Supervised Speech Separation Based on Deep Learning – Overview

(2017), Wang, Chen

Index terms:

**Speech separation, speaker separation, speech enhancement, supervised speech separation, deep learning, deep neural networks, speech dereverberation, time-frequency masking, array separation, beamforming**

Definitions / Acronyms:

General:

* SNR: Signal to Noise Ratio
* SRT: ?
* CASA: Computational Auditory Scene Analysis 🡪 Bergman
* T-F: time-frequency

Masks:

* IBM: Ideal Binary Mask
* TBM: Target Binary Mask
* IRM: ideal ratio mask
* SMM: spectral magnitude mask
* cIRM :complex ideal ratio mask
* PSM: phase sensitive mask

Mappings:

* TMS: target magnitude spectrum
  + GF-TPF: Gammatone frequency target power spectrum

Neural networks:

* DNN: Deep Neural Network
* CNN: Convolutional Neural Network
* RNN: Recurrent Neural Network
* GAN: Generative Adversarial Network
* MLP: feedforward multilayer perceptron
* MSE: Mean Square Error
* MLE: maximum Likelihood Estimator
* RBM: Restrictive Boltzmann Machine
* ReLU: Rectified Linear Unit
* Depth: number of layers in a NN (without counting the 1st one)
* LSTM : long short-term memory

Evaluation methods:

* SDR: source-to-distortion ratio
* SIR: source-to-interference ratio
* SAR: source-to-artifact ratio
* STOI: short-time objective intelligibility
* PESQ: perceptual evaluation of speech quality
* MOS: mean opinion score

Goal of speech separation: extract *target speech* (message) from *background interferences* (noise). Corresponds to auditory stream segregation.

Other situation: cocktail party: superposition of several messages over noise. For each of these messages, noise corresponds to background noise + the other messages.

Facts: intelligibility decreases with the increase of voices in the environment and with broadband noise as interferer. Narrow-band signals are easily identifiable.

-----------------------------------------------------------------------------------

Separation methods:

1. Monaural
   1. Speech enhancement: analyses general statistics of speech and noise 🡺 estimation of clean speech from noisy speech with noise estimate
      1. Ex: Spectral subtraction: power spectrum(noise estimate) - PS(noisy speech)
      2. Hyp: stationary background noise (or more stat. than speech)
   2. CASA
      1. Concept: Based on perceptual principles of auditory streaming: grouping etc. 🡪 SMC7
      2. Tandem algo: separates speech by alternating pitch estimation and pitch-based grouping
2. Array-based
   1. Beamforming: boosts the signal that arrives with a specific direction
      1. “Delay-and-sum” technique: adds multiple mic signal from the target direction in phase and uses phase differences to attenuate signals from the other directions.

Parameters: spacing + size + config of the array

Limitations: if target & interfering sources are close. If reverb

* 1. Supervised learning problem:
     1. Concept: T-F masking in CASA: a 2D- mask (weighting) is applied to the T-F representation of a source mixture to extract the source
     2. Ex: IBM: IBM as computational goal ~= binary classif

Improves speech intelligibility for both normal-hearing and hearing-impaired listeners.

* + 1. Deals with the following components:
       1. Learning machine
       2. Training targets
       3. Acoustic features

-------------------------------------------------------------------------------------------------------------------------

Classifiers and learning machines: types of DNNs for supervised speech separation:

1. Feedforward MLP
   1. Architecture
      1. feedforward connections from the input layer to the output layer
      2. layer-by-layer
      3. consecutive layers are fully connected.
      4. Some hidden layers
   2. Algo: backpropagation algorithm (Vanilla if I’m right)
   3. Loss function:
      1. Cross Entropy: −1/𝑁\*ΣΣ𝐼𝑖,𝑐𝑙𝑜𝑔(𝑝𝑖,𝑐)

I: output model neuron

𝑝𝑖, : predicted probability of 𝑖 belonging to class *c*.

*N:* number of output neurons

*C:* number of classes

𝐼𝑖,: binary indicator, which takes 1 if the desired class of neuron 𝑖 is 𝑐 and 0 otherwise.

* + 1. MSE: 1/𝑁\*Σ(𝑦𝑖−𝑦𝑖̂)²

𝑦𝑖̂: predicted output for neuron i

𝑦 : desired output for neuron i

* 1. Problem:
     1. Vanishing gradient 🡺 use single-hidden layer
  2. Solutions:
     1. layerwise unsupervised pretraining (=fine tuning) before supervised training: RBM
     2. ReLu activation function
     3. Skip connections

1. CNN
   1. Concept
      1. Suits especially visual-domain (can we generalize this affirmation to all 2D-domains?)
      2. incorporates invariances in pattern recognition (ex: translation invar.)
   2. Architecture:
      1. cascade of pairs of a convolutional layer
         1. Convolutional layer: made of multiple feature maps that extract a local feature regardless of its position in the previous layer = connection weights sharing (even if have different receptive fields).
         2. Receptive field: local area of the previous layer that is connected to the neuron,
      2. + a subsampling layer

performs local averaging or maximization over the receptive fields of the neurons in the convolutional layer

Subsampling: to reduce resolution and sensitivity to local variations.

* 1. Benefit:

Little number of parametres

1. RNN:
   1. Concept:
      1. Allows feedback connections
      2. treats input samples as a sequence and models the changes over **time 🡺** adapted to speech sep. pb
      3. RNN = DNN with infinite depth
   2. Problems:
      1. Vanishing gradient
      2. Exploding gradient
   3. Solutions:
      1. LSTM: memory cells through 3 gates:
         1. Input gate: how much current information should be added to the memory cell.
         2. Forget gate: how much former information to keep
         3. Output gate
2. GAN
   1. Concept
      1. simultaneously trained models: a generative model *G* and a discriminative model *D*
         1. *G:* learns to model labelled data
         2. *D* (usually a binary classifier): learns to discriminate between generated samples and target samples from training data.
      2. (Analogous to the minimax theory)
      3. Both models improve their accuracy until generated samples are indistinguishable from real ones.
      4. Use D to shape the loss function of G

DNN with 1 hidden layer:

* SVM
* MLP
* GMM
* => All of these methods have been tested in order to perform supervised Speech Sep.

Here: nb layers >=2

---------------------------------------------------------------------------------------------------------------------------

Evaluation metrics:

1. signal-level: quantify the degrees of signal enhancement or interference reduction:
   1. SNR
   2. Speech distorson
   3. Noise residue
   4. SDR
   5. SIR
   6. SAR
2. perception-level: separately evaluate speech intelligibility and speech quality
   1. Intelligibility
      1. IBM
      2. Articulation index
      3. A.+B. = HIT-FA rate
         1. HIT: percent of speech-dominant T-F units in the IBM that is correctly classified
         2. FA (false-alarm) : percent of noise-dominant units that is incorrectly classified.
      4. STOI: measures the correlation between the short-time temporal envelopes of a reference (clean) utterance and a separated utterance ϵ[0 1]
         1. Rk: overpredicts intelligibility scores
         2. Correlates best with human perception
   2. Quality
      1. PESQ: applies an auditory transform to produce a **loudness** spectrum, and compares the loudness spectra of a clean reference signal and a separated signal to produce a score in a range of -0.5 to 4.5, corresponding to the prediction of the perceptual MOS (mean opinion score).

Training targets

1. Masking-based targets: describe the time-frequency relationships of clean speech to background interference
   1. IBM: = local TBM
      1. Concept (aginn): based on the auditory masking phenomenon in audition and the exclusive allocation principle in auditory scene analysis
      2. LC: local criterion

𝐼𝐵𝑀=1, if 𝑆𝑁(𝑡,𝑓)>𝐿𝐶,   
 0 otherwise

* + 1. Loss function: cross-entropy
  1. TBM = global IBM with criterion= stationary signal corresponding to the average of all speech signals.
  2. IRM: soft version of the IBM [looks like a Winer filter]:
     1. 𝐼𝑅𝑀=(𝑆(𝑡,𝑓)² / (𝑆(𝑡, 𝑓)2+𝑁(𝑡,𝑓)2)𝛽
     2. Hyp: S and N uncorrelated 🡪 ok if just additive noise; not ok for reverb (at least early reverb)
     3. 𝛽 = 0.5
     4. <1
     5. Loss function: RME
  3. SMM: based on STFT:
     1. SMM(𝑡,𝑓)=|𝑆(𝑡,𝑓)||𝑌(𝑡,𝑓)|
     2. Method: To obtain separated speech, we apply the SMM or its estimate to the spectral magnitudes of noisy speech, and resynthesize separated speech with the phases of noisy speech
  4. PSM: SMM extended by considering a measure of phase
     1. PSM(𝑡,𝑓)= SMM(𝑡,𝑓)\*cos 𝜃
     2. 𝜃 : difference between clean speech phase and noisy speech phase (within the T-F unit)
     3. 🡺 higher SNR than SMM & better estimate of S
  5. cIRM:.
     1. Concept: ideal mask in the complex domain. Considers real and imaginary components of both noisy and clean speech
     2. Benefits:
        1. can perfectly reconstruct clean speech from noisy speech
        2. provides an estimate of the phase of the target signal
     3. Rk: not bounded 🡺 compression needed
  6. Signal Approx (SA)
     1. Concept: train a ratio mask estimator that minimizes the difference between the spectral magnitude of clean speech and that of estimated speech
     2. Formula : 𝑆𝐴(𝑡,𝑓)= [𝑅𝑀(𝑡,𝑓)|𝑌(𝑡,𝑓)|− |𝑆(𝑡,𝑓)|]2

𝑅(𝑡,𝑓) refers to an estimate of the SMM.

1. Mapping-based targets: spectral representations of clean speech
   1. TMS of clean speech:
      1. Concept:
         1. estimate the magnitude **spectrogram** of clean speech from that of noisy speech.
         2. Use mel- or power- spectrum instead of Magnitude-spectrum
      2. Method:
         1. Most used TMS: log-power spectrum normalized to zero mean and unit variance
         2. Compression: with fx ‘log’
         3. Combination of the estimated speech magnitude with noisy phase to produce the separated speech waveform.
      3. Cost function: MSE / MLE
   2. GF-TPS
      1. Concept:
         1. Base on a cochleagram based on a gammatone filterbank.
         2. Target: power of the cochleagram response to clean speech

---------------------------------------------------------------------------------------------------------------------------

Features:

1. Binaural separation: ITD & IID (ILD)
2. Monaural separation:
   1. **pitch**-**based** features
   2. amplitude modulation spec (AMS)
   3. cepstral-domain fetaures:
      1. mel-frequency cepstral coefficient MFCC
      2. delta-spectral cepstral coefficient (DSCC)
   4. gammatone-domain features:
      1. gammatone feature
      2. gammatone frequency cepstral coefficient GFCC
      3. gammatone frequency modulation coefficient (GFMC)
   5. linear prediction features:
      1. perceptual linear prediction (PLP)
      2. relative spectral transform PLP (RASTA- PLP)
   6. Autocorrelation features: calculated from other features🡪 e.g. ceps. Features
   7. medium-time filtering features 🡪 e.g. power normalized cepstral coefficients (PNCC)
   8. cochlea-based feature: Multi-Resolution Cochleagram (MRCG)
   9. Log-spectra
      1. Log-spectral magnitude(LOG-MAG)
      2. log mel-spectrum feature (LOG-MEL)
   10. raw waveform signal (WAV) [without any feature extraction]

# group Lasso??

* Exp1.: Looks like MRCG performed best for all types of noise; pitch features performed badly
* Exp.2: {PNCC + GF\*+LOG-MEL} but depend on the result

🡺 Importance of features for supervised speech separation. Using raw material is inefficient

-----------------------------------------------------------------------------------------------------------------------------------------

Monaural separation algorithms

Aims: speech enhancement / speech dereverberation /dereverberation plus denoising /speaker separation.

2012: Wang & Wang (1st one)

* used DNN for subband classification to estimate the IBM
* 64-channel gammatone filterbank to derive subband signals 🡪feature extraction 🡪 feed a subband DNN 🡪 more feature extraction🡪 linear SVM🡪 posterior mask
* Extended with a 2nd stage: 🡪 local window of the posterior mask feeds a T-F unit (no weight sharing)

2013: Lu et al.

* DAE: Deep Autoencoder: contains several AE – used for pretraining #mapping based method

2013: Xu et al.:

* MLP with RBM pretraining form log-ps

From then:

* RNN with LSTM 🡺 speech enhancement + robust ASR
* RNN 🡺 estimate PSM
* Deep stacking network 🡪 IBM estimation 🡺 pitch estimation
* DNN 🡪 simultaneously estimate the real and imaginary components of the cIRM🡺 better speech quality over IRM estimation
* 🡺Speech enhancement at the phoneme level
* 🡺Reproduce perceptual masking
* **hierarchical** DNN performing subband spectral mapping provides better enhancement than a **single** DNN performing fullband mapping
* skip connections between non-consecutive layers
* Multi-target training outperforms single-target training
* CNN🡪 IRM estimation + spectral mapping
* temporal mapping without resorting to a T-F representation.
  + Benefit: no need to use the phase of noisy speech in reconstructing enhanced speech
  + **“CNNs appear to be a natural choice for temporal mapping.”** [base on Fu et al. work, who removed connections to a fully-connected CNN]
* speech enhancement GAN (SEGAN) to perform temporal mapping (🡺enhancement or denoising):
  + D ~G: fully CNN, D provides the loss-function to G
  + Results: inconclusive and worse than masking or mapping methods. But another one got results equivalent to a DNN
* Not only DNN: NMF

Generalisation to untrained conditions:

1. Noise
2. Speaker
3. SNR