FRACTIONAL FOURIER TRANSFORM FEATURES FOR SPEECH RECOGNITION

Ruhi Sarikaya, Yuqing Gao and George Saon

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{sarikaya,yuqing,saon}@us.ibm.com

ABSTRACT

In this paper a novel speech signal representation method is presented. The proposed method is based on the Fractional Fourier transform (FrFT), which is a generalization of the classical Fourier transform (FT). Even though we use FrFT in feature extraction for speech recognition, it can very well be used in other areas such as enhancement, verification, and synthesis, where parametric representation of speech is needed. Experimental results conducted on the Aurora 2 database show significant improvements over MFCCs at high SNR conditions.

1. INTRODUCTION

There has been a significant amount of effort devoted to improving speech feature extraction. Even though a considerable number of alternative processing schemes have been proposed, Mel–frequency cepstral coefficients (MFCC) have remained as the most widely used feature extraction method [3]. MFCCs are computed in several steps. First, a Discrete Fourier Transform (DFT) of a frame of speech is computed to obtain the magnitude spectrum. Next, the magnitude spectrum is frequency–warped in order to transform the spectrum into Mel frequency where the filterbank is uniformly spaced. Then, filters are multiplied with the power spectra of the frame to compute the energy in each filter of the filterbank followed by the logarithmic compression. Finally, the discrete cosine transform (DCT) of the filterbank log energies are computed resulting into MFCCs. Linear Cepstral (LC) features are also obtained from the power spectrum by applying log–compression and DCT without Mel–warping and filterbank energy computation.

Speech is modeled not only across frequency but also across time. Modeling across frequency reflects properties of human auditory system as in the case of MFCC. Temporal properties are modeled by dynamic features or temporal filtering. Nevertheless, intonation and coarticulation introduce combined spectro–temporal fluctuations to speech even for the typical frame sizes used in the frontend analysis. For example, formant transitions exhibit diagonal trajectories.

The goal of this study is to introduce a new speech signal analysis tool called fractional Fourier transform (FrFT), which covers classical Fourier transform as a special case. To the best of our knowledge, the FrFT has not been applied to speech processing, recognition and enhancement before. While we use it to extract features for speech recognition, the use of FrFT should be of general applicability to a variety of problems within speech and audio processing. Moreover, digital implementation of FrFT is as efficient as that of the classical FT in the sense that it can also be computed in the order of $N \log N$ time, where N is signal sample length.

The FrFT has found many applications in the solutions of differential equations, quantum optics, optical and sonar signal processing [8, 4, 2, 9]. Its relationship to wide range of concepts has been established and it has been employed in conjunction with a variety of techniques [2, 4].

The Fourier analysis is one of the major tools used in signal processing, and in many other disciplines. Although Fourier transform is well suited for analysis and processing of time-invariant signals and systems, it can not achieve comparable results when the signal and/or system are time-varying. The classical Fourier analysis results in the frequency components of a signal. The fractional Fourier transform (FrFT) can reveal the mixed time and frequency components of signals [1]. For time-varying signals, filtering or processing in fractional Fourier domains might allow us to estimate the signal with smaller minimum square error (MSE) for certain classes of signals.

The rest of the paper is organized as follows. In Section 2, we explain the concept of fractional Fourier transform along with some important properties. The basics of discrete fractional Fourier transform is described in Section 3. Section 4 presents the experimental results and discussion. Finally, Section 5 summarizes the findings and possible future research directions.

2. THE CONCEPT AND DEFINITION OF THE FRACTIONAL FOURIER TRANSFORMATION

Time and frequency domains can be visualized as shown in Fig. 1, where the two domains are orthogonal. Each application of the Fourier transform rotates the representation of time domain signal $x(t)$ by $\pi/2$ in the counterclockwise direction.

$$F^1[x(t)] = X(w), \quad F^2[x(t)] = x(-t), \quad F^4[x(t)] = x(t) \qquad (1)$$

where F denotes the FT operator and $X(w)$ is the FT of $x(t)$. Hence, repeated applications of F corresponds to successive rotations of $\pi/2$. In this context, one could ask what linear operator corresponds to a rotation of ϕ that is not a multiple of $\pi/2$ as shown in Fig. 1. For the moment, let us assume that such an operator exists. This operator should possess the following properties as well:

1. $[F^a]^{-1} = F^{-a} = [F^a]^*$, where $(.)^*$ denotes Hermitian conjugation.
2. $F^\alpha F^\beta = F^{\alpha+\beta}$ (Additivity of rotations), α and β denote the rotation angles.

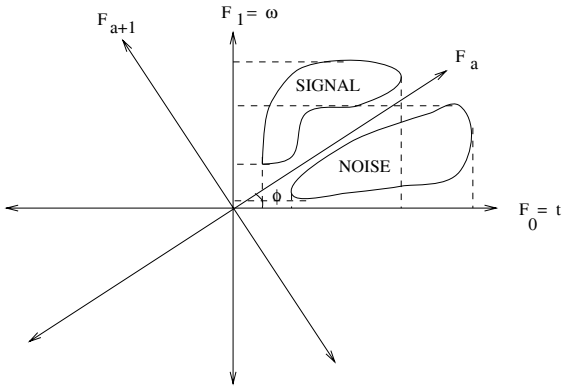

Figure 1: The fractional Fourier domain.

3. $F^{\pi/2} = F$ (Consistency with the FT)
4. $F^0 = I$ (Zero Rotation), I denotes the identity operator.
5. $F^{2\pi} = I$ ($2\ \pi$ rotation)
6. Linearity

The existence of such an operator satisfying the properties mentioned above, is shown in [5, 1] and named as Fractional Fourier Transformation. The transformation and the kernel of this transformation is given below, respectively, for $0 < |a| < 2$,

$$F^a[f(t_a)] \equiv \int_{-\infty}^{\infty} B_a(t_a, t) f(t) dt \tag{2}$$

$$B_a(t_a, t) \equiv A_\phi \exp[j\pi(t_a^2 \cot\phi - 2t_a t \csc\phi + t^2 \cot\phi)] \tag{3}$$

$$A_\phi \equiv \frac{\exp(-j\pi \mathrm{sgn}(\sin\phi)/4 + j\phi/2)}{|\sin\phi|^{1/2}} \tag{4}$$

where $\phi \equiv \frac{a\pi}{2}$ and j is the imaginary unit. The definition can be easily extended outside the interval, $0 < |a| < 2$, by using properties 2 and 5. One can analytically verify that $B_a(t_a, t)$ has the following properties:

$$a \to 0 \ \Rightarrow\ B_a(t_a, t) \to \delta(t_a - t) \tag{5}$$

$$a \to \pm 2 \ \Rightarrow\ B_a(t_a, t) \to \delta(t_a + t) \tag{6}$$

$$a \to 1 \ \Rightarrow\ B_a(t_a, t) \to \exp(-j2\pi t_a t) \tag{7}$$

The kernel has the following spectral expansion [8]:

$$B_a(t_a, t) = \sum_{k=0}^{\infty} \psi_k(t_a) e^{-j\frac{\pi}{2}ka} \psi_k(t) \tag{8}$$

where $\psi_k(t)$ is the kth Hermite–Gaussian function and t_a is the variable in the ath-order fractional Fourier domain. It is well known that the eigenvalues of the Fourier transform (FT) are $\lambda_k = e^{-j\frac{\pi}{2}k}$ and the corresponding eigenvectors are Hermite functions, which are normalized Hermite polynomials weighted by the factor $e^{-\frac{t^2}{2}}$. Note that $e^{-j\frac{\pi}{2}ka}$ is the ath power of the eigenvalue λ_k of the classical Fourier transform. Setting $a = 0$ results in the identity transform and $a = 1$ to classical FT. Because of the additional parameter a, whose optimal value will in general be different from $a = 1$ for different conditions, the FrFT is much more flexible and will in general offer better performance except in the special case when the optimal value coincidentally turns out to be equal to 1.

In addition to those properties listed above there are a number of additional properties of FrFT [1, 2, 5]. One interesting property is about rotation which states that the Wigner–Distribution (WD) of the ath order FrFT of a signal is the same as the WD of the original signal rotated counter clock–wise by an angle of $a\pi/2$ radians in the time–frequency plane [8]. This property may allow time domain warping concept, which is widely used in speech and speaker recognition, to be applied in fractional domains. Potentially, different interference, noise and other sources of variability may be beneficial to suppress in different fractional domains. The conceptual plot in Fig. 1 that is adopted from [4] shows the potential advantages of processing in the fractional domains. In the figure, the signal and noise representations overlap both in time and frequency domains. However, they may potentially be separated or have a smaller overlap in some fractional domains. Moreover, certain phonetic classes may have better representations in different fractional domains.

3. DISCRETE FRACTIONAL FOURIER TRANSFORM

The classical discrete Fourier transform (DFT) of a signal $f(n)$ can be represented in matrix notation as:

$$\mathbf{f}_1 = \mathbf{F}\mathbf{f} \tag{9}$$

where $\mathbf{f}$ is an $N \times 1$ column vector, $\mathbf{F}$ is the $N \times N$ DFT matrix and $\mathbf{f}_1$ is the DFT of $\mathbf{f}$. Similarly, the ath order discrete FrFT of $\mathbf{f}$, denoted as $\mathbf{f}_a$, is defined as:

$$\mathbf{f}_a = \mathbf{F}^a \mathbf{f} \tag{10}$$

where $\mathbf{F}^a$ is the $N \times N$ discrete FrFT matrix which corresponds to ath power of the classical DFT matrix $\mathbf{F}$. Note that there are certain subtleties and ambiguities in the definition of the power function [5].

The discrete FrFT is defined through a discrete analog of Eq. 8. Following the notation in [5], assuming $u_k[n]$ to be both the discrete Hermite–Gaussians and the orthonormal eigenvector set of the $N \times N$ $\mathbf{F}$ matrix and λ_k to be the associated eigenvalues, the discrete analog of Eq. 8 becomes:

$$\mathbf{F}^a[m, n] = \sum_{k=0}^{N-1} u_k[m] \lambda_k^a u_k[n] \tag{11}$$

This matrix is unitary since the eigenvalues, λ_k, of $\mathbf{F}$ have unit magnitude. Setting $a = 1$ reduces Eq. 11 to a spectral expansion of the classical DFT. Similarly, the index additivity property can be demonstrated by using the orthonormality of $u_k[n]$. As shown in [5] any definition that satisfies these three requirements, which are the first three properties given in Sec. 2, can be expressed in a spectral expansion form. Note that any arbitrary orthonormal eigenvector set of $\mathbf{F}$ satisfies these requirements. The fact that the eigenvector set of the DFT matrix is not unique is an ambiguity that should be resolved with Eq. 11. Note that $\mathbf{F}$ has only four distinct eigenvalues: $\{1, -1, j, -j\}$.

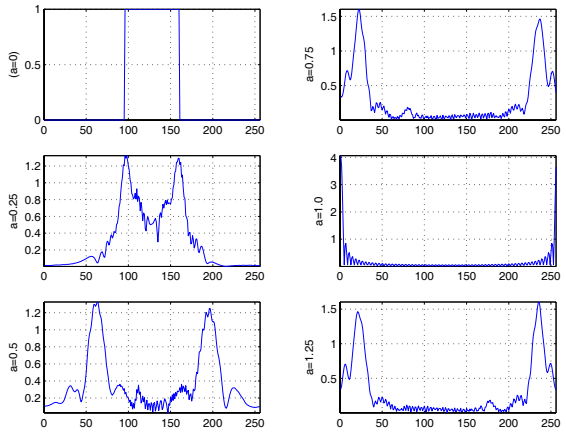

Figure 2: The magnitude of FrFT of rectangle function for various orders (domains).

The eigenvalues are in general degenerate. Therefore the eigenvector set is not unique. The same problem for the continuous case is resolved by choosing Hermite–Gaussian functions as the eigenfunctions of FT. Likewise, for the discrete case this ambiguity is resolved by choosing a common set of orthonormal eigenvectors for DFT and discrete Hermite–Gaussians.

The defining equation for the continuous Hermite–Gaussian functions is given below:

$$\frac{d^2 f(t)}{dt^2} - 4\pi^2 t^2 f(t) = \lambda f(t) \tag{12}$$

Likewise, the discrete Hermite–Gaussians are defined as the solutions of the corresponding difference equation. We refer the reader to [5] for the derivation of the following expression.

$$\mathbf{S} f[n] = \lambda f[n] \tag{13}$$

where $f[n]$ is the uniformly sampled values of $f(t)$ with a sampling interval of $1/\sqrt{N}$. Determining $\mathbf{S}$ is the main step in generating the $\mathbf{F}^a$ matrix. The remaining steps involve decomposing $\mathbf{S}$ into even and odd components and finding the eigenvalue and eigenvector set of each component. We refer the reader to [5] for a detailed explanation of these steps. Once $\mathbf{F}^a$ is determined the features used for speech recognition are computed using the following equation:

$$\mathrm{c}_a[i] = \sqrt{\frac{2}{M}} \sum_{j=1}^{M} \left\{ \mathrm{P}_{f^a}[j] \cos\left(\frac{\pi i}{M}(j-0.5)\right) \right\} \tag{14}$$

where c_a and P_{f^a} denote the cepstrum coefficients and the power spectrum of f in the ath fractional domain, respectively. The parameter M stands for the power spectrum dimension.

The example given in Fig. 2 demonstrates the relationship between time domain, frequency domain and some of the continuum of fractional domains in between. We evaluated the FrFT of the rectangle function for several orders. The first plot on the first column of Fig. 2 is the time domain function ($a = 0$). The second plot in the same column shows the magnitude of the transform in the $a = 0.25$ domain. The second plot in the second column is the regular DFT of the rectangular function which is a sinc function (after shifting the first half of the signal). Note that the last plot in the second column ($a = 1.25$) is the time–reversed copy of the first plot in the second column ($a = 0.75$) as stated by the properties given in Sec. 2. It is interesting to observe the evolution of the time domain signal into its frequency domain representation.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments are conducted on the Aurora 2 database. The database and the experimental configuration are described in [10]. The data consists of English digits spoken in the presence of additive noise and linear channel distortion. The distortions are artificially introduced to clean TI–digits data. There are two training conditions: clean–data and multi–style training, which uses noisy speech collected in various environments. The training data consists of 8440 utterances. The multi–style acoustic model training data consists of the same utterances artificially mixed with four different noise types at several signal–to–noise ratios (SNR). Three test sets consist of noise types similar to those seen in the training data (TSA), different from those seen in the training data (TSB), and with an additional convolutional channel distortion (TSC).

Speech is analyzed using 25 ms frames with a shift of 10 ms. Each frame is represented by a feature vector of 24 dimensions. MFCCs are computed from 24 Mel–filterbanks. Each frame is spliced with 4 proceeding and 4 succeeding frames (9 frames) and projected down to 39 dimensions using linear discriminant analysis (LDA). The range of this transformation is further diagonalized by using maximum likelihood linear transformation (MLLT) [7].

The acoustic model is trained as described in [11]. The model uses 22 phones, which are obtained from the words in the vocabulary. The system uses cross–word context dependent modeling with 198 leaves. The cross–word context dependency occurs only on the left of a word. Each phone is modeled by a 3–state left–to–right HMM topology. The output distributions for the 66 subphonetic units are modeled by a mixture of 3.2K Gaussians with diagonal covariance. The models are trained on the multi–style data. Given the small size of the vocabulary we performed unpruned Viterbi Decoding on a precompiled static HMM network obtained by expanding words into phones and leaves. The network contains 196 emitting and 78 null states.

The test data contains seven SNR conditions: {-5dB, 0dB, 5dB, 10dB, 15dB , 20DB, Clean}. The commonly used performance result includes the average word error rate (WER) for the conditions between 0dB and 20dB, which is shown as [0–20dB] in the third column of Table 1. Even though we chose Aurora 2 as the task for evaluations, we did not address issues related to noise robustness. The focus here is to introduce FrFT as an alternative frontend for a well defined compact experimental task. The FrFT, MFCC and LC are compared as frontends on the Aurora 2 database. Table 1 presents WERs averaged across different SNR conditions. In the table, [5dB–Clean] includes all conditions between 5dB and Clean: {5dB, 10dB, 15dB, 20dB, Clean}.

LDA objective function is used to determine the best fractional domain. We observed that the objective function is maximized in the neighborhood of $a = 1.0$. The fractional domain (order)

Performance Comparisons of the traditional and FrFT Based Features.(%)					
FrontEnd	Test	[0-20dB]	[5dB-Clean]	[10dB-Clean]	[15dB-Clean]
Linear Cepstra	TSA	11.54	3.78	1.65	1.13
No Adaptation	TSB	12.89	4.49	2.00	1.25
	TSC	11.97	4.56	2.36	1.47
MFCC	TSA	8.88	2.98	1.51	1.05
No Adaptation	TSB	9.70	3.29	1.62	1.08
	TSC	9.31	3.39	1.83	1.25
FrFT	TSA	8.96	2.92	1.44	1.02
No Adaptation	TSB	10.30	3.60	1.72	1.09
	TSC	9.22	3.34	1.78	1.20
Linear Cepstra	TSA	6.86	2.07	1.03	0.77
FMLLR Adaptation	TSB	7.86	2.42	1.13	0.78
	TSC	7.17	2.53	1.40	0.96
MFCC	TSA	6.30	1.94	0.98	0.73
FMLLR Adaptation	TSB	7.17	2.20	1.04	0.73
	TSC	6.46	2.31	1.29	0.90
FrFT	TSA	6.17	1.86	0.92	0.67
FMLLR Adaptation	TSB	7.21	2.23	1.01	0.65
	TSC	6.18	2.03	1.09	0.75

Table 1: Word Error Rates (WER) across different test conditions on the Aurora 2 database.

is determined empirically in the neighborhood of $a = 1.0$. We found $a = 1.025$ (equivalently $a = 0.975$) to be the best value for the minimum WER. Note that $a = 0.975$ results in the same transform coefficients but reversed in time. In Fig. 2, the transform outputs for the cases of $a = 0.75$ and $a = 1.25$ are identical, except for the time reversal of the axis.

In Table 1, we see that without adaptation MFCC features present the lowest WER for TSA and TSB across all SNR conditions. This is to be expected as MFCC includes a form of smoothing when filterbank energies are computed as opposed to using raw transform coefficients as in the cases of LC and FrFT. Nevertheless, FrFT performed slightly better than MFCC at high SNR conditions on TSC. We also performed feature space maximum likelihood linear regression (FMLLR) adaptation. The goal of this transform is to affinely transform the adaptation data so as to maximize their likelihood [6]. Performing FMLLR adaptation results in more improvement for LC and FrFT than MFCC. However, the improvement from adaptation is not enough for LC to match the performance of MFCC. On the other hand, FrFT matches the performance of MFCC for TSB and outperforms it for TSA and TSC. We observe that the FrFT based features obtained larger improvements as average SNR level is increasing.

We also notice the largest relative decrease in WER for TSC, which may suggest that FrFT is more robust to linear channel distortion than MFCC. The relative improvements in WER on TSC across different SNR conditions listed in the table are 4.3%, 12.1%, 15.5%, and 16.6%, respectively. The corresponding figures for TSA are moderate 2.1%, 4.1%, 6.1% and 11.9%. FrFT outperformed LC by about relative 10%, 9% and 19% on average for TSA, TSB and TSA, respectively across all SNR conditions.

5. CONCLUSIONS AND FUTURE WORK

We introduced a new speech analysis method for speech processing in general and speech recognition in particular. The proposed method is called fractional Fourier transform (FrFT), which generalizes classical Fourier transform. Even though we applied FrFT to extract features for speech recognition, we believe it will find other applications in such areas as enhancement, synthesis and others. The experiments conducted on the Aurora 2 database showed significant improvement compared to such traditional frontends as MFCC and linear cepstra in high SNR conditions. Our future research direction will focus on compensating different sources of speech variability in different fractional domains and joint modeling of speech in multitude of fractional domains.

Acknowledgment

The authors thank Michael Picheny for fruitful discussions.

References

[1] L.B. Almeida, "The fractional Fourier transform and time–frequency representation", *IEEE Trans. on Speech and Audio Proces.*, vol. 42, pp. 3084-3091, Nov. 1994.

[2] H. M. Ozaktas, M. A. Kutay and D. Mendlovic, *The fractional Fourier transform with applications in Optics and Signal Processing*, John Wiley & Sons, New York 2001.

[3] S. B. Davis and P. Mermelstein, "Comparisons of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on Speech. Signal Proces.*, vol.ASSP–28, pp. 357–366, 1980.

[4] H. M. Ozaktas, B. Barshan, D. Mendlovic, and L. Onural, "Convolution, filtering, and multiplexing in fractional Fourier domains and their relation to chirp and wavelet transform", *J. Opt. Soc. Amer. A.*, vol. 11, pp. 547–559, 1994.

[5] C. Candan, M. A. Kutay and H. M. Ozaktas, "The Discrete Fractional Fourier Transform", *IEEE Trans. on Speech and Audio Proces.*, vol. 48, pp. 1329–1337, May 2000.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM–based speech recognition", *Technical Report CUED/F–INFENG* Cambridge Univ. Eng. Dept., 1997.

[7] M. J. F. Gales, "Semi–tied covariance matrices for hidden Markov Models", *IEEE Trans. on Speech and Audio Proces.* vol. 7, pp. 272–281, 1999.

[8] V. Namias, "The fractional order Fourier transform and its application to quantum mechanics", *J. Inst. Math. Appl.*, vol. 25, pp. 241–265 , 1980.

[9] B. Ayrulu and B. Barshan, "Fractional Fourier transform pre–processing for neural networks and its application to object recognition", *Neural Networks*, vol. 15, pp. 131–140 , 2002.

[10] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *ICSLP–2000*, Beijing, China, vol. 4, pp. 29–32, October 2000.

[11] G. Saon and H. Huerta, "Improvements to the IBM Aurora 2 Multi–Condition System", *ICSLP–2002*, Denver, CO, Sept. 2002.