MONAURAL SPEECH SEPARATION BASED ON MULTI-SCALE FAN-CHIRP TRANSFORM

Pei Zhao, Zhiping Zhang, Xihong Wu

Speech and Hearing Research Center
State Key Laboratory of Machine Perception, Peking University
Beijing, 100871, China
{zhaopei,zhangzp,wxh}@cis.pku.edu.cn

ABSTRACT

A novel method for monaural speech separation is presented in this paper. Instead of the traditional Short Time Fourier Transform (STFT) for time-frequency analysis in speech separation, the Fan-Chirp Transform (FChT) has been applied to track the pitch and harmonics of the target speech. This method has two advantages over STFT. Firstly, the spectrum spread of dynamic harmonics within each analysis frame has been alleviated. Secondly, the FChT bases with proper chirp rate could be chosen according to different frequency modulation rates in the simultaneous speech. Furthermore, considering the changeability of frequency modulation rates, a multi-scale FChT is proposed to adaptively adjust the frame length of spectrum analysis. Experimental results prove the validity of the approach in monaural speech separation.

Index Terms— Speech analysis, Harmonic analysis, Frequency modulation, Chirp, Monaural speech separation

1. INTRODUCTION

In real world, sound is usually mixed by several sources. Separating the speech from other interference sources in monaural recordings is a challenging and meaningful work. Researchers have attacked this difficult problem in various research aspects, such as statistical machine learning [1], speech representation [2, 3], auditory scene analysis [4, 5], etc. Among these studies, speech representation attracts much attention.

The sinusoidal model [2, 6] is one of the most important tools for speech representation in speech separation [3]. It represents a speech signal as a linear combination of sinusoids with time varying amplitudes, frequencies, and phases for harmonic analysis. In voiced speech, the speech signal is represented by the sum of a finite number of corresponding sinusoidal parameters at the fundamental frequency and its harmonics. Recent implementations of the separation system with the sinusoidal model generally based on STFT [3]. Although the STFT is suitable for signals with fixed frequency components in an undertaking frame, for real speech, the pitch is time-variant, and the STFT may result in the spectrum spread of the harmonics. This problem is even more serious in high frequency harmonics, since the frequency modulation rates of high frequency harmonics are faster than the low ones.

In 2004, L.Weruaga and M.Képesi used the fast chirp transform for speech analysis [7], without the assumption that the pitch is constant in an analyzed frame. And then in 2006, FChT is proposed for spectrum analysis and is used to extract the pitch effectively both in clean and noisy speech [8]. In FChT, the bases are consisted of a set of comodulated sinusoids. When the modulation rate of the chirp basis is matched to that of the analyzed harmonics, FChT could get finer harmonic structure, and the spectrum spread of harmonics could be reduced significantly.

In the paper, we attempt to use FChT in speech separation for two advantages. Firstly, the spectrum spread of dynamic harmonics within each analysis frame has been alleviated in mixture speech. Secondly, the FChT bases with proper chirp rate could be chosen according to different Frequency Modulation (FM) rates in the simultaneous speech.

Furthermore, considering the changeability of frequency modulation rates, a multi-scale FChT is proposed to adaptively adjust the frame length of spectrum analysis in the paper. For the harmonics of speech with stable FM rates, enlarging the time scale of spectrum analysis so as to increase the frequency resolution leads to the efficiency of the separation of target speech spectrum in FChT analysis. On the contrary, in case that the FM rate is unstable, a smaller time scale is probably preferred to track subtle variation.

The rest of this paper is organized as follows. In Section 2 the overview of the proposed separation method is presented, and then related algorithms are explained in details. Section 3 shows the experimental results. Then the discussions and conclusions are followed in Section 4.

2. PROPOSED SPEECH SEPARATION METHOD AND RELATED ALGORITHMS

The proposed method adopts the multi-scale FChT to decompose the mixture speech into Chirp bases for feature extraction, and tracks the pitches one by one using a dynamic programming (DP) algorithm with a comb filter. Then after inte-

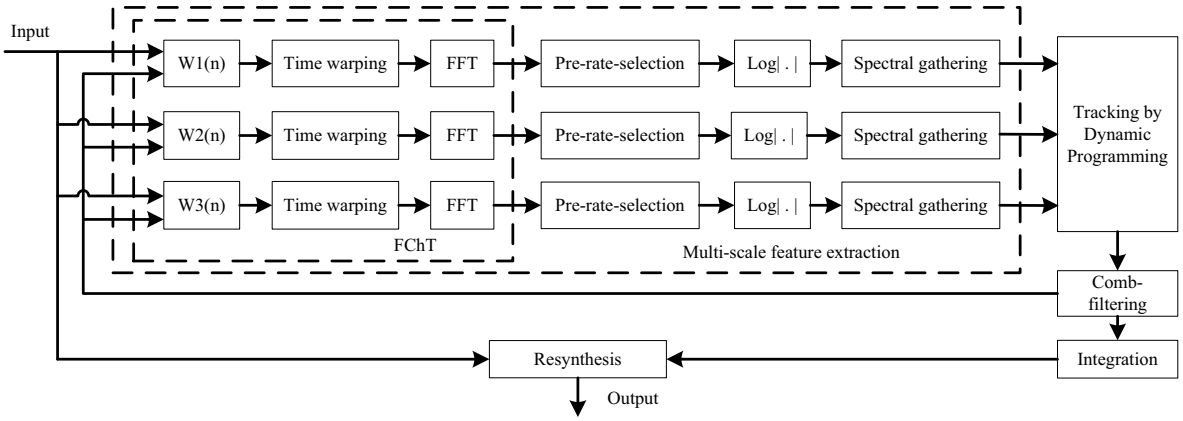

Fig. 1. *Overview of the proposed speech separation system based on multi-scale FChT.*

gration, the speech is resynthesized. The system overview is shown as figure 1, and the details are as follows.

2.1. FChT for mixture speech analysis

In the mixture speech, there are two important cues for speech separation. One is the pitch values of the different speakers in each analysis frame; the other is the FM rates of the simultaneous pitches. In the paper, we use FChT for mixture speech analysis to obtain these cues. The system decomposes the mixture speech into frames and the spectrum of chirp transform is obtained by FChT [7, 8], which is composed of an adaptive analysis basis of quadratic chirps, and could be implemented by time warping and FFT, shown in figure 1.

2.2. Multi-scale FChT and feature extraction

In the paper, multi-length analysized frames are adapted to various requirements of the time frequency resolution of mixture speech representation. The system decomposes the mixture speech into multi-length frames by the window functions denoted as W1(n), W2(n) and W3(n) in figure 1 in which the three levels (l = 1,2,3) of the frame lengths N_l contains 512, 768 and 1024 samples per frame for the l^{th} level. Then the spectrum of chirp transform is obtained by multi-scale FChT. The time warping process is shown as follow equations,

$$\mathbf{x}(n) \xrightarrow{\varphi_{\alpha,l}(.)} y_{\alpha,l}(n) = x(f_s \varphi_{\alpha,l}(n/f_s)) \quad (1)$$

$$\phi_{\alpha,l}(n) = (1 + \omega_\alpha(n - N_l))n \quad (2)$$

$$\phi_{\alpha,l}(\varphi_{\alpha,l}(n)) = n \quad (3)$$

where $\mathbf{x}(n)$ is an N_l samples length frame of mixture speech signal sampled at a Nyquist rate f_s , $\phi_{\alpha,l}(n)$ is the warping function with chirp rate ω_α with warping rate index α ($\alpha = 0, 1, \ldots A$), and $\varphi_{\alpha,l}$ is the inverse function of $\phi_{\alpha,l}$. Here l denotes the level index. The warped signal $y_{\alpha,l}(n)$ on each level l is then followed by a N_{max} point FFT as equation (4)

$$X_l(\alpha, k) = FFT_{N_{max}}\{y_{\alpha,l}(n)\} \quad (4)$$

where $N_{max} = 8192$ by adding zeros at the end of $y_{\alpha,l}(n)$. k is the frequency index, $k = 0, 1, \ldots N_{max} - 1$.

In one frame, the bases of FChT with the proper chirp rate, which match the FM rate of the pitch of the dominate speech, always make the spectra tighter than the other chirp rate bases when the dominate harmonics' energy is high enough. To save the computational cost, we select the chirp rates with the first M largest values of the normalized spectral square sum as the candidates, and in this paper, we select the first two ($M = 2$). This process is denoted as Pre-rate-selection module in figure 1 and is calculated by

$$S_l(\alpha) = \sum_{k=0}^{N_{max}/2} |X_l(\alpha, k)|^2 / \overline{X_l} \quad (5)$$

$$\overline{X_l} = \sum_{\alpha=0}^{A} \sum_{k=0}^{N_{max}/2} |X_l(\alpha, k)|^2 / N_l \quad (6)$$

where $S_l(\alpha)$ is the normalized spectral square sum. $\overline{X_l}$ is the sum of all rates on each level, and is normalized by the frame length N_l. It is used to normalize the feature value $S_l(\alpha)$ of various chirp rates and frame length levels in equation (5).

Then the logarithmic energy chirp spectra $SLog_l(\alpha, k)$ with the selected candidate with chirp rate index α on level l is given as equation (7).

$$SLog_l(\alpha, k) = log_{10}(|X_l(\alpha, k)|^2) \quad (7)$$

The gathered spectra of the pitch candidate, which is used as the feature for pitch tracking, shown in figure 1 as the Spectral gathering module, is calculated with equation (8) in [8]

$$\rho_l(\alpha, p) = \frac{1}{H} \sum_{h=1}^{H} SLog_l(\alpha, hp) \quad (8)$$

where H is the number of harmonics within the Nyquist bandwidth by assuming the candidate fundamental frequency index to be p, and h is the harmonic index. The values of the gathered spectra $\rho\{\Omega\} = \rho_l(\alpha, p)$ are the features with the

parameters $\Omega = \{\alpha, p, l\}$ of each level l , selected candidate chirp rate index α, and candidate pitch p. For the i^{th} frame, the parameters are denoted as $\Omega_i = \{\alpha_i, p_i, l_i\}$. To solve the double pitch errors, a weighted term $k(p)$ is used, which is decreased with the frequency p, and to solve the half pitch errors, delta value is used as the cost as equation (9)

$$c(\Omega) = k(p) \cdot (\rho_l(\alpha, p) - \rho_l(\alpha, p/2)) \tag{9}$$

2.3. Pitch tracking

After calculated the cost value, the system tracks the pitch in all levels by a DP algorithm. The process shown in figure 1 as Tracking module, is also illustrated in figure 2. The chirp rate ω_{α_i} of the i^{th} frame is given by

$$\omega_{\alpha_i} = \frac{(p_i - p_{i-1}) f_s}{p_i + p_{i+1}} \tag{10}$$

where p_{i-1} , p_i and p_{i+1} are the candidate pitches of frame $i-1$, i and $i+1$. The pitch candidates vary from 60 Hz to 400 Hz with equally spaced step in logarithmic domain.

The score $s_i(\Omega_i)$ of the i^{th} frame for each path of candidate pitch p_i, and FM rate index α_i obtained by equation (10) on level l_i is shown as equation (11) and (12)

$$s_i(\Omega_i) = s_{i-1}(\Omega^*_{i-1}) + c(\Omega_i) \tag{11}$$

$$\Omega^*_{i-1} = \arg\max_{\Omega_{i-1}}(s_{i-1}(\Omega_{i-1}) + c(\Omega_i)). \tag{12}$$

Then for each frame, the system selects the optimal Ω^* in the whole parameter space with the parameter Ω of pitch value and chip rate in all three levels for the largest score of the path. Then the system could get the optimal Ω^* sequence for all the frames of the utterance.

After tracking one continuous pitch contour in a segment, the system filters the harmonics belonging to the group of this estimated pitch contour using comb-filtering, and then tracks the continuous pitch contour of another source by the same method.

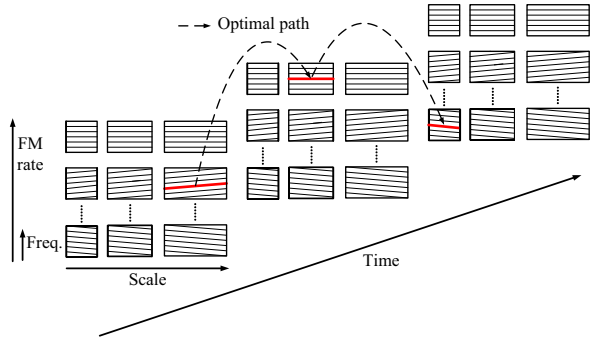

Fig. 2. *Illustration of multi-scale FChT based pitch tracking.*

2.4. Integration

For the segments which have no overlap in the time domain, the DP algorithm cannot get a solution for speech integration. Then the segments are integrated by the mean values of the pitch values obtained forward using a k-means algorithm, and the non-overlap segments are clustered into two source groups. However, this method is not suitable for the case of pitch ranges of the simultaneous talkers has much overlap.

2.5. Resynthesis

In this paper, the separated speech is resynthesized by FChT based sinusoidal model. The speech is composed by the connection of the frames with frame length N and no overlap. The i^{th} frame signal $s_i(n)$, $n = 1, \ldots N$ is shown as

$$s_i(n) = \sum_{h=0}^{H} c_i(h, n) \cos(\theta_i(h, n)) \tag{13}$$

$$c_i(h, n) = c_{i-1}(h, N) + n f_s (c_i(h, N) - c_{i-1}(h, N))/N \tag{14}$$

$$\theta_i(h, n) = \theta_{i-1}(h, N) + 2\pi n p_{i-1}(1 + n\omega_{\alpha_i}) \tag{15}$$

where $c_i(h, n)$ and $\theta_i(h, n)$ are the amplitude and phase of the h^{th} harmonic respectively at sample n of i^{th} frame, which are calculated as equation (14) and (15). The initial amplitude $c_{i-1}(h, N)$, final amplitudes $c_i(h, N)$, and the warping rate ω_{α_i} of the current i^{th} frame could be obtained according to the optimal Ω^* sequence and the FChT spectra. Both the phase and frequency of the harmonics are continuous over the conjoint frames in the model.

3. EXPERIMENTS AND RESULTS

In the experiments, a female utterance (n7) and a male utterance (n8) from the database of Cooke 100 utterances [9] are selected as test data, which are mixed by signal-to-noise ratios (SNRs) ranging from -6 dB to 6 dB in 2 dB increments. Both utterances in the mixtures were separated as the target speech, and the mixture SNR is defined as n7 to n8 ratio in the experiments. The mixture speech signals are separated by the method described above, and the source-to-interferences ratio (SIR) is employed to evaluate the efficiency of separation. It performs the measure described in [10] , which decomposes a given estimate $\hat{s}(n)$ of a source $s(n)$ as a sum shown as

$$\hat{s}(n) = s_{clean}(n) + e_{interf}(n) + e_{artif}(n) \tag{16}$$

where $s_{clean}(n)$ is the component of clean speech, $e_{interf}(n)$ is the component of interference, and $e_{artif}(n)$ is the artificial component. In [10], SIR is given by

$$SIR = 10log_{10} \frac{\| s_{clean} \|^2}{\| e_{interf} \|^2}. \tag{17}$$

In figure 3 (a-e), the spectrum from a frame of 0dB mixture speech is shown, and it contrasts the spectra from the

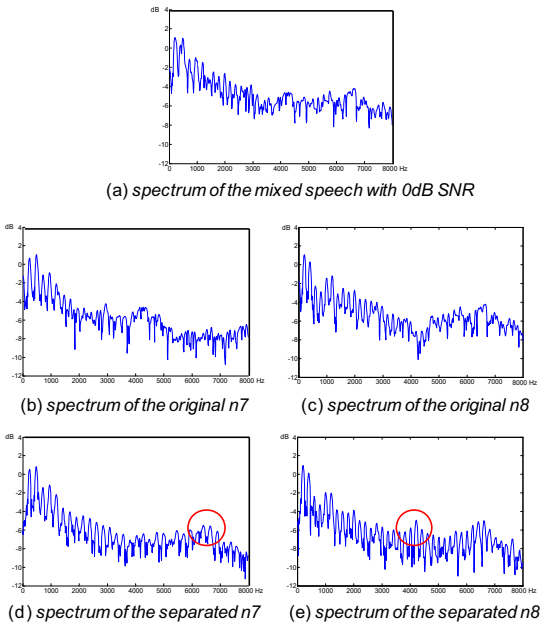

Fig. 3. *The spectra of mixture speech (a), original n7 (b), original n8 (c), separated n7 (d) and separated n8 (e) by the proposed system.*

separated n7 and n8 with those from the original utterances. In the gross, the separated spectra are similar to the original ones, especially in the low frequency. In some high frequency bands (denoted by circles), envelops of the separated spectra are interfered by the other speech source where the local SNR is lower.

The mean SIRs of the two separated sources by the STFT based and FChT based methods are shown in table 1, including 512 frame length STFT, 512, 768, 1024 frame length and multi-scale FChT. The frame shifts in each case are half of the frame lengths. The results of FChT based are better than the STFT based with the same frame length (512) by 0.98dB

Table 1. *Mean SIR (dB) results of the systems based on STFT with 512 samples per frame and FChT with variable frame lengths (512, 768, 1024 samples per frame and multi-scale).*

	STFT	FChT			
SNR	512	512	768	1024	multi-scale
-6dB	12.41	12.60	15.52	14.86	17.06
-4dB	12.00	13.34	14.40	15.07	17.09
-2dB	11.42	11.42	15.45	15.11	17.33
0dB	11.47	13.35	12.70	15.44	14.57
2dB	10.95	14.19	12.62	15.34	16.12
4dB	9.69	9.49	12.84	13.50	14.00
6dB	7.52	7.92	12.69	13.72	15.48
Average	10.78	11.76	13.75	14.72	15.95

increments averagely. Among the FChT based results, the mean SIR results of 1024 samples frame length is better than the other two frame lengths in the experiments, and compared with that result, the SIR of multi-scale increases 1.23dB averagely.

4. DISCUSSIONS AND CONCLUSIONS

In this paper, a multi-scale FChT is adopted for time frequency representation of mixture speech. This method is suitable to analyze the speech with fast variation and the mixture speech with multiple FM rates of the simultaneous pitches. Experimental results prove the validity of the approach in monaural speech separation. In this paper, the unvoiced speech is not considered, the computational cost is rather high, and the sequential grouping strategy should be improved. Therefore all above are our future works.

5. ACKNOWLEDGEMENTS

The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030), the National High Technology Research and Development Program of China (2006AA01Z 196; 2006AA010103), the National Key Basic Research Program of China (2004CB318105), and the New-Century Training Programme Foundation for the Talents by the Ministry of Education of China.

6. REFERENCES

[1] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. ASSP*, vol. 34, pp. 744–754, 1986.

[3] T. F. Quatieri and R. G. Daisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. ASSP*, vol. 38, pp. 56–69, 1990.

[4] A.S. Bregman, "Auditory scene analysis: The perceptual organization of sound," *in Cambridge,MA:MIT Press*, 1990.

[5] D. L. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms and applications," *in Wiley/IEEE Press*, 2006.

[6] J. Lim K. Kim, K. Hong, "Multiresolution sinusoidal speech model using elliptic band pass filter," *NOLISP*, pp. 70–75, 2005.

[7] L. Weruaga and M. Képesi, "Speech analysis with the fast chirp transform," *Proc. EUSIPCO*, pp. 1011–1014, 2004.

[8] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech Communication*, vol. 48, pp. 474–492, May 2006.

[9] M. P. Cooke, "Modeling auditory processing and organization," *in Cambridge University Press*, 1993.

[10] R. Gribonval C. Févotte and E. Vincent, "Bss_eval toolbox user guide," *IRISA Technical Report*, April 2005.