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

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 feasibility 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

\[
s[n, m] = (a_s + \cos(\Phi[n, m]))a[n, m]
\]

(1)

where \( \Phi[n, m] = \omega_s (n \cos \theta + m \sin \theta) + \phi \).

i.e., a sinusoid with spatial frequency \( \omega_s \), orientation \( \theta \), and phase \( \phi \) rests on a DC pedestal \( a_s \) and modulates a slowly-varying envelope \( a[n, m] \). The 2-D Fourier transform of \( s[n, m] \) (i.e., the GCT) is

\[
S(\omega, \Omega) = a_s A(\omega, \Omega) + 0.5e^{-j\varphi}A(\omega + \omega_s \sin \theta, \Omega - \omega_s \cos \theta) + 0.5e^{j\varphi}A(\omega - \omega_s \sin \theta, \Omega + \omega_s \cos \theta)
\]

(2)

where \( \omega \) and \( \Omega \) 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]:

\[
f_0 = (2\pi f_s)/(N_{STFT} a_s \cos \theta).
\]

(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**

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. **REPORT DATE**
   OCT 2009

2. **REPORT TYPE**

3. **DATES COVERED**
   00-00-2009 to 00-00-2009

4. **TITLE AND SUBTITLE**
   Towards Co-Channel Speaker Separation by 2-D Demodulation of Spectrograms

5a. **CONTRACT NUMBER**

5b. **GRANT NUMBER**

5c. **PROGRAM ELEMENT NUMBER**

5d. **PROJECT NUMBER**

5e. **TASK NUMBER**

5f. **WORK UNIT NUMBER**

6. **AUTHOR(S)**

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:**
   a. REPORT
      unclassified
   b. ABSTRACT
      unclassified
   c. THIS PAGE
      unclassified

17. **LIMITATION OF ABSTRACT**
   Same as Report (SAR)

18. **NUMBER OF PAGES**
   4

19a. **NAME OF RESPONSIBLE PERSON**

---

**Standard Form 298 (Rev. 8-98)**
Prepared by ANSI Bal Z39-18
A change in $f_0$ ($\Delta f_0$) across $\Delta t$ results in an absolute change in frequency of the $k^\text{th}$ pitch harmonic by $k\Delta f_0$. Therefore, in a localized time-frequency region (Figure 1b)

$$\tan \theta = \frac{k\Delta f_0}{\Delta t}.$$  

(4)

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

$$\frac{\partial f_0}{\partial t} \approx \Delta f_0/\Delta t = \left(f_0 \angle \theta\right)/f_{\text{center}}.$$  

(5)

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

$$\phi = \text{angle}[S(\omega = \sin \theta, \Omega = \cos \theta)].$$  

(6)

Our model maps harmonically related speech content in each $a[n,m]$ to concentrated entities in the STFTM 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

$$x[n,m] = \sum_{i=1}^{N} a_i[n,m]$$  

(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 $\theta$ in (7). 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.

\[ X(\alpha, \Omega) = \sum_{i=1}^{N} a_i[n,m] \sum_{i=1}^{N} \frac{\alpha_{ij} \cdot \alpha_{ij}}{\alpha_{ij}} + 0.5 \sum_{i=1}^{N} \left( \alpha_{ij} \cdot \alpha_{ij} \cdot \alpha_{ij} \cdot \alpha_{ij} \right) e^{-im} \]

(8)

For slowly-varying $A(\alpha, \Omega)$, the contribution to $X(\alpha, \Omega)$ from multiple speakers exhibits overlap near the GCT origin (Figure 2b); however, as in the single-speaker case, $A(\alpha, \Omega)$ 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 STFT (\(s_{n,m}\)), localized regions centered at \(k\) and \(l\) in time and frequency (\(s_{k,l}\)) of size 625 Hz by 100 ms are extracted using a 2-D Hamming window (\(w_k[m,l]\)) for GCT analysis. We then apply a high-pass filter \(h_{\alpha}[n,m]\) to each \(s_{k,l}\) to remove \(\alpha, A(\omega, \Omega)\) in (2); we denote this result as \(s_{k,l,\alpha}[n,m]\). \(h_{\alpha}[n,m]\) is a circular filter with cut-offs at \(\omega = \Omega = 0.1\pi\), corresponding to \(\omega\) to a ~300 Hz upper limit of \(f_0\) values observed in analysis.

For each \(s_{k,l,\alpha}[n,m]\), we aim to approximately recover \(\alpha, A(\omega, \Omega)\) using 2-D sinusoidal demodulation. The carrier (\(\cos(\Phi[n,m])\)) parameters are determined from the speaker’s pitch track using (3) for \(\omega\) and (6) for \(\varphi\). To determine \(\theta\), a linear least-squared error fit is applied to the pitch values spanning the 100-ms duration of \(s_{k,l,\alpha}[n,m]\). The slope of this fit approximates \(\partial s_{\alpha}[n,m]\) such that \(\theta\) is estimated using (5). \(s_{k,l,\alpha}[n,m]\) is multiplied by the carrier generated from these parameters followed by filtering with a circular low-pass filter \(h_{\alpha}[n,m]\) 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 \(s_{k,l}\)

\[
\hat{s}_{k,l}[n,m] = \left(\alpha_{\alpha} + \cos(\Phi[n,m])\right)\hat{a}[n,m].
\]

(9)

For each time-frequency unit of \(s_{k,l}[n,m]\), (9) corresponds to a linear equation in \(\alpha_{\alpha}\) since the values of \(s_{k,l}[n,m]\), \(\hat{a}[n,m]\), 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 \(s_{k,l}[n,m]\) using the estimated \(\alpha_{\alpha}\), \(\hat{a}[n,m]\), and \(\cos(\Phi[n,m])\) is denoted as \(\hat{s}_{k,l}[n,m]\). The full STFT estimate \(\hat{s}_f[n,m]\) is obtained using overlap-add (OLA) with a LSE criterion (OLA-LSE) [11]

\[
\hat{s}_f[n,m] = \sum_{N} \sum_{m} w_s[kT - n,IF - m] \hat{s}_{k,l}[n,m].
\]

(10)

OLA step sizes in time and frequency (\(T\) and \(F\)) are set to \(1/4\) the size of \(w_s[n,m]\). \(\hat{s}_f[n,m]\) 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 \(s_{k,l}[n,m]\) be a localized region of the full STFT computed for the mixture signal centered at \(k\) and \(l\) in time and frequency. \(x_{k,l}[n,m]\) is filtered with \(h_{\alpha}[n,m]\) to remove the overlapping \(\alpha_{\alpha}, A(\omega, \Omega)\) terms at the GCT origin (Figure 2b); we denote this result as \(x_{k,l,\alpha}[n,m]\). A cosine carrier for each speaker is generated using the corresponding pitch track and multiplied by \(x_{k,l,\alpha}[n,m]\) to obtain

\[
x_{k,l}[n,m] = x_{k,l,\alpha}[n,m]\cos(\omega[n\sin \theta + m\cos \theta] + \varphi)
\]

\[
= \hat{a}[n,m] + c[n,m].
\]

(11)

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

\[
x_{k,l}[n,m] = \sum_{i} (\alpha_{\alpha,i} + \cos(\Phi[n,m]))\hat{a}_i[n,m]
\]

(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 \(\alpha_{\alpha,i}\) 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. PRELIMINARY 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 \(s_{k,l}[n,m]\) with an adaptive filter

\[
h[n,m] = h_{\alpha}[n,m](1 + 2\cos(\omega[n\sin \theta + m\cos \theta] + \varphi))
\]

(13)

where \(\omega, \theta, \varphi\) are determined for each localized time-frequency region using the speaker's pitch track and \(h_{\alpha}[n,m]\) 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) 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>Exp1 SNR (dB)</th>
<th>Exp1 Demod.</th>
<th>Exp2 SNR</th>
<th>Exp2 Demod.</th>
<th>Exp2 TruePhase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtering</td>
<td>11.24</td>
<td>12.51</td>
<td>3.62</td>
<td>4.09</td>
<td>5.96</td>
</tr>
</tbody>
</table>

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 ([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