HARMONICITY PLAYS A CRITICAL ROLE IN DNN BASED VERSUS IN BIOLOGICALLY-INSPIRED MONAURAL SPEECH SEGREGATION SYSTEMS

Rahil Parikh[1], Ilya Kavalerov[2], Carol Espy-Wilson[1], Shihab Shamma[1]

[1] Institute for Systems Research, University of Maryland, College Park, MD, USA
[2] Google Inc, Mountain View, CA, USA

ABSTRACT

Recent advancements in deep learning have led to drastic improvements in speech segregation models. Despite their success and growing applicability, few efforts have been made to analyze the underlying principles that these networks learn to perform segregation. Here we analyze the role of harmonicity on two state-of-the-art Deep Neural Networks (DNN)-based models- Conv-TasNet and DPT-Net [1, 2]. We evaluate their performance with mixtures of natural speech versus slightly manipulated inharmonic speech, where harmonics are slightly frequency jittered. We find that performance deteriorates significantly if one source is even slightly harmonically jittered, e.g., an imperceptible 3% harmonic jitter degrades performance of Conv-TasNet from 15.4 dB to 0.70 dB. Training the model on inharmonic speech does not remedy this sensitivity, instead resulting in worse performance on natural speech mixtures, making inharmonicity a powerful adversarial factor in DNN models. Furthermore, additional analyses reveal that DNN algorithms deviate markedly from biologically inspired algorithms [3] that rely primarily on timing cues and not harmonicity to segregate speech.

Index Terms— speech segregation, Conv-TasNet, harmonic structure, temporal coherence, adversarial inputs

1. INTRODUCTION

The 'cocktail party problem' is the task of attending to a source of interest, usually speech, in a complex acoustic environment with concurrent sounds. Despite the apparent ease with which humans can group acoustic cues from such an environment and organize them to meaningfully perceive them [4], the complexity of this problem has inspired generations of neuroscientists, psychologists and engineers. The multi-disciplinary nature of this problem has motivated Computational Auditory Scene Analysis (CASA) algorithms using traditional pitch modulations extractions [5–7] and biologically-inspired systems which are based on frameworks such as 'temporal coherence' (TC) [3, 8]. In contrast to these approaches, there are strictly engineering frameworks such as signal factorization, statistical-speech modelling [9, 10] and recently deep learning [1, 2, 11–16] which ignore biological plausibility. Advances in deep learning have facilitated drastic improvements in speech segregation [1, 2, 12–16].

These models can now be used at scale, to separate individual speech from mixtures, enabling downstream audio processing tasks such as Automatic Speech Recognition (ASR), allowing us to use speech technologies in the *real-world*. These model architectures have also shown promising results in music segregation [17] and sound segregation [18]. Despite their unquestionable success, no work has been done to investigate the underlying principles they learn to perform grouping and consequently segregation.

We attempt to bridge the gap between our understanding of traditional CASA systems, biologically inspired algorithms and deep-learning systems by investigating the role of harmonicity on the latter. Since harmonicity of natural speech is linked with pitch extraction, it is a crucial cue for grouping in certain CASA systems [5, 7]. However, biologically plausible algorithms based on TC depend solely on the temporal correlation of features from the same source and are capable of performing segregation regardless of harmonicity [3]. While recent findings in [19] found that human subjects perform worse in segregating inharmonic acoustic mixtures, performance deficits were relatively modest and often attributable to source assignment confusion rather than true segregation difficulty of the overlapping words in the mixtures. In our work, we begin by exploring the importance of harmonicity on two prominent speech segregation models Conv-TasNet [1] and Dual-Path Transformers (DTP) Net [2]. We investigate how these models perform on mixtures of harmonic speech, inharmonic speech, and a variety of synthetic simple stimuli that exemplify these properties more clearly. We also extend this analysis to the same models when trained on inharmonic speech. We find that-

- DNN based segregation systems trained on natural speech fail to segregate inharmonic speech. The performance of DPT-Net degrades from 20.2 dB to 0.7 dB for imperceptible amounts of harmonic jitter.

- DNNs find it hard to learn to segregate sources when trained on inharmonic speech, instead becoming worse at segregating natural speech.

- They differ from biologically inspired systems in heavily relying on harmonic patterns to segregate speech.

2. BACKGROUND

2.1. Monaural Speech Segregation Models

Monaural speech segregation involves estimating C sources $s_1(t), \cdots, s_C(t)$ for an audio mixture $x_{mix}(t) = \sum_{i=1}^{C} s_i$. In this work we focus on end-to-end speech segregation models [14] which share a similar pipeline– the input mixture is encoded using a learnt analysis transform. The segregation network generates C masks for this encoded input. The masked input is then decoded using a learnt synthesis transform to generate a time-domain estimate for each source. These networks typically use Permutation Invariant Training (PIT) [16] to maximize the Scale Invariant Source-to-Noise Ratio (SI-SNR) [14]. In Conv-TasNet the segregation network consists of dilated 1-D convolutional networks, whereas DPTNet consists of a cascade of transformers configured to maximize the temporal receptive field over the encoded mixture. In this work we set $C = 2$ for simplicity and illustrative purposes.

2.2. Inharmonic Sound Sources

Sounds that originate from a periodic process have a harmonic structure in their Fourier representation, consisting usually of a component at the fundamental frequency (F0) and its integer multiples. Inharmonic sounds instead do not have frequency components at exact integer multiples of F0. We use the modified STRAIGHT algorithm [20] to generate inharmonic speech. This algorithm models voiced excitations as a sum of multiple time varying sinusoids, the frequencies of which are manipulated using 'jittering', wherein a random offset is added to each harmonic. This offset is bounded by some fraction- J of F0. The nth harmonic is thus modified to:

$$f_n(t) = nf_0(t) + J_n f_0(t); \quad -J \leq J_n \leq J \qquad (1)$$

Since other characteristics of the speech remain unchanged, synthetically generated inharmonic speech is very intelligible[1]. To generate inharmonic tone complexes, we sample from $x_{tone}(t) = \sum_{k=1}^{N} A_k sin(2\pi f_n(t)t)$ where $f_n(t)$ is given by (1). Sources are harmonic at $J_n = 0$.

3. EXPERIMENTS

To demonstrate the importance of harmonicity as a cue for grouping, we first evaluate performance of Conv-TasNet (trained on native (harmonic) WSJ-2-mix) with mixtures of harmonic or inharmonic complex tones. To extend these results to speech, we evaluate Conv-TasNet and DPT-Net with inharmonic versions of WSJ-2-mix, with varying levels of inharmonicity. We also evaluate these models with mixtures where one speaker is natural and the other is inharmonic. Finally, we train a Conv-TasNet on only inharmonic mixtures.

---

[1] http://labrosa.ee.columbia.edu/projects/inharmonic/

### 3.1. Dataset Construction and Evaluation

We use the native WSJ-2-mix dataset [12] created using the WSJ0 corpus to train our models on harmonic speech. We create multiple inharmonic versions of the WSJ0-corpus, by varying the harmonic jitter J in (1) and create their corresponding inharmonic WSJ-2-mix datasets. We change J from $0.01 - 0.30$, corresponding to an average offset in the range of $\pm 1.2 - \pm 40$ Hz for male speakers and $\pm 2.1 - \pm 65$ Hz for female speakers. A higher value of J corresponds to a higher deviation from the native source. We also generate versions of the WSJ-2-mix datasets wherein one source is harmonic and the other is inharmonic.

We evaluate source-segregation performance using the metric signal-to-distortion-ratio improvement (SDRi) [21]. The results are baselined against the model's performance on the harmonic WSJ-2-mix dataset with $J = 0$. Since each of the speech datasets contain differently jittered versions of the same samples from WSJ-2-mix, we can fairly compare the evaluation performance across these datasets.

3.2. Results

3.2.1. Speech Segregation Models Trained on Natural Speech

We begin by demonstrating that Conv-TasNet trained on the native WSJ-2-mix relies primarily on harmonicty to perform separation. The test stimuli are mixtures of alternating tone complexes (with no overlap) as illustrated in Fig 1. The model
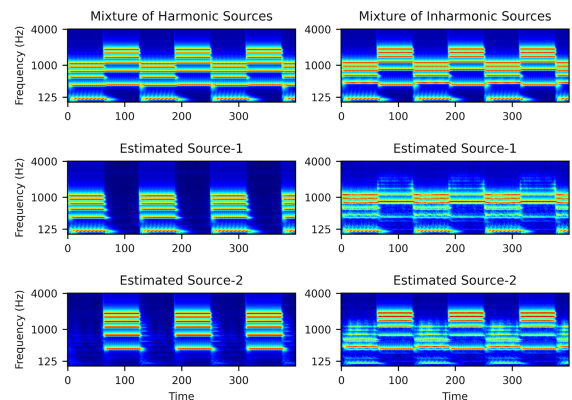

Fig. 1: Conv-Tasnet Segregates mixtures of harmonic tone complexes (left panels), but fails with inharmonic complexes (right panels).

performs well when both complexes are harmonic ($J = 0$ and $F0 = 110, 210$ Hz) as illustrated in the left panels of Fig. 1. It, however, fails completely when the complexes are harmonically-jittered (right panels) with $J = 0.20$.

This phenomenon applies equally when the pair of stimuli are very different as in Fig.2, where one is a tone complex and the other is speech. The segregation is perfectly effective

when both are harmonic (left panels) while it fails when one of the sources (the tone complex) is inharmonic (right panels) in spite of the speech being harmonic.
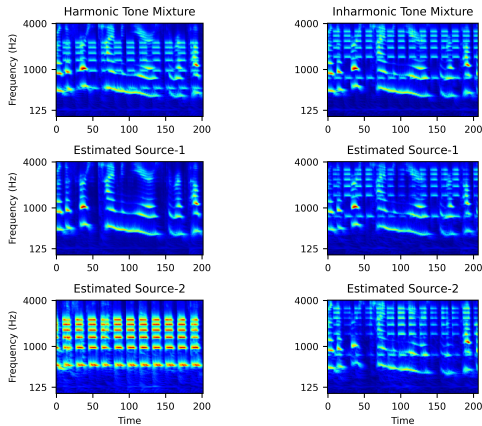

Fig. 2: Mixtures of natural speech with harmonic (left panels) and inharmonic (right panels) tone complexes are segregated only when both are harmonic.

Harmonicity may also be a transient phenomenon during speech mixtures due to the overlap of harmonic components. To illustrate the immediate sensitivity of the DNN models to the emergence (or absence) of harmonicity, we illustrate in Fig.3 two alternating mixtures with 50% overlap. In the left panels, both tone complexes are completely harmonic (F0=100 and 190). On the right, the complexes are inharmonic with frequencies $200, 600$ Hz and $100, 300, 500$ Hz. During the intervals of overlap, the effective instantaneous spectrum becomes the sum of the two complexes. On the left, the harmonic patterns still exist in the mixture, and hence are extracted out and segregated perfectly. But in the right panels, the tone mixture strongly resemble the harmonics of F0=100 Hz. The DNN model incorrectly groups this harmonic series of tones as one source, while the two individually inharmonic series become the other source. This demonstrates the strong dependence of the model on the harmonicity of the spectrum.

These tonal stimuli illustrate the vulnerabilities that affect models when harmonicity is disturbed. To extend our analysis to speech mixtures we evaluate the models using mixtures of: two inharmonic speakers (I+I), and one harmonic and one inharmonic speaker (H+I), using datasets discussed in Sec. 3.1. The results in Fig.4(a) illustrate that the models' performance degrades severely, especially on mixtures of two inharmonic speakers– 15 dB on natural speech to 0.7 dB for ConvTasnet and 20 dB to 0.7 dB for DPT-Net, for $J = 0.03$, which reveals a remarkable vulnerability.

3.2.2. *Training DNN-Models with Inharmonic Speech*

An important question is whether these DNN models can be made more robust to inharmonicity by training them on inharmonic speech. Here, we trained the Conv-TasNet on the WSJ-
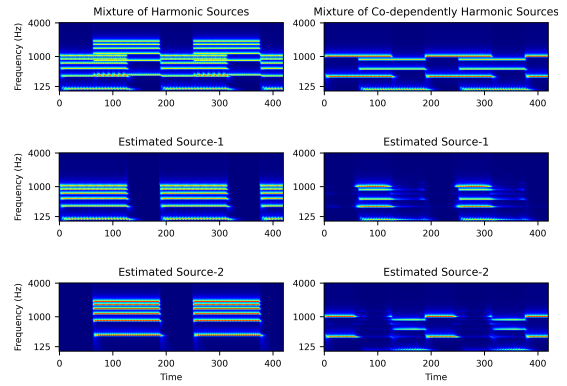
Fig. 3: Segregation of overlapped spectra. (Left panels) A mixture of 2 harmonic sources segregate well. (Right panels) Alternating inharmonic spectra become harmonic and incorrectly segregate during overlap.
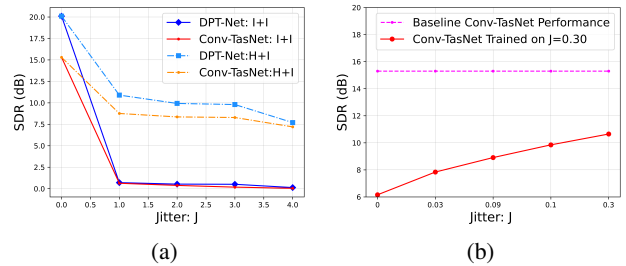
Fig. 4: Performance on mixtures of inharmonic speakers with increasing jitter J for models trained on (a) native WSJ-2-mix and (b) inharmonic WSJ-2-mix with $J = 0.3$

2-mix datasets with both inharmonic speakers. We note that the model finds it more challenging to learn to segregate with increasing inharmonicty. Our results from a model trained only on inharmonic data with $J = 30\%$ are illustrated in Fig. 4(b). In spite of being trained on data from the same distribution, the performance of the network drops from 15.3 dB to 10.6 dB. Interestingly, this network now performs poorly on harmonic and mildly inharmonic speech. Furthermore, when the model is trained on a combination of natural and inharmonic mixtures, it still finds it hard to segregate inharmonic mixtures but learns to segregate natural speech. We believe that learning to segregate inharmonic mixtures is a harder problem since the space of inharmonic patterns is very large, while that of harmonic patterns is sparse. This makes it difficult to cover all possibilities during training. These results underscore that the networks rely strongly on the constrained harmonic structure for segregation.

3.3. Discussion

Our experiments indicate that end-to-end models trained on natural speech are incapable of segregating an inharmonic source. Furthermore, we have shown that the models can-

not even be trained to segregate inharmonic mixtures.

These results provide clear evidence that DNN-based speech segregation is heavily reliant on learning the harmonic patterns that are available in natural speech, but that are absent or rarely available in inharmonic training. Their sensitivity to harmonicity implies the ability to track the F0, or pitch. But they seem to do so from the full spectral set of harmonic patterns and not just from a few harmonics as humans are readily able to do. For instance, in Fig.3, the network could not treat the sparser pattern of tones as part of a larger harmonic pattern, expecting instead to see the full patterns. To further test this, we generate a mixture of alternating harmonic tone complexes with and without the fundamental component and the second harmonic, as illustrated in Fig. 5. Since the model fails to segregate the tones, it is likely it has learned to recognize only the full harmonic patterns. They then proceed to *track* the trajectory of the pitches implied by these harmonics the varied mixtures of speakers and tones, and despite their often proximate values, patterns, and even inter-crossings and gaps and other well-known difficulties. These experiments provide evidence of the important cues involved in the remarkable performance of the DNNs, but they also demonstrate their vulnerabilities.

In additional experiments not described here for the sake of brevity, we also observed that the models have difficulties in segregating fricatives and unvoiced segments of speech. Fricatives are plentiful in the training set but are not harmonic in structure, indicating that these vulnerabilities are not a result of testing on 'out-of-distribution' data, but instead stem from the underlying mechanisms used by the DNNs to perform the segregation. Finally, we have also confirmed that these same vulnerabilities are not limited to time domain based models trained in an end-to-end fashion but also extend to spectrogram based segregation models.
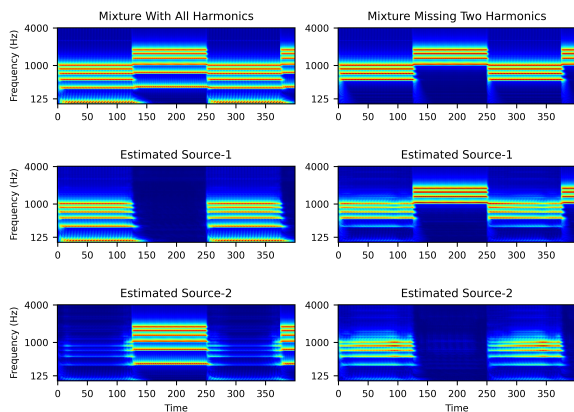
Fig. 5: Conv-Tasnet successfully segregates mixtures of harmonic sources (left panels). However, it fails when the fundamentals and some lower harmonics are absent (right panels).
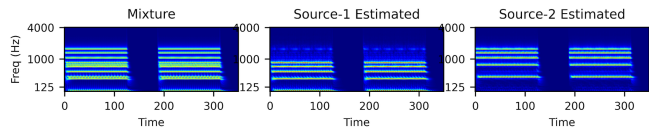
Fig. 6: Conv-TasNet segregates two harmonic sources even when synchronous.

3.3.1. Divergence of DNN Models from Temporal Coherence

Our results also indicate another significant divergence of the DNN-based segregation models from biologically plausible and perceptual models that perform speech segregation using TC. While TC partially leverages pitch information and consequently harmonicity, they primarily rely on the coincidence of the onsets and timing information to group the features emanating from a single source. This divergence between DNNs and such algorithms is evident in Fig. 6, which illustrates that Conv-TasNet can generally segregate two harmonic complexes that have the same exact onset and offset time. Humans and TC algorithms fail to do so because they would group all these components as coming form a single source. Conversely, unlike Conv-TasNet, TC algorithms would have been able to separate the tones in Fig. 3.

3.3.2. Inharmonic Sources as Adversarial Inputs

Since inharmonic speech can be made to be fairly similar sounding to natural speech, we believe that it can be used as an adversarial input to speech segregation models. Given that these networks perform poorly if an inharmonic speech or tone source is present in the mixture, an adversary can insert such sources, hindering source-segregation and thus all downstream-tasks. Since the presence of an inharmonic source is independent of the model architecture, we believe that adversarial nature of such sources is not limited to Conv-TasNet and DPT Net.

4. CONCLUSION

We have shown that DNN-based speech segregation models such as Conv-TasNet and DPT-Net rely on the harmonicity of the sources in the mixture to separate them. They, therefore, fail to segregate sources when their harmonic structure is disturbed, and also find it harder to learn to segregate mixtures with such inharmonic sources. This study demonstrates that these networks implicitly learn to *estimate* pitch by learning the harmonic patterns abundant in the training corpus. They then proceed to track these harmonic complexes to segregate speech. This effort to understand the underlying principles that these networks learn helps us narrow the gap between our understanding of auditory perception and end-to-end learning-based segregation models.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[3] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma, "Segregating complex sound sources through temporal coherence," *PLoS computational biology*, vol. 10, no. 12, pp. e1003985, 2014.

[4] Albert S Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[5] Srikanth Vishnubhotla and Carol Y Espy-Wilson, "An algorithm for speech segregation of co-channel speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, p. 109.

[6] Michael Stark, Michael Wohlmayr, and Franz Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2010.

[7] DeLiang L Wang and Guy J Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE transactions on neural networks*, vol. 10, no. 3, pp. 684–697, 1999.

[8] Mounya Elhilali and Shihab A Shamma, "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, 2008.

[9] Ron J Weiss and Daniel PW Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, vol. 24, no. 1, pp. 16–29, 2010.

[10] Martin Cooke, John R Hershey, and Steven J Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[11] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[12] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[13] Xiao-Lei Zhang and DeLiang Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 5, pp. 967–977, 2016.

[14] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[15] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[16] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[17] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[18] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 175–179.

[19] Sara Popham, Dana Boebinger, Dan PW Ellis, Hideki Kawahara, and Josh H McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[20] Daniel PW Ellis, Josh H McDermott, and Hideki Kawahara, "Inharmonic speech: a tool for the study of speech perception and separation," 2012.

[21] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.