TOWARDS CO-CHANNEL SPEAKER SEPARATION BY 2-D DEMODULATION OF SPECTROGRAMS[1]

Tianyu T. Wang

MIT Lincoln Laboratory
ttwang@ll.mit.edu

Thomas F. Quatieri

MIT Lincoln Laboratory
quatieri@ll.mit.edu

ABSTRACT

This paper explores a two-dimensional (2-D) processing approach for co-channel speaker separation of voiced speech. We analyze localized time-frequency regions of a narrowband spectrogram using 2-D Fourier transforms and propose a 2-D amplitude modulation model based on pitch information for single and multi-speaker content in each region. Our model maps harmonically-related speech content to concentrated entities in a transformed 2-D space, thereby motivating 2-D demodulation of the spectrogram for analysis/synthesis and speaker separation. Using a priori pitch estimates of individual speakers, we show through a quantitative evaluation: 1) Utility of the model for representing speech content of a single speaker and 2) Its feasibility for speaker separation. For the separation task, we also illustrate benefits of the model's representation of pitch dynamics relative to a sinusoidal-based separation system.

Index Terms— Grating Compression Transform, speaker separation, spectrogram demodulation, 2-D speech analysis

1. INTRODUCTION

Co-channel speaker separation is a challenging task in audio processing. For all-voiced speech, current methods operate on short-time frames of mixture signals (e.g., harmonic suppression, sinusoidal analysis, modulation spectrum [1 - 3]) or on single units of a time-frequency distribution (e.g., binary masking [4]). Alternatively, this paper proposes and assesses the feasibility of a 2-D analysis framework for this task. We analyze localized time-frequency *regions* of a narrowband spectrogram using 2-D Fourier transforms, a representation we refer to as the Grating Compression Transform (GCT).

The GCT has been explored by Quatieri [5], Ezzat et al [6, 7], and Wang and Quatieri [8] primarily for single-speaker analysis and is consistent with physiological modeling studies implicating 2-D analysis of sounds by auditory cortex neurons [9]. Ezzat et al. performed analysis/synthesis of a single speaker using 2-D demodulation of the spectrogram [7]. In [8], we proposed an alternative 2-D modulation model for formant analysis. Phenomenological observations in [5, 6] have also suggested that the GCT invokes *separability* of multiple speakers. Finally, in recent work, we have demonstrated the GCT's ability in analysis of multi-pitch signals [10]. This paper builds on these previous efforts in several ways.

First, in Section 2.1, we investigate GCT analysis of a single speaker using a 2-D amplitude modulation (AM) model based on pitch information. Section 2.2 extends this model to analysis of *multiple* speakers to account for the observations made in [5,6] regarding speaker separability in the GCT. Our framework motivates 2-D sinusoidal *demodulation* of the spectrogram for: 1) single-speaker analysis/synthesis and 2) speaker separation. Section 3 describes algorithms for these tasks. Section 4 presents a quantitative evaluation of these methods on real speech to assess: 1) *Utility* of the AM model in representing speech content of a single speaker and 2) Its *feasbility* for the separation task using a priori pitch estimates of individual speakers. As a baseline, we compare against a sinusoidal-based separation system that similarly uses such pitch estimates [2]. Section 5 concludes with future directions.

2. 2-D PROCESSING FRAMEWORK

2.1. Single-speaker Model

Consider a *localized* time-frequency region $s[n,m]$ (discrete-time and frequency n, m) of a narrowband short-time Fourier transform magnitude (STFTM) (Figure 1) computed for a single voiced utterance. Here, we extend a 2-D amplitude modulation (AM) model from our previous work [8] such that

$$\begin{aligned} s[n,m] &\approx (\alpha_0 + \cos(\Phi[n,m]))a[n,m] \\ \Phi[n,m] &= \omega_s(n\cos\theta + m\sin\theta) + \varphi. \end{aligned} \quad (1)$$

i.e., a sinusoid with spatial frequency ω_s, orientation θ, and phase φ rests on a DC pedestal α_0 and modulates a slowly-varying envelope $a[n,m]$. The 2-D Fourier transform of $s[n,m]$ (i.e., the GCT) is

$$\begin{aligned} S(\omega,\Omega) = \alpha_0 A(\omega,\Omega) &+ 0.5e^{-j\varphi}A(\omega+\omega_s\sin\theta, \Omega-\omega_s\cos\theta) \\ &+ 0.5e^{j\varphi}A(\omega-\omega_s\sin\theta, \Omega+\omega_s\cos\theta) \end{aligned} \quad (2)$$

where ω and Ω map to n and m, respectively. The sinusoid represents the harmonic structure associated with the speaker's pitch [5, 10]. Denoting f_s as the waveform sampling frequency and N_{STFT} as the discrete-Fourier transform (DFT) length of the STFT, the GCT parameters relate to the speaker's pitch (f_0) at the center (in time) of $s[n,m]$ (Figure 1b, c) [5, 10]:

$$f_0 = (2\pi f_s)/(N_{STFT}\omega_s\cos\theta). \quad (3)$$

[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.

Report Documentation Page

Form Approved
OMB No. 0704-0188

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

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Towards Co-Channel Speaker Separation by 2-D Demodulation of Spectrograms	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

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**	

A change in f_0 (Δf_0) across Δn results in an *absolute* change in frequency of the k^{th} pitch harmonic by $k\Delta f_0$. Therefore, in a localized time-frequency region (Figure 1b)

$$\tan\theta \approx (k\Delta f_0)/\Delta n \,. \qquad (4)$$

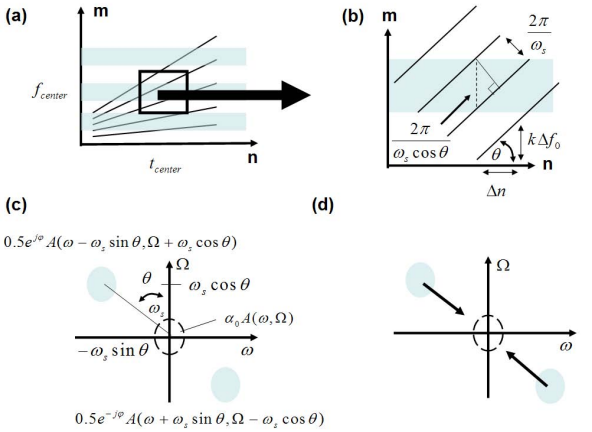

Figure 1. (a) Schematic of full STFTM with localized time-frequency region centered at t_{center} and f_{center} for GCT analysis (rectangle); (b) Localized region of (a) with harmonic structure (parallel lines) and envelope (shaded); triangle indicates spacing between harmonic lines; note also relation between θ, $k\Delta f_0$, and Δn; (c) GCT of (a) with baseband (dashed) and modulated (shaded) versions of the envelope; (d) Demodulation to recover near-DC terms.

For a particular $s[n,m]$ with center frequency f_{center} (Figure 1a), f_0 can be obtained from (3) such that $k \approx f_{center}/f_0$. The rate of change of f_0 ($\partial f_0/\partial t$) in $s[n,m]$ is then

$$\partial f_0/\partial t \triangleq \Delta f_0/\Delta n = (f_0 \tan\theta)/f_{center} \,. \quad (5)$$

Finally, φ corresponds to the position of the sinusoid in $s[n,m]$; for a non-negative DC value of $a[n,m]$, φ can be obtained by analyzing the GCT at $(\omega=\omega_s\sin\theta,\ \Omega=\omega_s\cos\theta)$

$$\varphi = angle[S(\omega_s\sin\theta,\omega_s\cos\theta)] \,. \quad (6)$$

Our model maps harmonically related speech content in each $s[n,m]$ to *concentrated* entities in the GCT near DC and at 2-D "carriers" (Figure 1c). Observe that if the near-DC terms were removed or corrupted, our model motivates approximate *recovery* of the near-DC terms from the carrier terms using sinusoidal demodulation (Figure 1d). Using demodulation, the full STFTM can then be recovered and combined with the STFT phase for approximate waveform reconstruction.

2.2. Multi-speaker Extension

In [5, 6], the GCT space was suggested to separate multiple speakers. To account for these observations, we approximate the STFTM computed for a *mixture* of *N* speakers in a localized time-frequency region $x[n,m]$ as the *sum* of their individual magnitudes. Using the model of (1), we then have

$$x[n,m] \approx \sum_{i=1}^{N} \alpha_{0,i} a_i[n,m] + \sum_{i=1}^{N} a_i[n,m]\cos(\omega_i[n\cos\theta_i + m\sin\theta_i] + \varphi_i). \qquad (7)$$

Equation (7) invokes the sparsity of harmonic line structure from distinct speakers in the STFTM (i.e., when harmonic components of speakers' are located at different frequencies). Nonetheless, separation of speaker content in the GCT can still be maintained when speakers exhibit harmonics located at *identical* frequencies (e.g., due to having the same pitch values, when pitch values are integer multiples of each other) due to its representation of pitch dynamics through θ in (7) [10]. An example of this is shown schematically in Figure 2a-b, where two speakers have equal pitch values but distinct pitch dynamics, thereby allowing separability in the GCT.

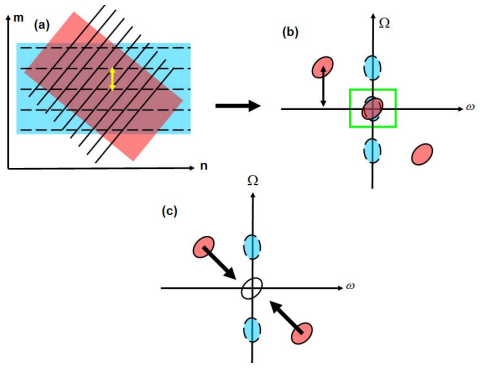

Figure 2. (a) Localized time-frequency region of STFTM computed on a mixture of two speakers; speaker with rising pitch and falling formant structure (solid, red); speaker with stationary pitch and stationary formant (dashed, blue); yellow arrow denotes same pitch value at center of region (b) GCT of (a) showing overlap of near-DC terms (green rectangle); speakers exhibit the same vertical distances (black arrow) from the ω-axis corresponding to equal pitch values; separability is maintained due to distinct angular positions off of the Ω-axis; (c) Demodulation to recover near-DC terms of one speaker.

The 2-D Fourier transform of (7) is

$$X(\omega,\Omega) = \sum_{i=1}^{N} \alpha_{0,i} A_i(\omega,\Omega) + 0.5\sum_{i=1}^{N} A_i(\omega+\omega_i\sin\theta_i, \Omega-\omega_i\cos\theta_i)e^{-j\varphi_i} + 0.5\sum_{i=1}^{N} A_i(\omega-\omega_i\sin\theta_i, \Omega+\omega_i\cos\theta_i)e^{j\varphi_i} \,. \qquad (8)$$

For slowly-varying $A_i(\omega,\Omega)$, the contribution to $X(\omega,\Omega)$ from multiple speakers exhibits *overlap* near the GCT origin (Figure 2b); however, as in the single-speaker case, $A_i(\omega,\Omega)$ can be estimated through sinusoidal demodulation according to the proposed model. This model therefore motivates localized 2-D *demodulation* of the STFTM computed for a mixture of speakers for the speaker separation task (Figure 2c).

3. ALGORITHMS

Herein we discuss algorithms motivated by the models of Section 2. Section 3.1 discusses 2-D demodulation of the STFTM for analysis/synthesis of a single speaker. Our approach is distinct from work by Ezzat et al. in which scattered data interpolation was used for demodulation [7]. In this work, we apply sinusoidal demodulation in conjunction with a least-squared error fit to estimate the gain parameter in (1). Section 3.2 describes a similar algorithm for the speaker separation task. Both methods assume a priori pitch estimates of individual speakers.

3.1. Single-speaker Analysis/Synthesis

To assess the AM model's ability to represent speech content of a single speaker, an STFT is computed for the signal using a 20-ms Hamming window, 1-ms frame interval, and 512-point DFT. From the *full* STFTM ($s_F[n,m]$), localized regions centered at *k* and *l* in time and frequency ($s_{kl}[n,m]$) of size 625 Hz by 100 ms are extracted using a 2-D Hamming window ($w_h[n,m]$) for GCT analysis. We then apply a high-pass filter $h_{hp}[n,m]$ to each $s_{kl}[n,m]$ to *remove* $\alpha_0 A(\omega,\Omega)$ in (2); we denote this result as $s_{kl,hp}[n,m]$. $h_{hp}[n,m]$ is a circular filter with cut-offs at $\omega=\Omega=0.1\pi$, corresponding in ω to a ~300 Hz upper limit of f_0 values observed in analysis.

For each $s_{kl,hp}[n,m]$, we aim to approximately *recover* $\alpha_0 A(\omega,\Omega)$ using 2-D sinusoidal demodulation. The carrier ($\cos(\Phi[n,m])$) parameters are determined from the speaker's pitch track using (3) for ω_s and (6) for φ. To determine θ, a linear least-squared error fit is applied to the pitch values spanning the 100-ms duration of $s_{kl,hp}[n,m]$. The slope of this fit approximates $\partial f_0/\partial t$ such that θ is estimated using (5). $s_{kl,hp}[n,m]$ is multiplied by the carrier generated from these parameters followed by filtering with a circular low-pass filter $h_{lp}[n,m]$ with cut-offs at $\omega=\Omega=0.1\pi$; we denote this result as $\hat{a}[n,m]$. $\hat{a}[n,m]$ is combined with the carrier using (1) and set equal to $s_{kl}[n,m]$

$$s_{kl}[n,m]=(\alpha_0+\cos(\Phi[n,m]))\hat{a}[n,m]. \quad (9)$$

For each time-frequency unit of $s_{kl}[n,m]$, (9) corresponds to a linear equation in α_0 since the values of $s_{kl}[n,m]$, $\hat{a}[n,m]$, and $\cos(\Phi[n,m])$ are known. This overdetermined set of equations is solved in the least-squared error (LSE) sense. The resulting estimate of $s_{kl}[n,m]$ using the estimated α_0, $\hat{a}[n,m]$, and $\cos(\Phi[n,m])$ is denoted as $\hat{s}_{kl}[n,m]$. The *full* STFTM estimate $\hat{s}_F[n,m]$ is obtained using overlap-add (OLA) with a LSE criterion (OLA-LSE) [11]

$$\hat{s}_F[n,m]=\frac{\sum_k\sum_l w_h[kT-n,lF-m]\hat{s}_{kl}[n,m]}{\sum_k\sum_l w_h{}^2[kT-n,lF-m]}. \quad (10)$$

OLA step sizes in time and frequency (*T* and *F*) are set to $1/4$ the size of $w_h[n,m]$. $\hat{s}_F[n,m]$ is then combined with the STFT phase for waveform reconstruction using OLA-LSE [11].

3.2. Speaker Separation

For speaker separation, the demodulation steps are nearly identical to those in Section 3.1 but applied to the *mixture* signal. Briefly, let $x_{kl}[n,m]$ be a localized region of the full STFTM computed for the *mixture* signal centered at *k* and *l* in time and frequency. $x_{kl}[n,m]$ is filtered with $h_{hp}[n,m]$ to remove the overlapping $\alpha_{0,i}A_i(\omega,\Omega)$ terms at the GCT origin (Figure 2b); we denote this result as $x_{kl,hp}[n,m]$. A cosine carrier for each speaker is generated using the corresponding pitch track and multiplied by $x_{kl,hp}[n,m]$ to obtain

$$\begin{aligned} x_{kl,i}[n,m] &= x_{kl,hp}[n,m]\cos(\omega_i[n\sin\theta_i+m\cos\theta_i]+\varphi_i) \\ &= \hat{a}_i[n,m]+c[n,m]. \end{aligned} \quad (11)$$

If the speakers' carriers are in distinct locations of the GCT, $c[n,m]$ summarizes cross terms *away* from the GCT origin such that $\hat{a}_i[n,m]$ can be obtained by filtering $x_{kl,i}[n,m]$ with $h_{lp}[n,m]$. For each speaker, $\hat{a}_i[n,m]$ is combined with its respective carrier using (1). These results are summed and set equal to $x_{kl}[n,m]$ to solve for $\alpha_{0,i}$ in the LSE sense:

$$x_{kl}[n,m]=\sum_{i=1}^{N}(\alpha_{0,i}+\cos(\Phi[n,m]))\hat{a}_i[n,m] \quad (12)$$

Recall that the GCT represents pitch *and* pitch dynamics; it may therefore invoke improved speaker separability over representations relying solely on harmonic sparsity (Section 2.2). In a region where speakers have *equal* pitch values and the *same* temporal dynamics, however, (12) invokes a near-singular matrix. To address this, we compute the angle between the $\hat{a}_i[n,m]$ columns of the matrix. When this angle is below a threshold of $\pi/10$, the $\alpha_{0,i}$ is solved for by reducing the matrix rank to that corresponding to a single speaker.

Finally, the estimated full STFTMs of the target speakers are reconstructed using (10). Speaker waveforms are then reconstructed using OLA-LSE by combining the estimated STFTMs with the STFT phase of the *mixture* signal.

4. PRELIMNARY EVALUATION

This section describes preliminary evaluations of the algorithms of Sections 3.1 (denoted as Exp1) and 3.2 (Exp2). We analyzed two all-voiced sentences sampled at 8 kHz ("Why were you away a year, Roy?" and "Nanny may know my meaning") spoken by 10 males and females (40 total sentences). Pitch estimates of the individual sentences were determined prior to analysis from an autocorrelation-based pitch tracker.

In Exp1, we perform analysis/synthesis of a single speaker as described in Section 3.1. For comparison, we also generated a waveform by filtering $s_F[n,m]$ with an adaptive filter

$$h_s[n,m]=h_{lp}[n,m](1+2\cos(\omega_s[n\sin\theta+m\cos\theta]+\varphi_s)) \quad (13)$$

where ω_s, θ, and φ_s are determined for each localized time-frequency region using the speaker's pitch track and $h_{lp}[n,m]$ is that described in Section 3.1. The filtered STFTM is used to recover the waveform as in Section 3.1. This method assesses the value of the model for representing speech content of a single speaker, independent of the 2-D LSE fitting procedure.

To assess the feasibility of GCT-based speaker separation (Exp2), we analyzed mixtures of two sentences (Nanny + Roy) spoken by 10 males and females mixed at 0 dB (90 mixtures total). For comparison, we used a baseline sinewave-based separation system (SBSS); SBSS models sinewave amplitudes and phases given their frequencies (e.g., harmonics) for each

speech signal [2]. We chose this baseline for comparison as it similarly uses a priori pitch estimates to obtain the sinusoidal frequencies, and to assess potential benefits of the GCT's explicit representation of pitch dynamics (Section 3.2).

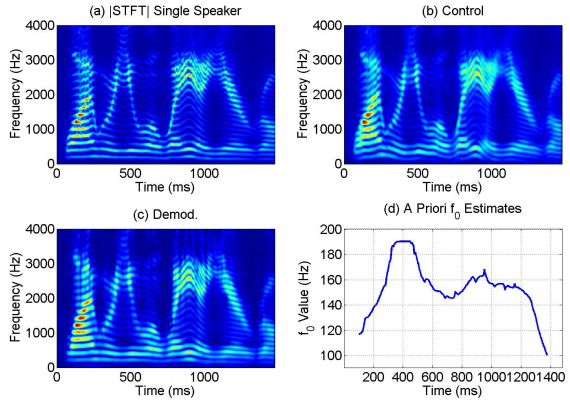

Figure 3. (a) STFT magnitude of single speaker sentence (Roy); (b) Recovered STFT magnitude using control method; (c) As in (b) but using demodulation; (d) A priori pitch estimates of sentence in (a) - (c).

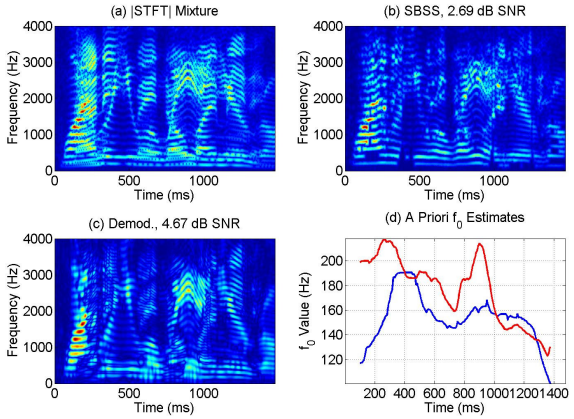

Figure 4. (a) STFT magnitude of mixture (Nanny + Roy); (b) Recovered STFT magnitude of Roy sentence using SBSS with resulting SNR listed; (c) As in (b) but using demodulation; (d) A priori pitch estimates of target (blue) and interfering (red) speakers; pitch tracks exhibit crossings throughout mixture.

Table I. Average SNRs across 40 single sentences (Exp1) and 90 mixtures (Exp2).

	Exp1 Filtering	Exp1 Demod.	Exp2 SBSS	Exp2 Demod.	Exp2 TruePhase
SNR (dB)	11.24	12.51	3.62	4.09	5.96

Figure 3 shows STFTMs obtained in the single-speaker experiment and a priori pitch estimates. In this example, demodulation appears to provide a similar reconstruction as the control method. In Figure 4, we show the resulting STFTMs for the separation task using the single-speaker sentence as the target. In this example, the pitch tracks of the target and interferer exhibit *crossings* (Figure 4d), thereby leading to overlapping harmonic structure in the mixture STFTM. Qualitatively, GCT demodulation appears to provide a more faithful reconstruction of the target than SBSS. To quantify the performance in Exp1 and Exp2, we computed average signal-to-noise ratios (SNR) of the original and reconstructed waveforms (Table I). In Exp1, demodulation provides a better reconstruction than filtering by ~1.3 dB. One possible cause for this is the introduction of negative magnitude values in the filtered STFTM. These effects are likely minimized in demodulation through the LSE fitting procedure. Nonetheless, both methods provide good reconstruction of the waveform with overall SNR > 11 dB. In Exp2, consistent with the recovered STFTMs (Figure 4), demodulation affords a larger gain in SNR than SBSS in the example shown (captions, Figure 4b, c) and on average. This is presumably due to the GCT's explicit representation of pitch dynamics. In informal listening for Exp1, subjects (non-authors) reported no perceptual difference between the filtering and demodulation methods in relation to the original signal. In Exp2, subjects reported intelligible reconstructions of the target speech for both methods with a reduced amplitude of the interferer. However, in assessing SBSS, subjects reported that the interferer sounded "metallic" while this synthetic quality was not perceived for the GCT system. Though more formal listening tests are needed, these observations demonstrate the utility of the AM model for representing speech content of a single speaker. Furthermore, they demonstrate the GCT's feasibility for speaker separation and its advantages in representing pitch dynamics for this task.

5. CONCLUSIONS

This paper has introduced a 2-D processing approach for single- and multi-speaker analysis. We have quantitatively shown that a 2-D modulation model accounting for near-DC terms of the GCT provides good representation of speech content of a single speaker. We have also shown that this model is a promising representation for co-channel speaker separation. For the separation task, one limitation of the current implementation is its use of the STFT phase computed for the *mixed* signal in reconstruction. Table I shows results of applying the STFTM obtained through demodulation with the *true* phase of the target resulting in an average SNR of ~6 dB. Future work will explore magnitude-only reconstruction [11] methods to address this discrepancy. We also aim to incorporate existing methods for multi-pitch analysis and estimation (e.g., [10, 12]) with the current framework towards a full separation system. Finally, the current framework may be extended for analysis/synthesis and separation of speech-like sources (e.g., musical instruments) due to its representation of harmonic (e.g., an instrument's pitch) and slowly-varying structure (e.g., an instrument's timbre, analogous to speech formants in localized regions of the STFTM.

6. REFERENCES

[1] D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel Speaker Separation by Harmonic Enhancement and Suppression," IEEE TSAP, v5, pp. 407-424, 1997.

[2] T. Quatieri and R. Danisewicz, "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech," IEEE TASSP, v38, 1990.

[3] S. Schimmel, L. Atlas, K. Nie, "Feasibility of Single Channel Speaker Separation Based on Modulation Frequency Analysis," ICASSP 2007, Honolulu, HI, USA.

[4] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE TNN, vol. 15, 2004.

[5] T. Quatieri, "2-D Processing of Speech with Application to Pitch Estimation," ICSLP 2002.

[6] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectrotemporal Analysis of Speech Using 2-D Gabor Filters," Interspeech 2007, Antwerp, Belgium.

[7] T. Ezzat, J. Bouvrie, and T. Poggio, "AM-FM Demodulation of Spectrograms Using Localized 2-D Max-Gabor Analysis," ICASSP, 2007.

[8] T. Wang and T. Quatieri, "Exploiting Temporal Change of Pitch in Formant Estimation," ICASSP, 2008, Las Vegas, NV, USA.

[9] T. Chi, P. Ru, S. Shamma, "Multiresolution Spectrotemporal Analysis of Complex Sounds," JASA v118, pp. 887 - 906, 2005

[10] T. Wang and T. Quatieri, "2-D Processing of Speech for Multi-Pitch Analysis," Interspeech 2009, Brighton, UK.

[11] T. Quatieri, Discrete-time Speech Signal Processing: Principles and Practice. Upper Saddle River, NJ: Prentice-Hall, 2001.

[12] Wu, M., Wang, D., and Brown, G., "A Multipitch Tracking Algorithm for Noisy Speech," IEEE TASL, v11, pp. 229 - 241, 2003.