Tutorial 1 - T1
Sunday Morning, Sep 2, 9:00AM - 12:30PM

Deep Learning Based Speech Separation

DeLiang Wang - The Ohio State University
Deep Learning Based Speech Separation

Speech separation is the task of separating target speech from background interference. In contrast to the traditional signal processing perspective, speech separation can be formulated as a supervised learning problem, where the discriminative patterns of speech, speakers, and background noise are learned from training data. The recent introduction of deep learning to supervised speech separation has dramatically accelerated progress and boosted separation performance, including the first demonstration of substantial speech intelligibility improvements by hearing impaired listeners in noisy environments. This tutorial provides a comprehensive overview of the research on deep learning based supervised speech separation in the last several years. We systematically introduce three main components of supervised separation: learning machines, training targets, and acoustic features. Much of the tutorial will be on separation algorithms where we describe monaural methods, including speech enhancement (speech-non-speech separation), speaker separation (multi-talker separation), and speech dereverberation, as well as multi-microphone techniques. In addition, we discuss a number of conceptual issues, including what constitutes the target source.

DeLiang Wang received the B.S. degree in 1983 and the M.S. degree in 1986 from Peking (Beijing) University, Beijing, China, and the Ph.D. degree in 1991 from the University of Southern California, Los Angeles, CA, all in computer science. Since 1991, he has been with the Department of Computer Science & Engineering and the Center for Cognitive and Brain Sciences at Ohio State University, Columbus, OH, where he is currently a Professor and University Distinguished Scholar. He also holds a visiting appointment at the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi’an, China. He has been a visiting scholar to Harvard University, Oticon A/S (Denmark), and Starkey Hearing Technologies. Wang’s research interests include machine perception and deep neural networks. Among his recognitions are the Office of Naval Research Young Investigator Award in 1996, the 2005 Outstanding Paper Award from IEEE Transactions on Neural Networks, and the 2008 Helmholtz Award from the International Neural Network Society. He serves as Co-Editor-in-Chief of Neural Networks, and on the editorial boards of several journals. He is an IEEE Fellow.
Deep Learning Based Speech Separation

DeLiang Wang

Perception & Neurodynamics Lab
Ohio State University
& Northwestern Polytechnical University

http://www.cse.ohio-state.edu/pnl/

Outline of presentation

I. Introduction
II. Training targets
III. Features
IV. Separation algorithms
V. Concluding remarks
Real-world audition

What?
- Speech message
- Speaker
- Age, gender, linguistic origin, mood, ...
- Music
- Car passing by

Where?
- Left, right, up, down
- How close?

Channel characteristics

Environment characteristics
- Room reverberation
- Ambient noise

Sources of intrusion and distortion

additive noise from other sound sources
reverberation from surface reflections
channel distortion
Cocktail party problem

- **Term coined by Cherry**
  - “One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it ‘the cocktail party problem’…” (Cherry’57)
  - “For ‘cocktail party’-like situations… when all voices are equally loud, speech remains intelligible for normal-hearing listeners even when there are as many as six interfering talkers” (Bronkhorst & Plomp’92)

- **Ball-room problem by Helmholtz**
  - “Complicated beyond conception” (Helmholtz, 1863)

- **Speech separation problem**
  - Separation and enhancement are used interchangeably when dealing with nonspeech interference

---

Human performance in different interferences

[Graph showing word intelligibility score (%) vs. signal to noise ratio (dB) with different conditions: Complex tones, Broadband noise, 1 Voice, 2 Voices, 8 Voices.]

SRT 23 dB difference in Speech Reception Threshold!

Some applications of speech separation

- Robust automatic speech and speaker recognition
- Noise reduction for hearing prosthesis
  - Hearing aids
  - Cochlear implants
- Noise reduction for mobile communication
- Audio information retrieval

Traditional approaches to speech separation

- Speech enhancement
  - Monaural methods by analyzing general statistics of speech and noise
  - Require a noise estimate
- Spatial filtering with a microphone array
  - Beamforming
    - Extract target sound from a specific spatial direction with a sensor array
  - Independent component analysis
    - Find a demixing matrix from multiple mixtures of sound sources
- Computational auditory scene analysis (CASA)
  - Based on auditory scene analysis principles
  - Feature-based (e.g. pitch) versus model-based (e.g. speaker model)
Supervised approach to speech separation

• **Data driven, i.e. dependency on a training set**

• **Born out of CASA**
  • Time-frequency masking concept has led to the formulation of speech separation as a supervised learning problem

• **A recent trend fueled by the success of deep learning**

• **Focus of this tutorial**

---

Ideal binary mask as a separation goal

• **Motivated by the auditory masking phenomenon and auditory scene analysis, we suggested the ideal binary mask as a main goal of CASA (Hu & Wang’04)**

• **The idea is to retain parts of a mixture where the target sound is stronger than the acoustic background, and discard the rest**

• **The definition of the ideal binary mask (IBM)**

\[
IBM(t, f) = \begin{cases} 
1 & \text{if } SNR(t, f) \geq \theta \\
0 & \text{otherwise}
\end{cases}
\]

• **\(\theta\): A local SNR criterion (LC) in dB, which is typically chosen to be 0 dB**

• **Optimal SNR:** Under certain conditions the IBM with \(\theta = 0\) dB is the optimal binary mask in terms of SNR gain (Li & Wang’09)

• **Maximal articulation index (AI) in a simplified version (Loizou & Kim’11)**

• **It does not actually separate the mixture!**
IBM illustration

Subject tests of ideal binary masking

- IBM separation leads to large speech intelligibility improvements
  - Improvement for stationary noise is above 7 dB for normal-hearing (NIH) listeners (Brungart et al.’06; Li & Loizou’08; Cao et al.’11; Ahmadi et al.’13), and above 9 dB for hearing-impaired (HI) listeners (Anzalone et al.’06; Wang et al.’09)
  - Improvement for modulated noise is significantly larger than for stationary noise
- With the IBM as the goal, the speech separation problem becomes a binary classification problem
  - This new formulation opens the problem to a variety of pattern classification methods
Speech perception of noise with binary gains

- Wang et al. (2008) found that, when LC is chosen to be the same as the input SNR, nearly perfect intelligibility is obtained when input SNR is \(-\infty\) dB (i.e. the mixture contains noise only with no target speech)
- IBM modulated noise for ???

![Speech shaped noise](image)

Deep neural networks

- **Why deep?**
  - As the number of layers increases, more abstract features are learned and they tend to be more invariant to superficial variations
  - Superior performance in practice if properly trained (e.g., convolutional neural networks)
- **Deep structure is harder to train**
  - Vanishing gradients: Error derivatives tend to become very small in lower layers
  - Restricted Boltzmann machines (RBMs) can be used for unsupervised pretraining
  - However, RBM pretraining is not needed with large training data
Different DNN architectures

- **Feedforward networks**
  - Multilayer perceptrons (MLPs) with at least two hidden layers
    - With or without RBM pretraining
  - Convolutional neural networks (CNNs)
    - Cascade of pairs of convolutional and subsampling layers
    - Invariant features are coded through weight sharing
  - Backpropagation is the standard training algorithm

- **Recurrent networks**
  - Backpropagation through time is commonly used for training recurrent neural networks (RNNs)
  - To alleviate vanishing or exploding gradients, LSTM (long short-term memory) introduces memory cells with gates to facilitate the information flow over time

Part II: Training targets

- **What supervised training aims to learn is important for speech separation/enhancement**
  - Different training targets lead to different mapping functions from noisy features to separated speech
  - Different targets may have different levels of generalization

- **While the IBM is first used in supervised separation (see Part IV), many training targets have since been used**
Different training targets

- **TBM** (Kjems et al.’09; Gonzalez & Brookes’14) is similar to the IBM except that interference is fixed to speech-shaped noise (SSN)
- **IRM** (Srinivasan et al.’06; Narayanan & Wang’13; Wang et al.’14; Hummersone et al.’14)

\[
IRM(t, f) = \left( \frac{S^2(t, f)}{S^2(t, f) + N^2(t, f)} \right) = \left( \frac{\text{SNR}(t, f)}{\text{SNR}(t, f) + 1} \right)^eta
\]

- \(S\) and \(N\) denote speech and noise
- \(\beta\) is a tunable parameter, and a good choice is 0.5
- With \(\beta = 0.5\), the IRM becomes a square root Wiener filter, which is the optimal estimator of the power spectrum

Different training targets (cont.)

- **Spectral magnitude mask** (Wang et al.’14)

\[
SMM(t, f) = \frac{|S(t, f)|}{|Y(t, f)|}
\]

- \(Y\) denotes noisy signal
- **Phase-sensitive mask** (Erdogan et al.’15)

\[
PSM(t, f) = \frac{|S(t, f)|}{|Y(t, f)|} \cos \theta
\]

- \(\theta\) denotes the difference of the clean speech phase and noisy speech phase within the T-F
- Because of phase sensitivity, this target usually leads to a better estimate of clean speech than the SMM mask
Complex Ideal Ratio Mask (cIRM)

- This mask is defined so that, when applied, it results in clean speech (Williamson et al.’16)
  \[ S(t, f) = cIRM * Y(t, f) \]

- With complex numbers, solve for mask components
  \[ cIRM(t, f) = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i \frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \quad x \in \{r, i\} \]

- Some form of compression (e.g. tangent hyperbolic function) should be used to bound mask values

Different training targets (cont.)

- **Target magnitude spectrum (TMS)** (Lu et al.’13; Xu et al.’14; Han et al.’14)
  \[ |S(t, f)| \]
  - A common form of the TMS is the log-power spectrum of clean speech

- **Gammatone frequency target power spectrum (GF-TPS)** (Wang et al.’14)
  \[ S_{GF}^2(t, f) \]

- The estimation of these two targets corresponds to spectral mapping, as opposed to T-F masking for earlier targets
Signal approximation

- In signal approximation (SA), training aims to estimate the IRM but the error is measured against the spectral magnitude of clean speech (Weninger et al.'14)

\[ SA(t, f) = (RM(t, f)|Y(t, f)| - |S(t, f)|)^2 \]

- \( RM(t, f) \) denotes an estimated IRM
- This objective function maximizes SNR

Illustration of various training targets

(a) IBM  
(b) TBM  
(c) IRM  
(d) GF-TPS

(e) SMM  
(f) PSM  
(g) TMS

Factory noise at -5 dB
Evaluation of training targets

- **Wang et al. (2014)** studied and evaluated a number of training targets using the same DNN with 3 hidden layers, each with 1024 units

- **Other evaluation details**
  - Target speech: TIMIT
  - Noises: SSN + four nonstationary noises from NOISEX
    - Training and testing on different segments of each noise
  - Trained at -5 and 0 dB, and tested at -5, 0, and 5 dB
  - Evaluation metrics
    - STOI: standard metric for predicted speech intelligibility
    - PESQ: standard metric for perceptual speech quality
    - SNR

Comparisons

- **Comparisons among several different training targets**
  - An additional comparison for the IRM target with multi-condition (MC) training of all noises (MC-IRM)

- **Comparisons with different approaches**
  - Speech enhancement (SPEH) (Hendriks et al.’10)
  - Supervised NMF: ASNA-NMF (Virtanen et al.’13), trained and tested in the same way as supervised separation
STOI comparison for factory noise

PESQ comparison for factory noise
Summary among different targets

- Among the two binary masks, IBM estimation performs better in PESQ than TBM estimation
- Ratio masking performs better than binary masking for speech quality
  - IRM, SMM, and GF-TPS produce comparable PESQ results
- SMM is better than TMS for estimation
  - Many-to-one mapping in TMS vs. one-to-one mapping in SMM, and the latter should be easier to learn
  - Estimation of spectral magnitudes or their compressed version tends to magnify estimation errors

Part III: Features for supervised separation

- For supervised learning, features and learning machines are two key components
- Early studies only used a few features
  - ITD/ILD (Roman et al.’03)
  - Pitch (Jin & Wang’09)
  - Amplitude modulation spectrogram (AMS) (Kim et al.’09)
- Subsequent studies expanded the list
  - A complementary set is recommended: AMS+RASTA-PLP+MFCC (Wang et al.’13)
  - A new feature, called MRCG (multi-resolution cochleagram), is found to be discriminative (Chen et al.’14)
A systematic feature study

- Extending Chen et al. (2014), Delfarah and Wang (2017) recently conducted a feature study that considers room reverberation and both speech enhancement and speaker separation
  - Evaluation done at the low SNR level of -5 dB, with implications for speech intelligibility improvements

Evaluation framework

- Each frame of features is sent to a DNN to estimate the IRM
Features selected for evaluation

- Features examined before include GF (gammatone feature), GFCC (Shao et al.’08), GFMC, AC-MFCC, RAS-MFCC, PAC-MFCC, PNCC (Kim & Stern’12), GFB, and SSF
- In addition, newly studied ones include
  - Log spectral magnitude (LOG-MAG)
  - Log mel-spectrum (LOG-MEL)
  - Waveform signal (WAV)

STOI improvements (%)
Result summary

- Best performing features are MRCG, PNCC, and GFCC under different conditions
- Gammatone-domain features (MRCG, GF and GFCC) perform strongly
- Modulation-domain features do not perform well
- Waveform signal without any feature extraction is not a good feature
- The most effective feature sets:
  - PNCC, GF, and LOG-MEL for speech enhancement
  - PNCC, GFCC, and LOG-MEL for speaker separation

Part IV. Separation algorithms

- **Monaural separation**
  - Speech-nonspeech separation
  - Speaker separation
  - Separation of reverberant speech
- **Multi-channel separation**
  - Spatial feature based separation
  - Masking based beamforming
Early monaural attempts at IBM estimation

- **Jin & Wang (2009) proposed MLP-based classification to separate reverberant voiced speech**
  - A 6-dimensional pitch-based feature is extracted within each T-F unit
  - Classification aims at the IBM, but with a training target that takes into account of the relative energy of the T-F unit
- **Kim et al. (2009) proposed GMM-based classification to perform speech separation in a masker dependent way**
  - AMS features are extracted within each T-F unit
  - First monaural speech segregation algorithm to achieve speech intelligibility improvement for NH listeners

DNN as subband classifier

- **Y. Wang & Wang (2013) first introduced DNN to address the speech separation problem**
  - DNN is used for as a subband classifier, performing feature learning from raw acoustic features
  - Classification aims to estimate the IBM
DNN as subband classifier

Extensive training with DNN

- **Training on 200 randomly chosen utterances from both male and female IEEE speakers, mixed with 100 environmental noises at 0 dB (~17 hours long)**
  - Six million fully dense training samples in each channel, with 64 channels in total
- **Evaluated on 20 unseen speakers mixed with 20 unseen noises at 0 dB**
- **DNN based classifier produced the state-of-the-art separation results at the time**
Speech intelligibility evaluation

- **Healy et al. (2013)** subsequently evaluated the classifier on speech intelligibility of hearing-impaired listeners
  - A very challenging problem: “The interfering effect of background noise is the single greatest problem reported by hearing aid wearers” (Dillon’12)
- **Two stage DNN training to incorporate T-F context in classification**

![Diagram showing the process of speech intelligibility evaluation](image)

Results and sound demos

- **Both HI and NH listeners showed intelligibility improvements**
- **HI subjects with separation outperformed NH subjects without separation**

![Sound levels and intelligibility graphs](image)
Generalization to new noises

• While previous speech intelligibility results are impressive, a major limitation is that training and test noise samples were drawn from the same noise segments
  • Speech utterances were different
  • Noise samples were randomized
• This limitation can be addressed through large-scale training for IRM estimation (Chen et al.’16)

Large-scale training

• Training set consisted of 560 IEEE sentences mixed with 10,000 (10K) non-speech noises (a total of 640,000 mixtures)
  • The total duration of the noises is about 125 h, and the total duration of training mixtures is about 380 h
  • Training SNR is fixed to -2 dB
• The only feature used is the simple T-F unit energy
• DNN architecture consists of 5 hidden layers, each with 2048 units
• Test utterances and noises are both different from those used in training
STOI performance at -2 dB input SNR

<table>
<thead>
<tr>
<th></th>
<th>Babble</th>
<th>Cafeteria</th>
<th>Factory</th>
<th>Babble2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>0.612</td>
<td>0.596</td>
<td>0.611</td>
<td>0.611</td>
<td>0.608</td>
</tr>
<tr>
<td>100-noise model</td>
<td>0.683</td>
<td>0.704</td>
<td>0.750</td>
<td>0.688</td>
<td>0.706</td>
</tr>
<tr>
<td>10K-noise model</td>
<td>0.792</td>
<td>0.783</td>
<td>0.807</td>
<td>0.786</td>
<td>0.792</td>
</tr>
<tr>
<td>Noise-dependent model</td>
<td>0.833</td>
<td>0.770</td>
<td>0.802</td>
<td>0.762</td>
<td>0.792</td>
</tr>
</tbody>
</table>

- DNN model with large-scale training provides similar results to noise-dependent model

DNN as spectral magnitude estimator

- *Xu et al. (2014) proposed a DNN-based enhancement algorithm*
  - DNN (with RBM pretraining) is trained to map from log-power spectra of noisy speech to those of clean speech
  - More input frames and training data improve separation results
Xu et al. results in PESQ

<table>
<thead>
<tr>
<th>SNR</th>
<th>Noisy</th>
<th>L-MMSE</th>
<th>SNN</th>
<th>DNN_1</th>
<th>DNN_2</th>
<th>DNN_3</th>
<th>DNN_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.99</td>
<td>3.32</td>
<td>3.48</td>
<td>3.46</td>
<td>3.59</td>
<td>3.60</td>
<td>3.59</td>
</tr>
<tr>
<td>15</td>
<td>2.65</td>
<td>2.99</td>
<td>3.26</td>
<td>3.24</td>
<td>3.35</td>
<td>3.36</td>
<td>3.36</td>
</tr>
<tr>
<td>10</td>
<td>2.32</td>
<td>2.65</td>
<td>2.99</td>
<td>2.97</td>
<td>3.08</td>
<td>3.10</td>
<td>3.09</td>
</tr>
<tr>
<td>5</td>
<td>1.98</td>
<td>2.30</td>
<td>2.68</td>
<td>2.65</td>
<td>2.76</td>
<td>2.78</td>
<td>2.78</td>
</tr>
<tr>
<td>0</td>
<td>1.65</td>
<td>1.93</td>
<td>2.32</td>
<td>2.29</td>
<td>2.38</td>
<td>2.41</td>
<td>2.41</td>
</tr>
<tr>
<td>5</td>
<td>1.38</td>
<td>1.55</td>
<td>1.92</td>
<td>1.89</td>
<td>1.95</td>
<td>1.97</td>
<td>1.97</td>
</tr>
<tr>
<td>Ave</td>
<td>2.16</td>
<td>2.46</td>
<td>2.78</td>
<td>2.75</td>
<td>2.85</td>
<td>2.87</td>
<td>2.87</td>
</tr>
</tbody>
</table>

With trained noises

<table>
<thead>
<tr>
<th>SNR</th>
<th>Noisy</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>3.15</td>
<td>2.89</td>
<td>3.52</td>
<td>3.19</td>
<td>3.43</td>
<td>3.24</td>
<td>3.58</td>
<td>3.90</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.81</td>
<td>2.55</td>
<td>3.23</td>
<td>2.85</td>
<td>3.19</td>
<td>2.96</td>
<td>3.31</td>
<td>3.61</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.47</td>
<td>2.21</td>
<td>2.89</td>
<td>2.51</td>
<td>2.93</td>
<td>2.66</td>
<td>3.03</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.14</td>
<td>1.87</td>
<td>2.57</td>
<td>2.11</td>
<td>2.60</td>
<td>2.30</td>
<td>2.71</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.81</td>
<td>1.56</td>
<td>2.21</td>
<td>1.72</td>
<td>2.24</td>
<td>1.92</td>
<td>2.38</td>
<td>1.93</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.52</td>
<td>1.28</td>
<td>1.82</td>
<td>1.34</td>
<td>1.85</td>
<td>1.52</td>
<td>1.96</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>Ave</td>
<td>2.32</td>
<td>2.06</td>
<td>2.70</td>
<td>2.29</td>
<td>2.71</td>
<td>2.43</td>
<td>2.83</td>
<td>2.47</td>
<td></td>
</tr>
</tbody>
</table>

With two untrained noises (A: car; B: exhibition hall)

---

Xu et al. demo

Street noise at 10 dB (upper left: DNN; upper right: log-MMSE; lower left: clean; lower right: noisy)
Part IV. Separation algorithms

- **Monaural separation**
  - Speech-nonspeech separation
  - Speaker separation
  - Separation of reverberant speech

- **Multi-channel separation**
  - Spatial feature based separation
  - Masking based beamforming

---

**DNN for two-talker separation**

- **Huang et al. (2014; 2015)** proposed a two-talker separation method based on DNN, as well as RNN (recurrent neural network)
  - Mapping from an input mixture to two separated speech signals
  - Using T-F masking to constrain target signals
    - Binary or ratio
Network architecture and training objective

- A discriminative training objective maximizes signal-to-interference ratio (SIR)

\[
\|\mathbf{y}_{1c} - \mathbf{y}_{1c}\|^2 - \gamma \|\mathbf{y}_{1c} - \mathbf{y}_{2c}\|^2 + \|\mathbf{y}_{2c} - \mathbf{y}_{2c}\|^2 - \gamma \|\mathbf{y}_{2c} - \mathbf{y}_{1c}\|^2
\]

Huang et al. results

- **DNN and RNN perform at about the same level**
  - About 4-5 dB better in terms of SIR than NMF, while maintaining better SDRs and SARs (input SNR is 0 dB)
  - Demo at https://sites.google.com/site/deeplearningsourceseparation
Talker dependency in speaker separation

- Huang et al.’s speaker separation is talker-dependent, i.e. the same two talkers are used in training and testing
  - Speech utterances are different between training and testing
- Speaker separation can be divided into three classes
  - Talker dependent
  - Target dependent
  - Talker independent

Target-dependent speaker separation

- In this case, the target speaker is the same between training and testing, while interfering talkers are allowed to change
- Target-dependent separation can be satisfactorily addressed by training with a variety of interfering talkers (Du et al.’14; Zhang & Wang’16)
Talker-independent speaker separation

- This is the most general case, and it cannot be adequately addressed by training with many speaker pairs
- Talker-independent separation can be treated as unsupervised clustering (Bach & Jordan ’06; Hu & Wang ’13)
  - Such clustering, however, does not benefit from discriminant information utilized in supervised training
- Deep clustering (Hershey et al.’16) is the first approach to talker-independent separation by combining DNN based supervised feature learning and spectral clustering

Deep clustering

- With the ground truth partition of all T-F units, an affinity matrix is defined as

\[ A = YY^T \]

  - \( Y \) is the indicator matrix built from the IBM. \( Y_{i,c} \) is set to 1 if unit \( i \) belongs to (or dominated by) speaker \( c \), and 0 otherwise
  - \( A_{i,j} = 1 \) if units \( i \) and \( j \) belong to the same speaker, and 0 otherwise
- To estimate the ground truth partition, DNN is trained to produce embedding vectors such that clustering in the embedding space provides a better partition estimate
Deep clustering (cont.)

- **DNN training minimizes the following cost function**

\[ C_Y(V) = \| \hat{A} - A \|_F^2 = \| VV^T - YY^T \|_F^2 \]

- \( V \) is an embedding matrix for T-F units, and each row represents an embedding vector for one T-F unit
- \( \| \cdot \|_F^2 \) denotes the squared Frobenius norm
- **During inference (testing), the K-means algorithm is applied to cluster T-F units into speaker clusters**
- **Isik et al. (2016) extend deep clustering by incorporating an enhancement network after binary mask estimation, and performing end-to-end training of embedding and clustering**

Permutation invariant training (PIT)

- **Recognizing that talker-dependent separation ties each DNN output to a specific speaker (permutation variant), PIT seeks to untie DNN outputs from speakers in order to achieve talker independence (Kolbak et al.’17)**
  - Specifically, for a pair of speakers, there are two possible assignments, each of which is associated with a mean squared error (MSE). The assignment with the lower MSE is chosen and the DNN is trained to minimize the corresponding MSE
PIT illustration and versions

- Frame-level PIT (tPIT): Permutation can vary from frame to frame, hence needs speaker tracing (sequential grouping) for speaker separation
- Utterance-level PIT (uPIT): Permutation is fixed for a whole utterance, hence needs no speaker tracing

Interspeech'18 tutorial

CASA based approach

- Limitations of deep clustering and PIT
  - In deep clustering, embedding vectors for T-F units with similar energies from underlying speakers tend to be ambiguous
  - uPIT does not work as well as tPIT at the frame level, particularly for same-gender speakers, but tPIT requires speaker tracing

- Speaker separation in CASA is talker-independent
  - CASA performs simultaneous (spectral) grouping first, and then sequential grouping across time

- Liu & Wang (2018) proposed a CASA based approach by leveraging PIT and deep clustering
  - For simultaneous grouping, tPIT is trained to predict the spectra of underlying speakers at each frame
  - For sequential grouping, DNN is trained to predict embedding vectors for simultaneously grouped spectra

Interspeech'18 tutorial
Sequential grouping in CASA

- **Differences from deep clustering**
  - In deep clustering, DNN based embedding is done at the T-F unit level, whereas it is done at the frame level in CASA.
  - Constrained K-means in CASA ensures that the simultaneously separated spectra of the same frame are assigned to different speakers.

Speaker separation performance

- **Talker-independent separation produces high-quality speaker separation results, rivaling talker-dependent separation results**

<table>
<thead>
<tr>
<th>SDR improvements (in dB)</th>
<th>Same Gender</th>
<th>Different Gender</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep clustering++ (Isik et al., 2016)</td>
<td>9.4</td>
<td>12.0</td>
<td>10.8</td>
</tr>
<tr>
<td>PIT (Kolbeek et al., 2017)</td>
<td>7.5</td>
<td>12.2</td>
<td>10.0</td>
</tr>
<tr>
<td>CASA (Liu and Wang, 2018)</td>
<td>10.3</td>
<td>12.6</td>
<td>11.3</td>
</tr>
</tbody>
</table>
Talker-independent speaker separation demos

New pair of male-male speaker mixture

Speaker1 - uPIT  Speaker2 - uPIT
Speaker1 – DC++  Speaker2 – DC++
Speaker1 - CASA  Speaker2 - CASA
Speaker1 - clean  Speaker2 - clean

Part IV. Separation algorithms

- Monaural separation
  - Speech-nonspeech separation
  - Speaker separation
  - Separation of reverberant speech
- Multi-channel separation
  - Spatial feature based separation
  - Masking based beamforming
Reverberation

- **Reverberation is everywhere**
  - Reflections of the original source (direct sound) from various surfaces
- **Adverse effects on speech processing, especially when mixed with noise**
  - Speech communication
  - Automatic speech recognition
  - Speaker identification
- **Previous work**
  - Inverse filtering (Avendano & Hermansky’96; Wu & Wang’06)
  - Binary masking (Roman & Woodruff’13; Hazrati et al.’13)

DNN for speech dereverberation

- **Learning the inverse process (Han et al.’14)**
  - DNN is trained to learn the mapping from the spectrum of reverberant speech to the spectrum of clean (anechoic) speech
    - Work equally well in spectrogram and cochleagram
  - Straightforward extension to separate reverberant and noisy speech (Han et al.’15)
Learning spectral mapping

- **DNN**
  - Rectified linear + sigmoid
  - Three hidden layers
  - Input: current frames + 5 neighboring frame on each side
  - Output: current frame

Dereverberation

- Clean
- Reverberant ($T_{60} = 0.6$ s)
- Dereverberated
Two-stage model for reverberant speech enhancement

- **Noise and reverberation are different kinds of interference**
  - Background noise is an additive signal to the target speech
  - Reverberation is a convolutive distortion of a speech signal by a room impulse response
- **Following this analysis, Zhao et al. (2017) proposed a two-stage model to address combined noise and reverberation**

---

DNN architecture

- **The system has three modules:**
  - Denoising module performs IRM estimation to remove noise from noisy-reverberant speech
  - Dereverberation module performs spectral mapping to estimate clean-anechoic speech
  - Time-domain signal reconstruction (TDR) module (Wang & Wang’15) performs time-domain optimization to improve magnitude spectrum estimation
Results and demo

- Average across different reverberation times (0.3 s, 0.6 s, 0.9 s), different SNRs (-6 dB, 0 dB, 6 dB), and four noises (babble, SSN, DLIVING, PCAFETER)

<table>
<thead>
<tr>
<th></th>
<th>STOI (in %)</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td>61.8</td>
<td>1.22</td>
</tr>
<tr>
<td>masking</td>
<td>77.7</td>
<td>1.81</td>
</tr>
<tr>
<td>proposed</td>
<td>82.6</td>
<td>2.08</td>
</tr>
</tbody>
</table>

- Recent tests with hearing-impaired listeners show substantial intelligibility improvements

- Demo (T60 = 0.6 s, babble noise at 3 dB SNR)

Part IV. Separation algorithms

- **Monaural separation**
  - Speech-nonspeech separation
  - Speaker separation
  - Separation of reverberant speech

- **Multi-channel separation**
  - Spatial feature based separation
  - Masking based beamforming
Binaural separation of reverberant speech

- **Speech separation has extensively used binaural cues**
  - Localization, and then location-based separation
- **T-F masking for separation**
  - Roman et al. (2003) proposed the first supervised method for binaural separation
  - DUET (Yilmaz & Richard'04) is the first unsupervised method

DNN based approach

- **Jiang et al. (2014) proposed DNN classification based on binaural features for reverberant speech separation**
  - Interaural time difference (ITD) and interaural level difference (ILD)
- **Using binaural features to train a DNN classifier to estimate the IBM**
  - Monaural GFCC features are also used
- **Systematic examination of generalization to different configurations and reverberation conditions**
Training and test corpus

- **ROOMSIM is used with KEMAR dummy head**
  - A library of binaural impulse responses (BIRs) is created with a room configuration: $6m \times 4m \times 3m$
  - Three reverberant conditions with $T_{60}$: 0s, 0.3s and 0.7s

- **Seventy-two source azimuths**
  - Azimuth angles from 0° to 355°, uniformly sampled at 5°, with distance fixed at 1.5m and elevation at zero degree
  - Target fixed at 0°
  - Number of sources: 2, 3 and 5

- **Target signals: TIMIT utterances**
  - Interference: Babble
  - Training SNR is 0 dB (with reverberant target as signal)
  - Default test SNR is -5 dB

Generalization to untrained configurations

- **Two-source configuration with reverberation (0.3s $T_{60}$)**

  - With reasonable sampling of spatial configurations, DNN classifiers generalize well
Combining spatial and spectral analyses

- Recently, Zhang & Wang (2017) used a more sophisticated set of spatial and spectral features
  - Separation is based on IRM estimation
  - Spectral features are extracted after fixed beamforming
  - Substantially outperforming conventional beamformers

Masking based beamforming

- Beamforming as a spatial filter needs to know the target direction for steering purposes
- Steering vector is typically supplied by direction-of-arrival (DOA) estimation of the target source, or source localization
- However, sound localization in reverberant, multi-source environments itself is difficult
  - For human audition, localization depends on separation (Darwin’08)
- A recent idea is to use monaural T-F masking to guide beamforming (Heymann et al.’16; Higuchi et al.’16)
  - Supervised masking helps to specify the target source
  - Masking also helps by suppressing interfering sound sources
MVDR beamformer

- To explain the idea, look at the MVDR beamformer
  - MVDR (minimum variance distortionless response) aims to minimize the noise energy from nontarget directions while maintaining the energy from the target direction
- Array signals can be written as

\[
y(t, f) = c(f)s(t, f) + n(t, f)
\]

- \(y(t, f)\) and \(n(t, f)\) denote the spatial vectors of the noisy speech signal and noise at frame \(t\) and frequency \(f\)
- \(s(t, f)\): Speech source
- \(c(f)s(t, f)\): received speech signal by the array
- \(c(f)\): steering vector of the array

MVDR beamformer (cont.)

- We solve for an optimal weight vector

\[
w_{opt} = \arg\min_w \{w^H \Phi_n w\}, \quad \text{subject to } w^H c = 1
\]

- \(H\) denotes the conjugate transpose
- \(\Phi_n\) is the spatial covariance matrix of the noise
- As the minimization of the output power is equivalent to the minimization of the noise power:

\[
w_{opt} = \frac{\Phi_n^{-1} c}{c^H \Phi_n^{-1} c}
\]
MVDR beamformer (cont.)

- The enhanced speech signal (MVDR output) is

\[ \hat{s}(t) = \mathbf{w}_{opt}^H \mathbf{y}(t) \]

- Hence, the accurate estimation of \( \mathbf{c} \) and \( \mathbf{\Phi}_n \) is key
- \( \mathbf{c} \) corresponds to the principal component of \( \mathbf{\Phi}_x \), the spatial covariance matrix of speech
- **With speech and noise uncorrelated, we have**

\[ \mathbf{\Phi}_x = \mathbf{\Phi}_y - \mathbf{\Phi}_n \]

- A noise estimate is crucial for beamforming performance, just like in traditional speech enhancement

Masking based beamforming

- **A T-F mask provides a way to more accurately estimate noise (and speech) covariance matrix from noisy input**
  - Heymann et al. (2016) use RNN with LSTM for monaural IBM estimation
  - Higuchi et al. (2016) compute a ratio mask using a spatial clustering method

From Erdogan et al. (2016)
Masking based beamforming (cont.)

- Zhang et al. (2017) trained a DNN for monaural IRM estimation, and multiple ratio masks are combined into one via maximum selection

\[
\Phi_n(t, f) = \frac{1}{\sum_{l=-L}^{t+L} (1 - RM(l, f))} \sum_{l=-L}^{t+L} (1 - RM(l, f)) y(l, f)y(l, f)^H
\]

- \(RM(l, f)\) denotes the estimated IRM from the DNN at frame \(l\) and frequency \(f\)
- An element of the noise covariance matrix is calculated per frame by integrating a window of neighboring \(2L+1\) frames. Per-frame estimation of \(\Phi_n\) is better than over the entire utterance or a signal segment.

Interspeech'18 tutorial

Masking based beamforming (cont.)

- Masking based beamforming is responsible for impressive ASR results on CHiME-3 and CHiME-4 challenges
  - These results are much better than using traditional beamformers
  - Masking based beamforming represents a major advance in beamforming based separation and multi-channel ASR
Part V: Concluding remarks

- Formulation of separation as classification, mask estimation, or spectral mapping enables the use of supervised learning
- Advances in supervised speech separation in the last few years are truly impressive (Wang & Chen’18)
  - Large improvements over unprocessed noisy speech and related approaches
  - The first demonstrations of speech intelligibility improvement in noise
  - Elevation of beamforming performance

A solution in sight for cocktail party problem?

- What does a solution to the cocktail party problem look like?
  - A system that achieves human auditory analysis performance in all listening situations (Wang & Brown’06)
- An ASR system that matches the human speech recognition performance in all noisy environments
  - Dependency on ASR
A solution in sight (cont.)?

- A speech separation system that helps hearing-impaired listeners to achieve the same level of speech intelligibility as normal-hearing listeners in all noisy environments
  - This is my current working definition – see my IEEE Spectrum cover story in March, 2017

Further remarks

- **Supervised speech processing is the mainstream**
  - Signal processing provides an important domain for supervised learning, and it in turn benefits from rapid advances in machine learning

- **Use of supervised processing goes beyond speech separation and recognition**
  - Multipitch tracking (Huang & Lee’13, Han & Wang’14)
  - Voice activity detection (Zhang et al.’13)
  - SNR estimation (Papadopoulos et al.’16)
  - Localization (Pertila & Cakir’17; Wang et al.’18)
Review of presentation

I. Introduction
II. Training targets
III. Features
IV. Separation algorithms
   • Monaural separation
     • Speech-nonspeech separation
     • Speaker separation
     • Separation of reverberant speech
   • Multi-channel separation
     • Spatial feature based separation
     • Masking based beamforming
V. Concluding remarks

Resources and acknowledgments

• This tutorial is based in part on the following overview

• DNN Matlab toolbox for speech separation
  • http://www.cse.ohio-state.edu/pnl/DNN_toolbox

• Source programs for some algorithms discussed in this tutorial are available at OSU Perception & Neurodynamics Lab’s website
  • http://www.cse.ohio-state.edu/pnl/software.html

• Thanks to Jitong Chen, Donald Williamson, Yuzhou Liu, and Zhong-Qiu Wang for their assistance in the tutorial preparation
Cited literature and other readings

Ahmadi, Groß, & Sinex (2013) JASA 133: 1687-1692.
Cao et al. (2011) JASA 129: 2227-2236.
Chen et al. (2016) JASA 139: 2604-2612.
Du et al. (2014) ICSP: 65-68.
Han et al. (2014) ICASSP: 4628-4632.
Hazrati et al. (2013) JASA 133: 1607-1614.
Helmholtz (1863) On the Sensation of Tone. Dover.
Herahey et al. (2016) ICASSP: 31-35.
Higuchi et al. (2016) ICASSP: 5210-5214.
Huang et al. (2014) ICASSP: 1581-1585.
Isik et al. (2016) Interspeech: 545-549.

Cited literature and other readings (cont.)

Lu et al. (2013) Interspeech: 555-559.
Wenninger et al. (2014) GlobalSIP MLASP Symp.
Zhao et al. (2017) ICASSP: 5580-5584.