. The whole process of Feature extraction has been shown in Figure 1.

FRACTIONAL FOURIER TRANSFORM

The FrFT of signal x(t) is represented as:

$$X_p = F_p[x(t)] = \int_{-\infty}^{\infty} x(t) K_\alpha(t,u)\,dt, \quad (1)$$

Where p is a real number and called the order of the FrFT, $\alpha = p\pi/2$ is the tranform angle, $F_p = [\blacksquare]$ denotes the FrFT operator, and $K_\alpha(t,u)$ is the kernel of the FrFT:

$$K_\alpha(t,u) = \begin{cases} \sqrt{\frac{1-j\cot\alpha}{2\pi}} \exp\left(j\frac{t^2+u^2}{2}\cot\alpha - jut\csc\alpha\right), & \alpha \neq n\pi \\ \delta(t-u), & \alpha = 2n\pi \\ \delta(t+u), & \alpha = (2n \pm 1)n\pi \end{cases} \quad (2)$$

The Kernel has the following properties:

$$K_{-\alpha}(t,u) = K_\alpha^*(t,u) \quad (3)$$

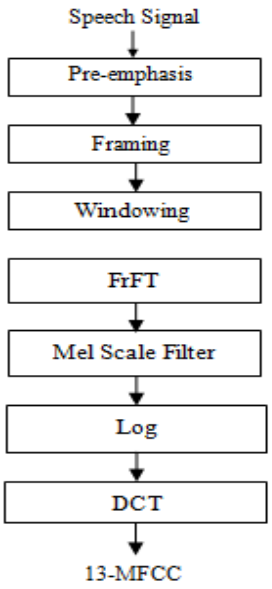

Figure 1: MFCC + FrFT Based Feature Extraction

$$\int_{-\infty}^{\infty} K_\alpha(t,u) K_\alpha^*(t,u')\,dt = \delta(u-u') \quad (4)$$

Hence the inverse FrFT is

$$x(t) = F_{-p}[X_\alpha(u)] = \int_{-\infty}^{\infty} X_\alpha(u) K_{-\alpha}(t,u)\,du, \quad (5)$$

Eq.(5) indicates that signal $x(t)$ can be interpreted as a decomposition to a basis formed by orthonormal Linear Frequency Modulated (LFM) [5] functions in the u domain, which means an LFM signal with a chirp rate corresponding to the tranform order p can be tranformed into an impulse in a certain fractional domain. Therefore, the FrFT has excellent localization performance for LFM signals.

III. RESULTS AND DISCUSSIONS

A database of fifty speakers, twenty one males and twenty nine females, for a total of ten isolated Hindi digits ("shunya", "ek", "do", "teen", "chaar", "paanch", "chheh", "saat", "aath" and "nao") has been prepared with sampling frequency of 16 kHz and 16 bits per sample. Ten samples of each Hindi digit for all fifty speakers were recorded in order to prepare the database. Speakers from different social classes and of different age groups (18-26 years) were chosen. In the training stage features are calculated frame wise. For each speaker features are derived from all the 10 samples of all the 10 digits (shunya -nau). In the next step GMM [6] model is prepared $(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$GMM models are prepared for all 50 speakers. Now, every speaker has its own independent GMM model. Testing is performed by taking first twenty frames (approx. 25 ms) of second sample of all the 10 Hindi digit of all the 50 speakers (i.e. 20x10x50=10000 features) were used for text-dependent speaker identification. In the text-independent case first nine digits are used for training purpose and the last one is used for training. Then these features were tested with the 50 GMM models prepared in the earlier stage of training. Now, the maximum log likelihood [7] is calculated for every testing

feature with the GMM model and decision is made about the speaker recognition. The rate of speaker identification is calculated based on this formula given below

$$\% \text{ Speaker Identification} = \frac{\text{total no of feature vectors} - \text{no of incorrect identified feature vectors}}{\text{total no of vectors}} \times 100$$

Result for speaker identification for both text independent and text dependent case are shown in table 1.

IV. CONCLUSION

The FFT combined with MFCC giving better results as compared to FrFT with MFCC. Reason for such difference is due to the effectiveness of the FFT analysis on formant determination. The speech signal is combination of chirp signals and in FrFT based method only harmonics of speech signal is taken into account. Also the optimal transform order is a crucial factor in FrFT based detection. Considering the effectiveness of the FFT analysis on formant determination and of the FrFT analysis on emphasizing harmonics, one possible approach is to combine the FFT and FrFT to get an improved representation of speech features for speech analysis and recognition.

REFERENCES

[1] B.S. Atal, "Effectiveness of Linear Prediction Characteristic of the Speech Wave for Automatic Speaker Identification and Verification," Journal of the Acoustical Society of America,Vol. 55, No. 6, pp. 1304–1312, 1974.

[2] Tomi Kinnunen, Evgeny Karpov and Pasi Franti "Real-time speaker identification and verification", IEEE Transaction on Audio, Speech and Language Processing, Vol. 14, No. 1, pp. 277-278, 2006.

[3] Wei-Qiang, Zhang Liang He, Tao Hou and Jia Liu, "Fractional Fourier transform based auditory feature for language identification", *IEEE Asia Pacific Conference on Circuits and Systems*, pp. 209-212, 2008.

[4] Ezzaidi H et al, "Pitch and MFCC dependent GMM models for speaker identification system", *Canadian conference on Electrical and Computer Engineering*, vol. 1, pp 43-46, 2004.

[5] Qi Lin, Tao Ran, Zhou Si-yong, "Detection and parameter estimation of multicomponent LFM signal based on the fractional Fourier transform", Science in China, 47(2), 184-198, 2004.

[6] Douglas A. Reynolds, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[7] J.D. Wise et al, " Maximum likelihood pitch estimation," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-24, no. 5, pp.418-423, October 1976.

Table 4.1 % Speaker Identification using FFT & FRFT with 50 Speakers

Feature Extraction Technique	Text dependent	Text independent
MFCC+FFT	92.60 %	81.40 %
MFCC+FRFT	78.98 %	77.99 %