Rheinisch-Westfälische Technische Hochschule Aachen
Lehrstuhl für Informatik 6
Prof. Dr.-Ing. Hermann Ney

Seminar
Selected Topics in Human Language Technology and Pattern Recognition
im WS 2017

ANN Supported Source Separation

Jérôme Lenßen

Matrikelnummer 359 221

Datum des Vortrages

Betreuer: Tobias Menne

Inhaltsverzeichnis

1 Introduction

Present the cocktailparty problem and make analogy to the human sound perception. Briefly mention the history of the problem and importance of a solution for automatic speech recognition (ASR) (e.g. for speech enhancement or applications of ASR like Amazon Echo, etc.).

2 Background

2.1 Representation of mixed speech signal

Mathematical description of the problem and the signal representation as time-frequency bins in the spectogram (e.g. as in [1] or [2, p.5-7]).
Given K sources and signal $x_k(t)$ of k-th source in the time domain we get from our recording a linear mix denoted as $y(t) = \sum_K x_k(t)$.
Mention short time Fourier transform (STFT) and give an overview of important parameters like window function, etc. .
Describe masc M_k for each of our sources defined as $|M_k| = \frac{|X_k(t,f)|}{|Y(t,f)|}$ where $X_k(t,f)$ is the time-frequency bin we get from the spectrogram at (t,f) computed by the STFT from source $k = 1 \ldots K$.

2.2 Performance measurement for speech separation models

Definition of SDR and SNR [SAR, SIR] as in [3]. Quick reminder of WER. [Introduce PESQ].

2.3 Non-neural network based models

2.3.1 Computational auditory scene analysis

Computational auditory scene analysis (CASA) as a rule based approach. Model of human sound perception. Focus on characteristics of human voice such as pitch, harmonics and vocal tract modelling. [4]

2.3.2 Non-negative matrix factorization

Brief overview of non-negative matrix factorization (NMF) and sparse NMF (SNMF) in the context of speech separation as proposed in [5], [6]. Idea: Spectogram Y can be factorized into activation matrix H and basis W for each source, so that $Y = \sum_K W_k H_k$.

2.3.3 Hidden Markov model

Present functions that hidden Markov models (HMM) and factorial HMMs (fHMM) compute and the idea behind using it for speech separation/ enhancement as in [7].
Idea: train one HMM per speaker and combine them to a fHMM later on.

3 Using deep neural networks for speech separation

3.1 Generic neural network architectures

Present most frequently used neural network architectures like RNN and especially (B)LSTM as in [8]. [Reference to presentations ID:01-05 for more details on ReLU, Backprop, etc.].

3.2 Challenges of neural networks in speech separation

3.2.1 Permutation problem

Speaker dependent speech separation faces the permutation problem. Give an example of the proplem: Speaker A, B and C with given mixtures $A+B, A+C$ and $B+C$ will cause confusion if A is assigned to first output in the first two mixtures.

3.2.2 Output dimension problem

Neural networks with fixed output size cannot cope with different numbers of speakers.

3.3 First attempts of DNN in speech separation

Present proposed models in [9] and highlight the CASA background (They still try to use characteristics of the speech signal and don't leave it to the DNN to learn them).

3.3.1 Average energy

Idea: Decoder needs information to differentiate between target and interference. In this case we assume that two signals have different SNR and use two DNNs to recognize the positive/negative SNR. Describe the objective function.

3.3.2 Energy-dependent denoisers

Use idea from above and minimize the mean squared error (MSE).

3.3.3 Average pitch

Assumption: there is signal with average higher/lower pitch like in the average energy based model.

3.3.4 Instantaneous energy

Describe weakness of previous models (Speakers can have same average energy/pitch) and introduce the binary masc used in training such as the training criterion used. Also intro to joint decoding strategy.

3.4 State of the art DNN models

Present the three competing deep learning based models.

3.4.1 Deep Clustering

Describe binary affinity matrix $A = YY^T$, embedding $V = f_\theta$ and objective function (high intrarelationship, low interrelationship). Discuss improvements of the model in follow up paper e.g. the enhancement network, optimizing signal reconstruction and end-to-end training (Present soft K-means). [10], [11].

3.4.2 Permutation Invariant Training

Present idea as illustrated in Fig. 2 in [12, p.5]. Discuss different applicable masks and training criterion.
Explain the solution to the permutation problem and the model used (CNN, ReLU).
Present improvements achieved by uPIT. [1], [12].

3.4.3 Deep Attractor Network

Idea: time-frequency bins are projected to an embedding space and attracted by attractor points. Paper presents how attractor poinst and projection are computed.
Mention analogy to the auditory process in the human brain [not really important but interesting]. Discuss different methods on how to compute attractor points (fixed/anchor attractor points). [13], [14]. Highlight relationship to DPCL as a special case of the attractor network (Proof: use that Y is orthogonal) and PIT [2, p.56]).

4 Comparison of proposed models

4.1 Datasets

Overview of WSJ0-2mix (Training and validation set), WSJ0-3mix based on the Wall Street Journal corpus and the Danish-2mix speaker set from the Danish corpus. Also mention the dataset from the 2006 Monaural Separation and Recognition Challenge.

4.2 Training

Describe the training process on known datasets and compare and discuss different training parameters/methods like:

- learning rate and training termination
- number of epochs
- regularization (e.g. dropout)
- normalization (e.g. gradient normalization)
- data preprocessing
- different network architectures (LSTM , BLSTM and RNN)
- mixed language models

4.3 Results

Compare SDR for different models which used the same dataset also try to compare models which do not share same test data. Performance on mixed language model.

5 Conclusion

Summarize main features and differences of DPCL, PIT and DANet and highlight the advancement in solving the single channel speech separation problem brought by those models.

Literatur

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

[2] Z. Chen, *Single Channel auditory source separation with neural network.* Dissertation, Columbia University, 2017.

[3] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation.,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[4] M. Cooke, *Modelling Auditory Processing and Organisation.* New York, NY, USA: Cambridge University Press, 1993.

[5] M. N. Schmidt and R. K. Olsson, “Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization,” in *Interspeech*, sep 2006.

[6] H. J. L. Roux, Weninger, “Sparse NMF - half-baked or well done?,” Mar. 2015.

[7] T. Virtanen, “Speech recognition using factorial hidden markov models for separation in the feature space.,” 01 2006.

[8] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[9] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, “Deep neural networks for single-channel multi-talker speech recognition,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, pp. 1670–1679, Oct. 2015.

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

[11] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” *CoRR*, vol. abs/1607.02173, 2016.

[12] 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*, vol. abs/1703.06284, 2017.

[13] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” *CoRR*, vol. abs/1611.08930, 2016.

[14] Z. Chen, Y. Luo, and N. Mesgarani, “Speaker-independent Speech Separation with Deep Attractor Network,” *CoRR*, vol. abs/1707.03634, 2017.