Towards co-channel speaker separation BY 2-D demodulation of spectrograms

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/ASPAA.2009.5346526">http://dx.doi.org/10.1109/ASPAA.2009.5346526</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sat Dec 08 01:18:34 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/71798">http://hdl.handle.net/1721.1/71798</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
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 in concentrated areas to a transformed 2-D space, thereby motivating 2-D 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 mixture spectrograms 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.

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

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

\[
\Phi[n,m] = \alpha_s(\cos\theta + m\sin\theta) + \phi.
\] (1)

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

\[
S(\omega, \Omega) = \alpha_s A(\omega, \Omega) + 0.5e^{-i\phi}A(\omega + \alpha_s \sin\theta, \Omega - \omega_s \cos\theta) + 0.5e^{i\phi}A(\omega - \alpha_s \sin\theta, \Omega + \omega_s \cos\theta)
\] (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 \( N_{\text{STFT}} \) as the discrete-Fourier transform (DFT) length of the STFT, the GCT parameters relate to the speaker’s pitch \( f_s \) at the center (in time) of \( s[n,m] \) (Figure 1b, c) [5, 10]:

\[
f_s = (2\pi f_s)/(N_{\text{STFT}}\alpha_s \cos\theta).
\] (3)
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)

\[
\tan \theta = \left( k\Delta f_0 \right)/\Delta t .
\]  

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) \([10]\). 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.

For a particular \( s[n,m] \) 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 \( s[n,m] \) is then

\[
\partial f_0/\partial t \approx \Delta f_0/\Delta t = (f_0 \tan \theta)/f_{\text{center}} .
\]  

Finally, \( \varphi \) corresponds to the position of the sinusoid in \( s[n,m] \); for a non-negative DC value of \( a[n,m] \), \( \varphi \) can be obtained by analyzing the GCT at \( (\omega = \omega_0 \sin \theta, \Omega = \omega_0 \cos \theta) \)

\[
\varphi = \text{angle}[S(\omega_0 \sin \theta, \omega_0 \cos \theta)].
\]  

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

\[
x[n,m] = \sum_{i=1}^{N} \alpha_i a_i[n,m] + \sum_{i=1}^{N} a_i[n,m] \cos \left( \frac{n \cos \theta + m \sin \theta + \varphi}{\Delta t} \right).
\]  

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

For each $s_{ul}[m,n]$, we aim to approximately recover $\alpha_{k,l}A(\omega,\Omega)$ using 2-D sinusoidal demodulation. The carrier (cos($\Phi[m,n]$)) parameters are determined from the speaker’s pitch track using (3) for $\varphi$ 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 $s_{ul}[m,n]$. The slope of this fit approximates $\partial f/s\partial t$ such that $\theta$ is estimated using (5). $s_{ul}[m,n]$ is multiplied by the carrier generated from these parameters followed by filtering with a circular low-pass filter $h_{up}[m,n]$ with cut-offs at $\omega=\Omega=0.1\pi$; we denote this result as $\hat{a}[m,n]$. $\hat{a}[m,n]$ is combined with the carrier using (1) and set equal to $s_{kl}[m,n]$

$$s_{kl}[m,n] = (\alpha_{k,l} + \cos(\Phi[m,n]))\hat{a}[m,n].$$  \hspace{1cm} (9)

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

$$\hat{s}_{f}[m,n] = \sum_{k,l} w_{km}[kT-n,LF-m]|s_{ul}[m,n]|.$$  \hspace{1cm} (10)

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

$$x_{kl}[m,n] = x_{ul}[m,n]\cos(\varphi[n\sin\theta + m\cos\theta] + \varphi)$$  \hspace{1cm} (11)

where $\varphi$, $\theta$, and $\varphi$ are determined for each localized time-frequency region using the speaker's pitch track and $h_{up}[m,n]$ 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 and Roy” spoken by 10 males and females 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.

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 $s_{f}[m,n]$ with an adaptive filter

$$h_{up}[m,n] = h_{up}[m,n](1 + 2\cos(\varphi[n\sin\theta + m\cos\theta] + \varphi)).$$  \hspace{1cm} (13)
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](image1)

**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](image2)

**Figure 4.** (a) STFT magnitude of mixture (Nanny + Roy); (b) Recovered STFT magnitude 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></th>
<th>Exp1 Filtering</th>
<th>Exp1 Demod.</th>
<th>Exp2 SBSS</th>
<th>Exp2 Demod.</th>
<th>Exp2 TruePhase</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</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 tones 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