Project Proposal: Monaural Single-Source Speaker Separation

E511 Machine Language for Signal Processing

Jonathan Branam
Indiana University
Bloomington, IN
jobranam@iu.edu

Jamie Israel
Indiana University
Bloomington, IN
jgisrael@iu.edu

Hellen Jiang
Indiana University
Bloomington, IN
heljiang@iu.edu

ABSTRACT

There have been several recent developments in the use of deep learning to perform speaker diarisation, or separation of multiple speech signals from a single source into separate homogenous signals representing each individual speaker. This task is particularly challenging where the input is monophonic, there is no control over environmental conditions, the number of speakers is unknown and there are no samples of the speakers from which to train. One such real-world application is for a call center seeking to analyze the speech of agents and customers where the only source is a monophonic recording of a call with unknown participants. The center would like to transcribe speech to text for content analysis, but they have been unable to obtain an accurate translation due in large part to overlap in the recorded signals (i.e., cross-talk) and the inability to associate utterances with speakers.

Our initial step will be to apply the non-negative matrix factorization technique illustrated in the class materials to perform a baseline for this source separation problem. From there, we intend to explore a neural network model using PyTorch to implement Permutation Invariant Training (PIT) from [4] or utterance-level Permutation Invariant Training (uPIT) from [5] as these seem to be the latest approaches to single source speaker separation in the academic literature. The call center recordings are not currently labeled, so we propose creating our own test and training dataset by mixing single-speaker recordings from either the TIMIT [2] or WSJ0 [1] recordings. The literature mentions the WSJ0-2mix data set that was used in [3] the seminal Deep Clustering (DPCL) paper but we have yet to find a detailed specification for how to generate this dataset.

1. REFERENCES

[1] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett. Csr-i (wsj0) complete ldc93s6a. https://catalog.ldc.upenn.edu/ldc93s6a, 1993.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus ldc93s1. https://catalog.ldc.upenn.edu/LDC93S1, 1993.

[3] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. *CoRR*, abs/1508.04306, 2015.

[4] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen. Multi-talker speech separation and tracing with permutation invariant training of deep recurrent neural networks. *CoRR*, abs/1703.06284, 2017.

[5] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *CoRR*, abs/1607.00325, 2016.