Inharmonic Sources

- Sounds with components not at integral multiples of fundamental frequency (F0)
- Inharmonic sound with components maximally jittered by ±1% F0
- Generate inharmonic WSI for different J
- Average spectral offset for male speakers: ±1.2 – ±40 Hz
- Average spectral offset for female speakers: ±12.1 – ±65 Hz

Evaluation Metric: Signal-Distortion Ratio (SDR)

Results: Divergence of DNNs from Temporal Coherence

• Humans and Temporal Coherence models (Krishnan et al. 2014) group all sources with the same timing onset and offset as one source.
• Unlike humans, Conv-Tasnet can segregate two synchronous, harmonic sources.

Results: Segregation Performance on Inharmonic Tones

• Conv-Tasnet can segregate a mixture of harmonic tones.
• It cannot segregate inharmonic tones.

• Conv-Tasnet can segregate a mixture of harmonic overlapping tones.
• A mixture of overlapping tones of 200Hz, 600Hz and 100Hz, 300Hz, 500Hz contain the harmonics of 100Hz during the overlap.
• Conv-Tasnet segregates this overlap as one single source.

• Conv-Tasnet can segregate a mixture of natural speech and harmonic tones.
• It cannot segregate mixtures of natural speech and inharmonic tones.

Introductions

• Traditional CASA algorithms: designed using established underlying principles
  - E.g., Temporal Coherence (Krishnan et al., 2014) models use timing cues  biologically inspired
  - E.g., Harmony and continuity in pitch (Vishnubhotla et al. 2009)
• Deep Neural Networks (DNN) models outperform CASA models but are black-boxes.
• Goal: Investigate the underlying principles of DNN based speech segregation models

Empirical Analysis

• Evaluate Conv-Tasnet [Luo et al., 2019] and DPT-Net [Chen et al. 2020] trained on natural speech with:
  - Mixtures of inharmonic tones
  - Mixtures of inharmonic speech (inharmonic speech + inharmonic speech)
  - Mixtures of natural and inharmonic speech (inharmonic speech + harmonic speech)
  - Baseline: Mixtures of natural speech (harmonic speech + harmonic speech)
• Train Conv-Tasnet and DPT-Net on inharmonic speech mixtures and evaluate with:
  - Mixtures of inharmonic speech
  - Mixtures of inharmonic speech

Conclusion

• Unlike Temporal Coherence models, DNNs do not rely on timing information.
• DNNs cue onto harmonicity for segregation.
• SOTA models completely fail with inharmonic inputs.
• DNNs implicitly perform pitch-tracking.
• DNNs find it challenging to learn from inharmonic speech.
• Inharmonic speech  adversarial input to DNN based models.

Takeaways

• Unlike Temporal Coherence models, DNNs do not rely on timing information.
• DNNs cue onto harmonicity for segregation.
• SOTA models completely fail with inharmonic inputs.
• DNNs implicitly perform pitch-tracking.
• DNNs find it challenging to learn from inharmonic speech.
• Inharmonic speech  adversarial input to DNN based models.

Next Steps

• Investigate how DNNs perform harmonic analysis.
• Investigate how DNNs perform harmonics tracking.
• Study spectrogram-based speech segregation models.

Acknowledgements

• This work was supported by NSF grant #1764010 and an AFOSR grant.
• The authors declare no conflict of interests.