Harmonicity Plays a Critical Role in DNN Based Versus in Biologically-Inspired Monaural Speech Segregation Systems

Rahil Parikh, Ilya Kavalerov, Carol Espy-Wilson, Shihab Shamma
University of Maryland College Park, MD, USA
Google Inc, CA, USA
Introduction
Monaural Speech Segregation Systems

Cocktail Party Problem → Computational Auditory Scene Analysis (CASA) → Speech Segregation

<table>
<thead>
<tr>
<th>Traditional CASA Systems</th>
<th>Deep Neural Network Based Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Harmonicity Model:</strong></td>
<td><strong>Harmonicity</strong></td>
</tr>
<tr>
<td>• Continuity of pitch in time</td>
<td>?</td>
</tr>
<tr>
<td>• Harmonic structure across frequency</td>
<td>?</td>
</tr>
<tr>
<td><strong>Temporal Coherence Model</strong></td>
<td><strong>Temporal Coherence</strong></td>
</tr>
<tr>
<td>• Biologically plausible</td>
<td>?</td>
</tr>
<tr>
<td>• Features of a single source are modulated</td>
<td>?</td>
</tr>
<tr>
<td>• Onset co-incidence and timing cues</td>
<td>?</td>
</tr>
</tbody>
</table>

Goal: Bridge the gap between CASA systems and Deep Neural Network based speech segregation models
Monaural Speech Segregation Systems

Cocktail Party Problem → Computational Auditory Scene Analysis (CASA) → Speech Segregation

<table>
<thead>
<tr>
<th>Traditional CASA Systems</th>
<th>Deep Neural Network Based Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Harmonicity Model:</strong></td>
<td><strong>Harmonicity</strong></td>
</tr>
<tr>
<td>• Continuity of pitch in time</td>
<td>✓</td>
</tr>
<tr>
<td>• Harmonic structure across frequency</td>
<td></td>
</tr>
<tr>
<td><strong>Temporal Coherence Model:</strong></td>
<td></td>
</tr>
<tr>
<td>• Biologically plausible</td>
<td></td>
</tr>
<tr>
<td>• Features of a single source are modulated</td>
<td></td>
</tr>
<tr>
<td>• Onset co-incidence and timing cues</td>
<td></td>
</tr>
</tbody>
</table>

Goal: Bridge the gap between CASA systems and Deep Neural Network based speech segregation models
Investigate the importance of harmonicity for DNN-based speech segregation models
Inharmonic Sources

• Inharmonic sounds: components not at integer multiples $F_0$

$$f_n(t) = n f_0(t) + J_n f_0(t); \quad -J < J_n < J$$ (1)

• Inharmonic Tones:

$$x_{tone} = \sum_{k=1}^{N} A_k \sin(2\pi f_n(t)t)$$ (2)

• Inharmonic sources: $J \neq 0$

• Natural speech: $J = 0$

Inharmonic Speech: Modified STRAIGHT Algorithm [Kawahara, 2018]
Experiments
Experiments

Dataset: WSJ0 and WSJ-2-Mix

Generate inharmonic versions of WSJ0 for each jitter: $0.01 < J < 0.30$:

- Average offset for male speakers: $\pm 1.2 - \pm 40$ Hz
- Average offset for female speakers: $\pm 2.1 - \pm 65$ Hz

Evaluate Conv-Tasnet and DPT-Net trained on natural (harmonic) speech mixtures with:

- Mixtures of inharmonic tones
- Inharmonic WSJ-2-Mix (inharmonic speech + inharmonic speech)
- Mixtures of inharmonic and natural WSJ0 (inharmonic speech + natural speech)
- Baseline: Natural WSJ-2-mix (natural speech + natural speech)

Evaluation Metric: Signal-Distortion Ratio (SDR)
Experiments

Dataset: WSJ0 and WSJ-2-Mix

Generate inharmonic versions of WSJ0 for each jitter: 0.01 < J < 0.30:
  • Average offset for male speakers: ±1.2 – ±40 Hz
  • Average offset for female speakers: ±2.1 – ±65 Hz

Evaluate Conv-Tasnet trained on inharmonic speech mixtures with:
  • Inharmonic WSJ-2-Mix (inharmonic speech + inharmonic speech)
  • Natural WSJ-2-mix (natural speech + natural speech)

Evaluation Metric: Signal-Distortion Ratio (SDR)
Results
DNN Models Trained on Natural Speech

Conv-Tasnet **fails** to segregate mixtures of inharmonic tones
DNN Models Trained on Natural Speech

Conv-Tasnet **can** segregate mixtures of natural speech + harmonic tones

Conv-Tasnet **cannot** segregate mixtures of natural speech + inharmonic tones

Both sources need to be harmonic
DNN Models Trained on Natural Speech

Tone 1: 200Hz, 600Hz, Tone 2: 100Hz, 300Hz, 500Hz
At overlap: harmonic series of 100 Hz

Network groups overlapping harmonic region as one source
DNN Models Trained on Natural Speech

- Model Performance drops to $\approx 0$ dB if both speakers are inharmonic
- Model Performance drops to $\approx 8$ dB if one speaker is inharmonic
DNN Models Trained on Inharmonic Speech

- The network finds it challenging to learn to segregate speech
- Model performance on natural speech deteriorates
- Harmonicity is critical for segregation
DNN Models Diverge from Temporal Coherence

- Humans and TC models (Krishnan et al. 2014) group all sources with the same timing onset and offset as one source, regardless of harmonicity

- Conv-Tasnet can segregate two synchronous, harmonic sources
Conclusion
Conclusion and Future Work

• DNNs cue onto the harmonic structure for segregation
• SOTA models completely fail with inharmonic inputs (adversarial inputs)
• DNNs implicitly learn the non-trivial task of pitch-tracking
• DNNs diverge from biologically inspired CASA models

Next Steps:
• Analysis on spectrogram-based DNN networks
• Investigation on how DNN models perform harmonic analysis