USING PITCH AND FORMANTS FOR ORDER ADAPTATION OF FRACTIONAL FOURIER TRANSFORM IN SPEECH SIGNAL PROCESSING

Hui Yin[1,2], *Climent Nadeu*[1], *Volker Hohmann*[1,3]

1. TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
2. Dept. of Electronic Engineering, Beijing Institute of Technology, Beijing 100081, China
3. Medical Physics, Universität Oldenburg, Germany

ABSTRACT

Fractional Fourier transform (FrFT) has been proposed to improve the time-frequency resolution in signal analysis and processing. However, selecting the FrFT transform order for the proper analysis of multi-component signals like speech is still debated. In this work, we investigated several order adaptation methods based on the pitch and formants of voiced speech. This study is motivated by the fact that speech is not stationary even in a short time interval, and the idea is shown using an AM-FM speech model. First, FFT and FrFT based spectrograms of an artificially-generated vowel are compared to indicate the merit of the methods. Second, a tonal vowel discrimination test is designed to compare the performances of the various proposed methods using MFCC features implemented with FrFT.

1. INTRODUCTION

Speech is a non-stationary signal. Traditional speech processing methods generally treat speech as short-time stationary, i.e., process speech in 20~30ms frames. In practice, intonation and coarticulation introduce combined spectro-temporal fluctuations to speech even for the typical frame sizes used in the front-end analysis. Modeling speech signals as frequency modulation signals accords better with speech characteristics from both production and perception views.

From the speech production view, traditional linear source-filter theory lacks the ability to explain the refined structure of speech in a pitch period. Maragos et al. therefore proposed an AM-FM modulation model for speech analysis, synthesis and coding [1]. From the perception view, neurophysiological studies show that the auditory system of mammals is sensitive to FM-modulated (chirpy) sounds. This fact explains the human sensitivity to non-stationary acoustic events with changing pitch (police and ambulance siren) [2].

Fractional Fourier transform (FrFT) can be considered as a generalization of the traditional Fourier transform [3]. Since FrFT can be considered as a decomposition of the signal in terms of chirps, FrFT is especially suitable for the processing of chirp-like signals [4]. The chirp rate (temporal derivative of instantaneous frequency) of the FrFT kernel functions is set by one free parameter, the transform order. The determination of the optimal transform orders is always critical. In this paper we show that the representation of the time-varying properties of speech may benefit from using the values of pitch and formants to set the order of the FrFT. Different order adaptation methods based on pitch and formants are proposed in this paper.

In tonal languages as Mandarin, the time evolution of pitch inside a syllable (the tone) is relevant for the meaning. Consequently, there are relatively fast changes of pitch which are usual and informative. As the use of the FrFT might help to better track the dynamic properties of speech harmonics, we have carried out a classification experiment using a small set of Mandarin vowels, where the classes correspond to the four basic types of tones, and the discrimination ability is measured using the MFCC features implemented with FrFT.

The rest of the paper is organized as follows. In section 2, the AM-FM model of speech is described, and the motivation of the proposed method is given. In section 3, the definition and some basic properties of FrFT are briefly introduced. In section 4, different order adaptation methods are described. One method is illustrated using FFT and FrFT based spectrograms of an artificially-generated vowel. In Section 5, a tonal vowel discrimination test is designed, and the results are given and analyzed. Some conclusions and future work are given in section 6.

2. THE AM-FM MODEL OF SPEECH

Considering the fluctuation of pitch and the harmonic structure, voiced speech can be modeled as an AM-FM signal

$$x(t) = \sum_{n=1}^{\infty} a_n(t) \cos(n(\omega_0 t + \int_0^t q(\tau) d\tau) + \theta_n) \qquad (1)$$

This research was partially supported by the Spanish project SAPIRE (TEC2007-65470), and a research grant to V.H. from the Spanish Ministry of Education and Science. H.Y. was partially supported by the National Nature Science Foundation of China under Grant NSFC 60605015.

where $q(t)$ is the frequency modulation function. Making the reasonable simplification that the frequency is changing linearly within the frame, i.e. $q(t)=kt$, where *k* is the chirp rate of the pitch (referred to as *pitch rate* in the rest of the paper), we can obtain:

$$x(t)=\sum_{n=1}^{\infty} a_n(t)\cos(\underbrace{n(\omega_0 t+\frac{1}{2}kt^2)+\theta_n}_{\varphi_n(t)}) \qquad (2)$$

The chirp rate of the *n*-th harmonic is the second derivative of the phase function

$$\frac{d^2\varphi_n(t)}{dt^2}=q_n=nk, \qquad (3)$$

which means that the chirp rate of the *n*-th harmonic is *n* times the pitch rate.

3. DEFINITION OF THE FRACTIONAL FOURIER TRANSFORM

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

$$X_\alpha(u)=F_p[x(t)]=\int_{-\infty}^{\infty} x(t)K_\alpha(t,u)dt, \qquad (4)$$

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

$$K_\alpha(t,u)=\begin{cases}\sqrt{\frac{1-j\cot\alpha}{2\pi}}\exp\left(j\frac{t^2+u^2}{2}\cot\alpha - jut\csc\alpha\right), & \\ & \alpha\neq n\pi \\ \delta(t\text{-}u), & \alpha=2n\pi \\ \delta(t+u), & \alpha=(2n\pm1)\pi\end{cases} \qquad (5)$$

The inverse FrFT is

$$x(t)=F_{-p}[X_\alpha(u)]=\int_{-\infty}^{\infty} X_\alpha(u)K_{-\alpha}(t,u)du \qquad (6)$$

Eq.(6) indicates that the signal $x(t)$ can be interpreted as a decomposition to a basis formed by the orthonormal Linear Frequency Modulated (LFM) functions in the *u* domain, which means an LFM signal with a chirp rate corresponding to the transform order *p* can be transformed into an impulse in a certain fractional domain. Therefore, the FrFT has excellent localization performance for LFM signals.

4. ORDER SELECTION METHODS

To test the proposed order selection methods informally, we produced an artificial vowel [i:] with time-varying pitch. The excitation of the vowel is a pulse train with linearly decreasing frequency from 450Hz to 100Hz, and the formants of the vowel are 384Hz, 2800Hz, and 3440Hz, which are extracted from a real female vowel. The sampling rate is 8000Hz.

We experimented three different classes of order adaptation methods based on the pitch and formants. They will be explained in detail and the spectrograms of the artificial vowel based on these methods are shown.

4.1. N times of pitch rate

Since the chirp rates for different harmonics are different, the FrFT is emphasizing the N-th harmonic when setting the transform order according to N times of the pitch rate *k*. The transform angle is determined by:

$$\alpha=\text{acot}(-2\pi*k*N). \qquad (7)$$

When the order is set according to N times of the pitch rate, the N-th harmonic and its neighbors will be emphasized, i.e. they have better concentration performance than the FFT-based spectrogram. On the other hand, the representation of harmonics whose chirp rates are not close to 10 times of pitch rates will be smeared. This is also true for the formants, because their frequency variations are generally smaller than the harmonics, i.e., the chirp rates of the formants are generally much smaller than N times of pitch rate when N gets larger.

4.2. Pitch and formants

The sub-band energies that are usually employed to compute the speech recognition features, e.g. in the widely used MFCC, are a representation of the envelope Since the FT-based spectral harmonics are an intermediate step in the computation of the envelope, a more precise representation of the harmonics in relevant regions of the spectral envelope may help to get more accurate formant estimates and also more discriminative speech features. This is the motivation for the order adaptation method based on pitch and formants that is introduced in the following. As in (11), the transform angle is determined by M times of the pitch rate *k*:

$$\alpha=\text{acot}(-2\pi*k*M). \qquad (8)$$

M will be computed from the frequency of a formant and the pitch frequency as

$$M=f_{formant}/f_{pitch}. \qquad (9)$$

Here, M is different for different analysis frames.

4.3. Multi-order multiplication

Since different optimal orders are needed for different harmonics, we can calculate the FrFT with the orders corresponding to 1, 2, 3… times of the pitch rate and multiply them together. This method can obtain a compromise among several harmonics. Alternatively, in our experiments, multi-order multiplication was also applied to the N FrFT spectrograms that target the first N formants according to the technique described in section 4.2. The resulting multiplied FrFT spectrogram is shown in figure 1 for N=3 (right panel). In this case, formant smearing is limited, while still enhancing the harmonics going through the formant resonances.

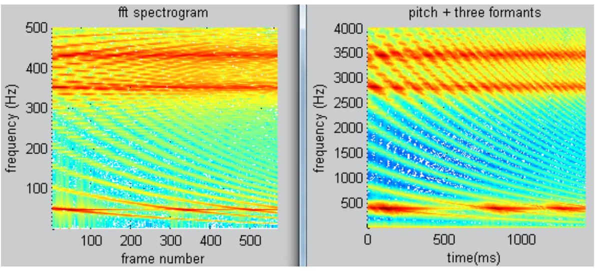

Figure 1: Left panel: *FFT-based spectrogram. Right panel: FrFT-based spectrogram with multi-order multiplication. The order multipliers M1,M2 and M3 (see eq. 9) correspond to the three formant frequencies.*

5. TONAL VOWEL DISCRIMINATION TEST

In Mandarin, there are four basic lexical tones and a neutral tone. The number of tonal syllables is about 1300, and it is reduced to about 410 when tone discriminations are discarded [5]. Fundamental frequency or pitch is the major acoustic feature to distinguish the four basic tones. In order to test the performance of the order adaptation methods, we designed a tonal vowel discrimination test. Since the proposed FrFT order adaptation methods may show a more accurate representation of the time-varying characteristics of the harmonics than the FT, we decided to test them in tone recognition for tonal languages.

5.1. Experiment design

We recorded the five Mandarin vowels [a], [i](yi), [u](wu), [e], [o](wo) with four tones: the flat tone (tone 1), the rising tone (tone 2), the falling and rising tone (tone 3), and the falling tone (tone 4). Each tone of each vowel from a female voice is recorded five times. The utterances are sampled at 8kHz, with a 16bit quantization. We use 16-dimensional standard MFCC features as the baseline. The features based on the FrFT are computed with the same processing used for the MFCCs, but substituting the Fourier transform by the FrFT (we will refer to them as FrFT-MFCC) [6].The performance of FrFT-MFCC using different order adaptation methods is compared with the baseline. Speech signals are analyzed using a frame length of 25ms and a frame shift of 10ms.

Because the recorded utterances have variable lengths, we use Dynamic Time Warping (DTW) to calculate the distances between all the utterances for the individual vowels. Thus, five 20x20 distance matrices are obtained (4 tones, 5 times). The discriminative ability of features can be analyzed using the Fisher score, which is defined as the ratio between the between-class variance and the within-class variance. Here, we take the distances calculated by DTW to compute a similar score (also referred to as Fisher Score):

$$F = \frac{\frac{1}{N_1}\sum_{m=1}^{5}\sum_{n=1}^{5}\sum_{i=1}^{4}\sum_{j\neq i,j=1}^{4} dist(v_i^m, v_j^n)}{\frac{1}{N_2}\sum_{m=1}^{5}\sum_{n=1}^{5}\sum_{i=1}^{4} dist(v_i^m, v_i^n)} \qquad (10)$$

v_i^m represents the token m of a vowel with tone i. N_1 and N_2 are the total numbers of the between-class and within-class tokens respectively. $dist(\bullet)$ represents the Euclidean Distance. By this analysis, the discriminability across different tones of the same vowel is assessed. The discrimination among different vowels is also assessed here for comparison.

5.2. Pitch rate and formant calculation

The speech is processed in overlapping frames. Each frame is further divided into several non-overlapping sub-frames. One pitch value is detected for one sub-frame. These pitch values are obtained using a robust pitch tracking algorithm described in [7]. In order to get the pitch rate of a frame, we first calculate the median value of the sub-frame pitch values for this frame to set a threshold, if any sub-frame pitch value is larger than twice this threshold, then it is divided by 2. If any pitch value is smaller than half the threshold, it is multiplied by 2. By this, octave confusions are largely eliminated. Then, a straight line was fitted to all the corrected pitch values in this frame. The pitch rate is taken as the slope of this fitted line.

The formants are determined as the frequencies of the LPC-based spectral peaks. The order for LPC analysis is set to be twice the number of formants used in the multi-order FrFT analysis. Note that when the number of formants exceeds 4, they may be not real formants but envelop peaks.

5.3. Experimental results

The Fisher scores for different vowels using the various methods are given in Table 1. In this experiment, the frame length is 25ms and frame shift is 10ms. Every frame is divided into 5 subframes. The experimental results show that FrFT analysis increases the tone discriminability for most of the order selection methods proposed here. We can see that:

(1) The average Fisher score over all vowels using MFCC is 4.43. This indicates that MFCC already has a good discriminability for different tones, but the FrFT-MFCC can get even better results, especially for the multi-order multiplication method with N=1*2*..*5, which obtains nearly 50% improvement. For comparison, the Fisher score for the discrimination of different vowels of the same tone is 12.20 on average across tones. This indicates that the discrimination of tones is more difficult than the discrimination of vowels,

as expected, and that the improvement of tone-discrimination by using the FrFT might provide a large benefit for speech analysis and recognition applications.

(2) When using a single N value for the N times of pitch rate method, the increases of the scores are moderate. Just as stated before, the formants may be dispersed when N gets larger, because the chirp rate of formants is not close to that value. There is always an optimal value of N. Generally N=1~3 can obtain a good compromise between tracking the dynamic speech harmonics and preserving the concentration of the formants.

	a	i	e	o	u	Average
MFCC	2.77	3.94	5.28	4.59	5.56	4.43
N=1	2.63	4.48	6.24	4.90	6.61	4.97
N=2	2.58	4.15	6.07	4.78	6.48	4.81
N=3	2.55	3.95	5.90	4.68	6.38	4.69
N=5	2.49	3.71	5.61	4.52	6.19	4.5
Pitch +MP	2.46	4.76	6	4.77	6.55	4.91
Pitch +2MP	2.25	3.91	5.53	6.94	8.74	5.47
Pitch +3MP	2.27	4.53	5.67	6.23	11.2	5.99
Pitch +5MP	2.44	4.52	5.85	6.00	12.0	6.16
Pitch+ 10MP	2.11	4.21	6.85	4.13	12.7	5.99
N=1*2	2.4	4.63	5.67	6.91	9	5.72
N=1*2*3	2.36	5.41	5.71	6.17	11.6	6.25
N=1*2*..*5	2.46	5.01	5.86	5.96	12.4	6.34
N=1*2*..*10	2.13	4.08	6.83	4.1	12.5	5.93

Table 1: *Fisher scores using MFCC and all variants of the FrFT-MFCC method. MP denotes the main peaks of the LPC spectrum, and Pitch + xMP refers to the technique presented in Section 4.2. When x>1, the transforms are multiplied as explained in Section 4.3 (right panel in Fig. 1).*

(3) The pitch + "formants" method can obtain significantly better results than the method only based on the pitch. Different vowels have their different optimal numbers of formants, e.g. for [u], even using 10 formants its maximum is still not achieved, but for [i], the maximum is achieved using one main formant, and for [o], two formants. The pitch + 5MP method can obtain good results on average for all vowels except [a].

(4) For the vowel [a], the FrFT-MFCC always performs worse than MFCC. This is possibly because the first formant of [a] is much higher than in the other vowels. A higher formant needs a larger N, but a larger N will smear the formant, so a good compromise can't be achieved.

(5) The multi-order multiplication method with different number of N's can significantly increase the scores for vowels [i] [e], [o] and [u] compared with MFCC. These four vowels achieve their best results with different numbers of order multipliers. Here, they are 3, 10, 1, 10 respectively. The best average result of all is obtained using the multi-order multiplication method with N=1*2*..*5.

(6) Compared with the pitch + MP method, the pitch + 2MP method improves the discriminability of FrFT-MFCC for vowels [o], [u], but not for the other three vowels, especially for [i]. The reason for this might be the frequencies of the first two formants of [o] and [u] are low and close, so a significant improvement can be obtained; but it's the opposite for [i], whose first formant is quite low and the second formant is rather high. The smearing effect prevails in the combination of the corresponding two orders. When more "formants" are taken, such situation is somewhat alleviated.

6. CONCLUSIONS AND FURTHER WORK

In this paper, we have proposed several order adaptation methods for FrFT in speech signal analysis and processing, which are based on the pitch and the formants (or just envelope peaks) of voiced speech. The FrFT results with the proposed order selection methods have some improvement over its FFT counterpart. We have also done some preliminary work applying the N times of pitch rate method to speech recognition, and the results show some improvement over the MFCC baseline [8]. 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.

10. REFERENCES

[1] P. Maragos, T. Quatieri, and J. F. Kaiser, “On Amplitude and Frequency Demodulation Using Energy Operators,” *IEEE Transaction on Acoustics, Speech and Signal Processing*, *41*(4), pp. 1532-1550, April 1993.

[2] M. Képesi, L. Weruaga, “High-resolution noise-robust spectral-based pitch estimation”, *Interspeech*, Lisbon, Portugal, 313-316, 2005.

[3] Namias V. The fractional order Fourier transform and its application to quantum mechanics. *J Inst Math Appl*, *25*, 1980, 241-265.

[4] 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

[5] Y.-R. Chao, ed., *A Grammar of Spoken Chinese*, University of California Press, Berkeley, 1968.

[6] Yin Hui, Xie Xiang, Kuang Jingming, "Adaptive-Order Fractional Fourier Transform Features for Speech Recognition", *Interspeech*, Brisbane, Australia, 2008

[7] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)" in "Speech Coding & Synthesis", W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.

[8] Hui Yin, Climent Nadeu, V. Hohmann, et al. "Order adaptation of the fractional Fourier transform using the intraframe pitch change rate for speech recognition". ISCSLP, Kunming, China, 2008.