2-D PROCESSING OF SPEECH FOR MULTI-PITCH ANALYSIS

Tianyu T. Wang and Thomas F. Quatieri

MIT Lincoln Laboratory
[*ttwang*, *quatieri*]@ll.mit.edu

ABSTRACT[1]

This paper introduces a two-dimensional (2-D) processing approach for the analysis of multi-pitch speech sounds. Our framework invokes the short-space 2-D Fourier transform magnitude of a narrowband spectrogram, mapping harmonically-related signal components to multiple concentrated entities in a new 2-D space. First, localized time-frequency regions of the spectrogram are analyzed to extract pitch candidates. These candidates are then combined across multiple regions for obtaining separate pitch estimates of each speech-signal component at a single point in time. We refer to this as *multi-region analysis* (MRA). By explicitly accounting for pitch dynamics within localized time segments, this separability is distinct from that which can be obtained using short-time autocorrelation methods typically employed in state-of-the-art multi-pitch tracking algorithms. We illustrate the feasibility of MRA for multi-pitch estimation on mixtures of synthetic and real speech.

Index Terms— ***2-D speech processing, Grating Compression Transform, multi-pitch analysis, segmental pitch dynamics***

1. INTRODUCTION

Estimating the pitch values of concurrent speech sounds from a single recording is a fundamental challenge in speech analysis. Typical approaches involve processing of short-time and band-pass signal components along single time or frequency dimensions. In contrast, in this paper, we address multi-pitch estimation using a two-dimensional (2-D) processing framework. 2-D analysis for pitch estimation was previously proposed in [1] where 2-D Fourier transforms were computed over localized time-frequency regions of a narrowband spectrogram, a representation referred to as the Grating Compression Transform (GCT). The GCT was observed to coherently represent pitch information in a transformed 2-D space and used for a single-pitch estimation task in noise. The current paper builds on this previous effort.

First, we show in Section 2 that the GCT of any localized time-frequency region is capable of accurately representing the pitch of a single source across all frequencies. Earlier approaches on the GCT [1] were shown to provide accurate pitch measurements only in low-frequency portions of the spectrogram. This finding leads to a novel multi-region analysis (MRA) method for multi-pitch signals described in Section 3. We show that MRA provides a form of separability of pitch information distinct from that obtained from short-time autocorrelation analysis used in state-of-the-art multi-pitch tracking methods [2][3]. Section 4 then demonstrates the value of MRA for multi-pitch estimation on synthetic and real speech with two-component mixtures having typical pitch relations. Our results show that MRA is a promising analysis framework for new or existing multi-pitch tracking systems[2]. Section 5 concludes with discussion of future directions.

2. BACKGROUND AND FRAMEWORK

Consider a localized time-frequency region $s[n,m]$ (discrete-time and frequency: *n*, *m*) of a narrowband short-time Fourier transform (STFT) magnitude (or log-magnitude) exhibiting harmonic line structure. A simple example of this condition is shown in Figures 1a and 1b for an impulse train with linearly increasing pitch (125-200 Hz). A 2-D sinewave model for $s[n,m]$ is [1]

$$s[n,m] \approx K + \cos(\omega_s \Phi[n,m]) \tag{1}$$

where ω_s denotes the local spatial frequency of the sinusoid, $\Phi[n,m]$ is a 2-D phase term indicating its orientation, and *K* is a constant DC term. $\Phi[n,m]$ is defined as

$$\Phi[n,m] = n\sin\theta + m\cos\theta \tag{2}$$

where θ is the angle of rotation of the harmonic lines relative to the time axis. The 2-D Fourier transform of $s[n,m]$ is then

$$\begin{aligned} S(\omega,\Omega) = 2\pi K\delta(\omega,\Omega) &+ 2\pi\delta(\omega + \omega_s \sin\theta, \Omega - \omega_s\cos\theta) \\ &+ 2\pi\delta(\omega - \omega_s\sin\theta, \Omega + \omega_s\cos\theta) \end{aligned} \tag{3}$$

such that the harmonic line structure maps to a set of impulses in the GCT (Figure 1c). In [1], ω_s (*radial* distance of impulses to the GCT origin) was observed to be inversely proportional to the pitch of $s[n,m]$ for low-frequency portions of the spectrogram: $f_{0,radial} = \frac{1}{N_{STFT}}\frac{2\pi fs}{\omega_s}$. We have found, however, that $\omega_s\cos\theta$ (*vertical* distance of impulses to the GCT origin) better represents pitch across *all* frequency regions: $f_{0,vertical} = \frac{1}{N_{STFT}}\frac{2\pi fs}{\omega_s\cos\theta}$. *fs* is the sampling rate of the waveform, and N_{STFT} is the discrete-Fourier transform (DFT) length used to compute the spectrogram.

[1] This work was supported by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. The work of T.T. Wang was additionally supported by the National Institutes of Deafness and Other Communicative Disorders under grant 5 T 32 DC00038.

[2] In this paper, we use *pitch estimation* as distinct from *pitch tracking* which entails assigning pitch estimates to mixture components.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **SEP 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **2-D Processing of Speech for Multi-Pitch Analysis**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Lincoln Laboratory,244 Wood Street,Lexington,MA,02420**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Interspeech 2009, Brighton, UK, Sept. 9, 2009. U.S. Government or Federal Rights License

14. ABSTRACT

see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

Figure 1d compares $f_{0,radial}$ and $f_{0,vertical}$ computed across multiple frequency regions (Figure 1a, rectangles) for a local time segment (Figure 1a, arrow) by peak-picking of the GCT magnitude. While $f_{0,vertical}$ remains constant across frequency regions and corresponds to the true pitch value at the center of segment (~160 Hz), $f_{0,radial}$ decreases across frequency regions by ~10 Hz. This effect is presumably due to the increased fanning of harmonic line structure in higher-frequency regions with changing pitch. Note that this comparison implies that rotation of the GCT components (i.e., θ) *increases* from low- to high-frequency regions.

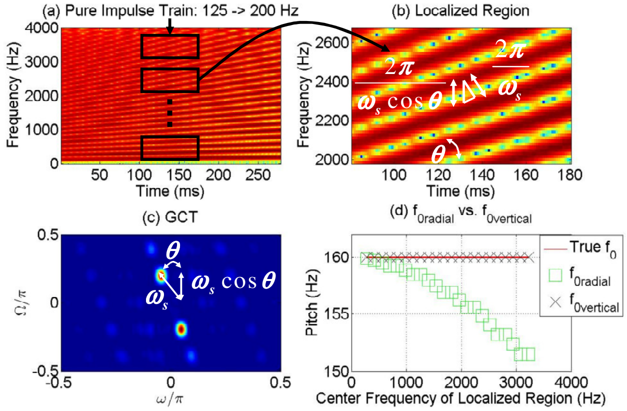

Figure 1. (a) Spectrogram with localized regions (rectangles) across frequency for a single time segment (arrow); (b) Zoomed-in region from (a); vertical ($\frac{2\pi}{\omega_s \cos\theta}$) and spatial ($\frac{2\pi}{\omega_s}$) distance between harmonic lines; 2-D sinewave orientation (θ); (c) GCT (magnitude) of (b) with ω_s , θ , $\omega_s \cos\theta$; DC component removed for display purposes. (d) Pitch estimates obtained from vertical vs. radial distances to GCT origin.

3. ANALYSIS OF MULTI-PITCH SIGNALS

This section discusses GCT analysis of multi-pitch signals. Section 3.A discusses separability of pitch information in the GCT while Section 3.B shows that this separability is distinct from that obtained from short-time autocorrelation-based analysis.

A. Separability of Pitch Information in the GCT

Extending (2) and (3) to the case of *N* concurrent signals,

$$s[n,m] \approx \sum_{i=1}^{N}\left(K_i + \cos(\omega_i \Phi[n,m;\theta_i])\right) \tag{4}$$

$$\begin{aligned} S(\omega,\Omega) = 2\pi\sum_{i=1}^{N} K_i\delta(\omega,\Omega) + 2\pi\sum_{i=1}^{N}\delta(\omega+\omega_i\sin\theta_i, \Omega-\omega_i\cos\theta_i) \\ +2\pi\sum_{i=1}^{N}\delta(\omega-\omega_i\sin\theta_i, \Omega+\omega_i\cos\theta_i). \end{aligned} \tag{5}$$

Here, we approximate the (log)-magnitude STFT of a mixture of signals as the *sum* of the STFT (log)-magnitudes computed for each individual signal. This approximation holds best when the contribution to the STFT from distinct sources occupies different frequency bands. Nonetheless, as we will show, separation of pitch in the GCT can be maintained even when these conditions do not necessarily hold, i.e., when a frequency band contains more than one source (with or without similar pitch values). The GCT's potential to separate pitch information was previously observed in phenomenological analyses by Quatieri [1] and Ezzat, et al [4].

To see how this separability can occur, consider a region of the spectrogram having two sets of harmonic lines corresponding to two distinct pitch trajectories that are *constant* through $s[n,m]$ (Figure 2a); the corresponding GCT (Figure 2b) would exhibit two sets of impulses along the Ω -axis. In this case, separability can only be achieved when the two pitch values are sufficiently different. The present argument also generalizes to the case when the two trajectories in $s[n,m]$ move at the same rate and direction.

Figures 2c and d illustrate a condition in which two pitch trajectories have *equal* pitch values defined at the center of $s[n,m]$ in time, but are moving in opposite directions at the same rate. Despite the overlap of harmonic structure in $s[n,m]$, the GCT maintains separability of pitch information due to its *explicit* representation of the underlying temporal trajectories of the two sources in the values of θ_i in (5) (i.e., $\theta_1 = -\theta_2 = \theta$). More generally, this separability holds under conditions where the rates of change of the two pitch trajectories are different (i.e., $|\theta_1| \neq |\theta_2|$).

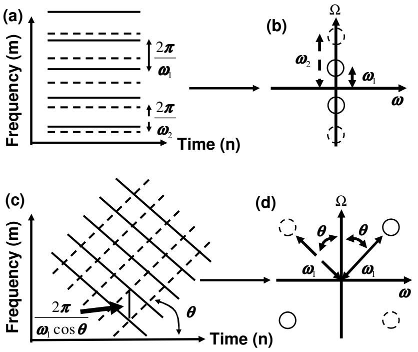

Figure 2. (a) Localized region with two distinct pitch values and no temporal change; (b) GCT corresponding to (a); separability occurs along the Ω -axis; (c) Localized region with two pitch candidates with the *same* pitch value but different temporal dynamics; (d) GCT corresponding to (c); separability from difference in temporal dynamics.

Finally, recall from Section 2 that for moving pitch trajectories, θ increases from low- to high-frequency regions (Figure 1d). Consequently, analysis of multiple regions across frequency and time is expected to provide better separability of pitch information than that of a *single* low-frequency region across time as in [1].

B. Comparison to Short-time Autocorrelation Analysis

Multi-pitch tracking systems typically obtain pitch candidates from autocorrelation analysis of band-pass filtered versions of the waveform on a *frame-by-frame* basis (e.g., [2][3]). This approach provides distinct pitch candidates for a single point in time but does not represent the pitch dynamics of multi-pitch signals. Here, we show that the GCT's representation of pitch dynamics within a local time segment invokes separability of pitch information distinct from that obtained in short-time autocorrelation analysis.

Figure 3 shows analyses of two synthesized concurrent vowels with rising and falling pitch contours of 150-200 Hz and 200-150 Hz across a 200 ms duration. Figure 3a shows the formant structure for the vowels. In Region 1 (R1), the rising vowel exhibits a formant peak while the falling vowel exhibits a valley; in Region 2 (R2), a formant peak is present for both vowels. Analyses are done at the center of the mixture where both sources have pitch values ~175 Hz.

For comparison, two linear-phase band-pass filters centered at the formant peaks of R1 and R2 were applied to the waveform. To obtain an envelope [2], filtered waveforms were then half-wave rectified and low-pass filtered (cutoff = 800 Hz). The normalized autocorrelation ($r_{xx}[n]$) was computed for a 30-ms duration of the envelopes (Figure 3b-c). For R1, a single distinct pitch estimate and its sub-harmonics are present (Figure 3b, arrow); however, $r_{xx}[n]$ for R2 (Figure 3c) reflects the interaction of closely-spaced periodicities and appears "noisy". These observations are similar to those observed by Wu, et al. [2] in which these "noisy" bands were discarded in favor of those exhibiting a dominant pitch to compute a summary correlogram at a single point in time.

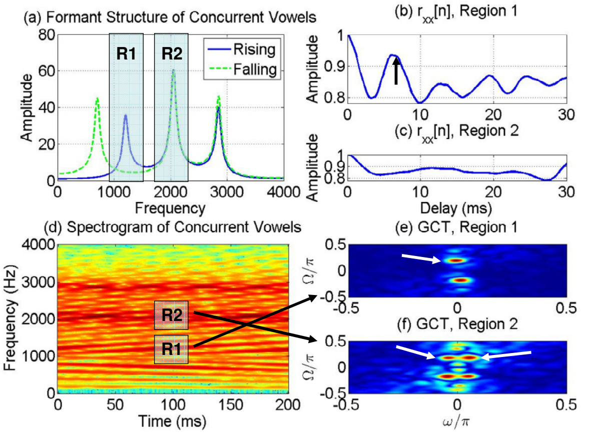

Figure 3. (a) Formant structure of rising- (dashed) and falling-pitched (solid) vowels in Region 1 (R1) and Region 2 (R2); (b) $r_{xx}[n]$ for R1 with lag (at arrow) corresponding to dominant pitch; (c) $r_{xx}[n]$ for R2; (d) Spectrogram of concurrent vowels; (e) GCT for R1 with dominant pitch (arrow); (f) GCT for R2 with two sets of pitch peaks (arrows).

Figures 3e-f show GCTs computed over localized time-frequency regions at R1 and R2 (Figure 3d). A single dominant set of impulses, corresponding to a single pitch value, is present in the GCT for R1, similar to $r_{xx}[n]$ for R1; however, *two* distinct sets of peaks can be seen for R2 (Figure 3e-f, arrows) corresponding to two similar pitch values. *The GCT can therefore separate pitch information of two speakers with similar energies and pitch values in a localized set of frequency bands by exploiting the temporal dynamics of their underlying pitch trajectories*. This separability is distinct from that obtained using short-time autocorrelation analysis (compare with Figure 3c). Recall that this separability generalizes to the case where source signals exhibit similar energies but different pitch values/temporal dynamics (Section 3).

4. MULTI-PITCH ESTIMATION

This section demonstrates multi-pitch estimation using GCT-based MRA. Our goal is to assess the value of GCT-based MRA for accurately obtaining pitch estimates rather than assigning estimates to distinct speakers across time (i.e., pitch *tracking*). Section 4.A describes the real and synthetic speech mixtures used. Section 4.B and C describe our analysis and post-processing methods, respectively. Section 4.D presents our results.

4.A Synthetic and Real Speech Mixtures

Concurrent vowels with linear pitch trajectories spanning 300 ms were synthesized using a glottal pulse train and an all-pole formant envelope with formant frequencies of 860, 2050, and 2850 Hz and bandwidths of 56, 65, 70 Hz (*/ae/*) [5]. For real speech, two all-voiced sentences spoken by a male and female were used. Two cases were analyzed to illustrate typical pitch-trajectory conditions: 1) *separate* or 2) *crossing* trajectories within the utterance. All signals were mixed at 0 dB overall signal-to-signal ratio (SSR) and pre-emphasized prior to analysis. True pitch values were obtained using a single-pitch estimator on the signals prior to mixing [6].

4.B GCT-based MRA Methodology

The log-STFT magnitude was computed for all mixtures with a 25-ms Hamming window, 1-ms frame interval, and 512-point DFT. Time-frequency regions of size 100 ms by 700 Hz were extracted from the spectrogram at a 5-ms and 140-Hz region interval in time and frequency, respectively. A 2-D gradient operator was applied to the spectrogram prior to extraction to reduce the contribution of the DC and near-DC components to the GCT. To obtain pitch candidates for each region, the GCT magnitude was multiplied by three binary masks derived from thresholding the 1) overall amplitude, 2) gradient (∇GCT), and 3) Laplacian (ΔGCT). The thresholds were chosen as $\max(GCT)/3$, $\max(\nabla GCT)/3$, and $\min(\Delta GCT)/3$. Region growing was performed on the masked GCT, and pitch candidates were obtained by extracting the location of the maximum amplitude in each resulting region. Candidates corresponding to the two largest amplitudes were kept for each time-frequency region. In the case where only a single pitch value is present, the value is assigned twice to the region.

4.C Post-processing

For synthetic speech, a simple *clustering* method was used to assign pitch values at each point in time from the candidates of GCT-based MRA. All candidates at a single point in time were collected and sorted, and the median of the top and bottom halves of the collection were then chosen as the two pitch values. A similar technique was used for real speech; however, due to the longer duration of these signals, we sought to exploit the temporal continuity of the underlying pitch contours in clustering. At each 5-ms interval for a time-frequency region, pitch candidates from its neighboring regions in time spanning 100-ms and across frequencies were combined for clustering. To compare GCT-based MRA with previous work [1], we also assigned to each 5-ms interval the two candidates from analyzing a *single* low-frequency region. Figure 4 illustrates these post-processing methods.

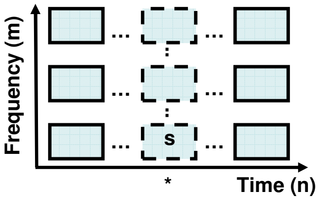

Figure 4. Post-processing methods for assigning pitch value at time *; 's' denotes single low-frequency region used as in [1]; dashed regions denote regions used in clustering for synthetic speech; shaded regions denote regions used in clustering for real speech.

Finally, *oracle* pitch values were obtained by assigning to each time point the pitch candidate from GCT-based MRA closest in frequency to the true pitch values. The accuracy of these estimates is viewed as assessing the value of GCT-based MRA for obtaining pitch candidates independent of post-processing (e.g., clustering).

4.C. Results

Figures 5 - 7 show estimates for the synthetic and real speech mixtures. The *total-best* percent error between estimates and truth for both source signals was computed at each time point:

$$\%error = 100\left[\frac{\left|f_1-\hat{f}\right|}{f_1}+\frac{\left|f_2-\hat{f}\right|}{f_2}\right]. \qquad (6)$$

$\hat{f}$ is the estimate from *clustering*, *single*, or *oracle* closest in frequency to the true pitch values f_1 and f_2. Table 1 gives average *%error*'s (*%error*$_{avg}$) computed across time for all cases.

For the synthetic concurrent-vowels task (Syn1-4, Figure 5), GCT-based MRA provides accurate estimates under a variety of mixed pitch trajectories. The oracle estimates follow the true pitch values with *%error*$_{avg}$ < 0.04% while the clustering scheme assigns pitch values across time for GCT-based MRA with *%error*$_{avg}$ < 1.75% (Table 1). Observe also that the oracle and clustering of pitch candidates derived from GCT-based MRA exhibits lower *%error*$_{avg}$ than single-region analysis in all cases.

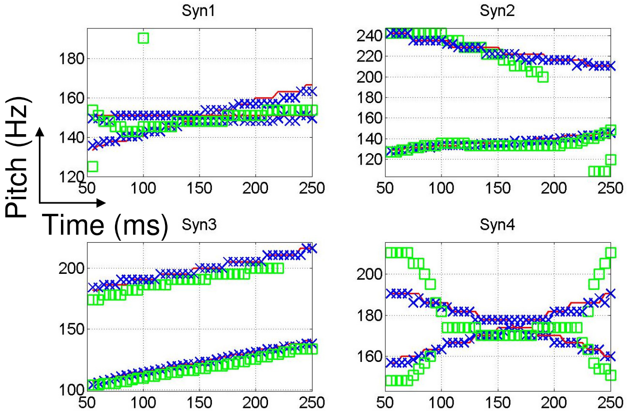

Figure 5. Concurrent vowel pitch estimates; *clustering* (x/blue), *single* (square/green), and true pitch values (solid/red): (1) rising 125-175 Hz + constant 150 Hz, (2) falling 250-200 Hz + rising 125-150 Hz, (3) rising 100-150 Hz + rising 175-225 Hz, (4) falling 200-150 Hz + rising 150-200 Hz. Estimates start (end) at 50 (250) ms to remove edge effects.

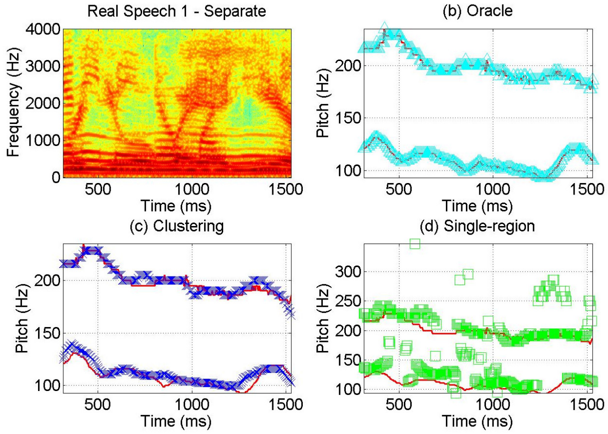

Figure 6. (a) All-voiced mixture spectrogram with separate pitch trajectories, Male - "Why were you away a year?" + Female - "Nanny may know my meaning."; first 250 ms and last 50 ms excluded to remove edge effects in clustering due to initial and final silent regions; (b-d) *truth* (solid/red), *oracle* (triangle/light blue, b), *clustering* (x/blue, c), *single* (square/green, d).

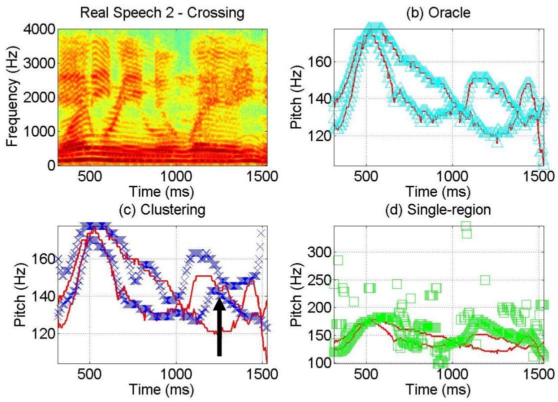

Figure 7. As in Figure 6 but with all-voiced mixture with crossing pitch trajectories: Male - "Why were you away a year?" + Male - "Nanny may know my meaning."; (c) arrow denotes "jump" due to differences in energy between sources in localized time-frequency regions.

Table 1. Average %errors across time for each mixture and method.

	Syn1	Syn2	Syn3	Syn4	Real1	Real2
oracle	0.00	0.00	0.03	0.00	0.00	0.00
clustering	1.74	1.44	1.00	1.08	5.46	7.91
single	5.98	14.97	13.24	9.13	42.18	20.13

For real speech, the oracle pitch values match truth with 0.00% average error in both separate and crossing conditions. Although close to truth for the separate case, it appears that median-based clustering is not optimal for exploiting the oracle candidates in the crossing case, with jumps in pitch values from distinct talkers (e.g., Figure 7c, arrow). This is likely due to the inability of the clustering method to account for points in time in which one speaker is dominant in energy. Nonetheless, the accuracy of the oracle estimates demonstrates the feasibility of employing GCT-based MRA for multi-pitch estimation with an improved post-processing method. Finally, as in the synthetic cases, the oracle and clustering of the GCT-based MRA pitch candidates outperform the single-region method, thereby further illustrating the benefits of exploiting multiple regions for analysis.

5. DISCUSSION

This paper has shown GCT-based MRA provides separability of pitch information for a variety of multi-pitch signals. Since the GCT can separate pitch information from multiple sources of similar energies, the assumption of a *single* dominant source does not need to be invoked when obtaining candidates in localized time-frequency regions as typically done for short-time autocorrelation analysis (e.g., [2]). The accuracy of the pitch estimates obtained using GCT-based MRA on real and synthetic mixtures demonstrates the feasibility of employing this analysis framework in conjunction with existing multi-pitch tracking techniques (e.g., those based on hidden Markov models [2]).

6. REFERENCES

[1] T.F. Quatieri, "2-D Processing of Speech with Application to Pitch Estimation," ICLSP, September 2002.

[2] M. Wu, D.L. Wang, and G. Brown, "A Multi-Pitch Tracking Algorithm for Noisy Speech," *IEEE TSAP*, 11:229-241, 2003.

[3] T. Tolonen and M. Karjalainen., "A Computationally Efficient Multi-pitch Analysis Model, " *IEEE TSAP*, 8:708-716, 2000.

[4] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal Analysis of Speech Using 2-D Gabor Filters," Interspeech 2007.

[5] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[6] Y. Medan, E. Yair, and D. Chazan, "Super Resolution Pitch Determination of Speech Signals," *IEEE TSP*, 39:40-48, 1991.