NEURAL TARGET SPEECH EXTRACTION

Marc Delcroix (NTT), Katerina (“Katka”) Zmolikova (BUT)

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

About us

Marc Delcroix (NTT)

- M.Eng. degree from the Free University of Brussels, and the École Centrale Paris, 2003.
- Ph.D. degree from Hokkaido University, Sapporo, Japan, 2007.
- Researcher at NTT since then working on various aspect of speech processing.
- Main Research interests:
  - Robust speech recognition
  - Model adaptation
  - Speech enhancement
  - Target speech extraction

Kateřina (“Katka”) Žmolíková (BUT):

- Ing. Degree from Brno University of Technology (BUT), Czech Republic, 2016.
- Currently working towards her Ph.D. degree.
- Main Research interests:
  - Speech enhancement
  - Speech recognition
  - Target speech extraction

Copyright 2021 NTT CORPORATION, BUT

Acknowledgments

- Thanks to our co-authors and our colleagues at
  - BUT Speech@FIT
  - NTT

Copyright 2021 NTT CORPORATION, BUT

Table of contents (1/2)

1. Introduction [Marc]
2. Target speech extraction: Audio clue-based approaches [Katka]

   [5 min break]

3. Multi-channel approaches [Katka]
4. Visual/multimodal approaches [Marc]
5. Other tasks
   - Target speaker VAD [Katka]
   - Target speaker ASR [Marc]
6. Conclusion & future directions [Marc]

   [Discussion and QA]

Copyright 2021 NTT CORPORATION, BUT

Table of contents (1/2)

1. Introduction [Marc] – 30 min
2. Target speech extraction: Audio clue-based approaches [Katka] – 40 min

   [10 min break/QA]

3. Multi-channel approaches [Katka] – 20 min
4. Visual/multimodal approaches [Marc] – 25 min
5. Other tasks –25 min
   - Target speaker VAD [Katka] – 5 min
   - Target speech recognition [Marc] – 20 min
6. Conclusion & future directions [Marc] – 8 min

   [Discussion and QA]

Copyright 2021 NTT CORPORATION, BUT

Questions

- Short questions at end of each part or through chat
- Discussion and longer questions at the end

Copyright 2021 NTT CORPORATION, BUT

List of acronyms(1/2)

AM	Acoustic Model
AMS	Anchor Mean Subtraction
ASR	Automatic Speech Recognition
BLSTM	Bidirectional Long Short-Term Memory
BSS	Blind Source Separation
CE	Cross-Entropy
CMS	Cepstral Mean Subtraction
CNN	Convolutional Neural Network
Conv-TasNet	fully-Convolutional Time-domain audio separation Network
CTC	Connectionist Temporal Classification
DANet	Deep Attractor Network
DENet	Deep Extractor Network
DNN	Deep Neural Network
DPRNN	Dual-Path RNN
E2E	End-to-End
EEG	Electroencephalogram
FC	Fully-Connected
FiLM	Feature-wise Linear Modulation
GAN	Generative Adversarial Network
GCC	Generalized Cross-Correlation
HMM	Hidden Markov Model
IBM	Ideal Binary Mask
ICA	Independent Component analysis
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
LM	Language Model
LRS	Lip Reading Sentences
LSTM	Long Short-Term Memory
MC	Multi-Channel
MSE	Minimum Squared Error
MVDR	Minimum Variance Distortion-less Beamformer
MWF	Multi-channel Wiener Filter

Copyright 2021 NTT CORPORATION, BUT

List of acronyms(2/2)

NMF	Non-negative Matrix Factorization
NN	Neural Network
PESQ	Perceptual Evaluation of Speech Quality
PIT	Permutation Invariant Training
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SDR	Signal to Distortion Ratio
SiSNR	Scale independent SNR
SiSDR	Scale-invariant Signal-to-Distortion Ratio
SIR	Signal to Interference Ratio
SE	Speech Enhancement
SNR	Signal to Noise Ratio
SOTA	State-Of-The-Art
STFT	Short-time Fourier Transform
STOI	Short-Time Objective Intelligibility
TasNet	Time-domain audio separation Network
TCN	Temporal Convolution Network
TF	Time-Frequency
TS-ASR	Target Speaker ASR
TSE	Target Speech Extraction
TS-VAD	Target Speaker VAD
VAD	Voice Activity Detection
VAE	Variational Autoencoder
WER	Word Error Rate
WSJ	Wall Street Journal

Copyright 2021 NTT CORPORATION, BUT

1. Introduction

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

Introduction

- Cocktail party problem and target speech extraction
- Notations
- Relation between TSE and speech enhancement tasks
- Datasets, toolkits and evaluation metrics

Copyright 2021 NTT CORPORATION, BUT

Introduction

- **Cocktail party problem and target speech extraction**
- Notations
- Relation between TSE and speech enhancement tasks
- Datasets, toolkits and evaluation metrics

Copyright 2021 NTT CORPORATION, BUT

Cocktail party-effect

Humans can focus their attention intentionally on a specific sound signal (Selective hearing)

Realized using various clues [Darwin+00]

- Locational,
- Speaker voice characteristics,
- Visual,
- Content

→ Can follow a conversation at a cocktail party, pick-up our name, etc.

Copyright 2021 NTT CORPORATION, BUT

Target speech extraction (TSE)

Computational selective hearing

i.e. Extract speech of a target speaker in a mixture given speaker clues

Copyright 2021 NTT CORPORATION, BUT

Target speech extraction (TSE)

Computational selective hearing

i.e. Extract speech of a target speaker in a mixture given speaker clues

Copyright 2021 NTT CORPORATION, BUT

Classical ways to tackle TSE

- Fixed beamformer
  - Extract signal from a fixed direction
  - ☹ Requires knowing the position of the target speaker
  - → Lack of flexibility

- Separation
  - Separate mixture into all its source signals
  - ☹ Requires knowing/estimating number of speakers
  - ☹ Speaker-output ambiguity
  - → Need to be combined with some speaker identification
  - ☹ Cascade Separation+ speaker identification is not optimal for TSE task

Copyright 2021 NTT CORPORATION, BUT

Classical ways to tackle TSE

- Fixed beamformer
  - Extract signal from a fixed direction
  - ☹ Requires knowing the position of the target speaker
  - → Lack of flexibility

- Separation
  - Separate mixture into all its source signals
  - ☹ Requires knowing/estimating number of speakers
  - ☹ Speaker-output ambiguity
  - → Need to be combined with some speaker identification
  - ☹ Cascade separation+ speaker identification is not optimal for TSE task

Copyright 2021 NTT CORPORATION, BUT

Advantages of TSE

By exploiting speaker clues, TSE avoids the limitation of previous schemes

- ☺ **Does not require knowing the speaker location[1]**
- ☺ **No estimation of nb of speakers required**
- ☺ **No speaker-output ambiguity**
- ☺ **Optimal for TSE task**

- TSE made possible recently thanks to progress in speech enhancement/separation, speaker identification & video processing
- Especially, deep-learning[2] enabled optimized TSE system
  - 2017, showed possibility with audio clues [Zmolikova17]
  - 2018, showed possibility with video clues [Afouras+18, Ephrat+18,Owens +18]

  Since then,
  - Rapid progress following development of neural speech enhancement/separation/speaker identification
  - Extension to new tasks (ASR, diarization etc.)

Target speaker clues

[1] Some approaches discussed in section 3 use speaker location as clue
[2] This tutorial focuses on neural network-based approaches

Copyright 2021 NTT CORPORATION, BUT

Demo video

Demo of audio-clue-based TSE (SpeakerBeam)

(1) Record enrollment
(2) TSE demo: man + woman
(3) TSE demo: man + man + music

Available on YouTube:
https://www.youtube.com/watch?v=7FSHgKip6vI

Copyright 2021 NTT CORPORATION, BUT

Innovative R&D by NTT

SpeakerBeam: Paying attention to the speaker you want to listen to
- Computational selective hearing based on deep learning -

NTT Communication Science Laboratories

Copyright©2018 NTT corp. All Rights Reserved.

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Voice recorder
- Teleconference

Target-Speaker ASR

- Smartphones
- Smart speakers

Meeting ASR

- Meeting recognition
- Minute generation

Copyright 2021 NTT CORPORATION, BUT

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Voice recorder
- Teleconference

Target-Speaker ASR

- Smartphones
- Smart speakers

Meeting ASR

- Meeting recognition
- Minute generation

Copyright 2021 NTT CORPORATION, BUT

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Voice recorder
- Teleconference

Target-Speaker ASR

- Smartphones
- Smart speakers

Meeting ASR

- Meeting recognition
- Minute generation

Copyright 2021 NTT CORPORATION, BUT

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Voice recorder
- Teleconference

Target-Speaker ASR

- Smartphones
- Smart speakers

Meeting ASR

- Meeting recognition
- Minute generation

Copyright 2021 NTT CORPORATION, BUT

Introduction

- Cocktail party problem and target speech extraction
- **Notations**
- Relation between TSE and speech enhancement tasks
- Datasets, toolkits and evaluation metrics

Copyright 2021 NTT CORPORATION, BUT

Speech mixture

Speech recorded at a distant microphone contains

- Target speaker
- Reverberation
- Background noise
- Interfering speakers

- Neglect reverberation

Copyright 2021 NTT CORPORATION, BUT

Notations - Signal model

Speech mixture

$$y^m[t] = x_s^m[t] + \sum_{\substack{i=1,\\ i\neq s}}^{I} x_i^m[t] + v^m[t]$$

$$\mathbf{y}^m \in \mathbb{R}^{1\times T}$$

$y^m[t]$	mixture signal at microphone m
$x_i^m[t]$	i-th source signal (include reverberation)
$v^m[t]$	noise signal at microphone m
m	microphone index
t	time index
T	duration of the mixture
I	number of speakers
s	index of target speaker

In case of single channel we omit microphone index m

Copyright 2021 NTT CORPORATION, BUT

Notations - Signal model

- Signal model in STFT domain

$$Y^m[n,f] = X_s^m[n,f] + \sum_{\substack{i=1,\\ i\neq s}}^{I} X_i^m[n,f] + V^m[n,f]$$

n frame index
f frequency index
F number of frequency bins

Copyright 2021 NTT CORPORATION, BUT

Notations - Signal model

- Signal model in STFT domain

Copyright 2021 NTT CORPORATION, BUT

Notations – Speaker clues

- Inform the system about the target with speaker clue $\mathbf{c}_s$

$c^{(a)}[t]$ *or* $\mathbf{c}_s^{(a)} \in \mathbb{R}^{1\times T_a}$

$\mathbf{c}^{(v)}[t]$ *or* $\mathbf{C}_s^{(v)} \in \mathbb{R}^{D^v\times T}$

Copyright 2021 NTT CORPORATION, BUT

Introduction

- Cocktail party problem and target speech extraction
- Notations
- **Relation between TSE and speech enhancement tasks**
- Datasets, toolkits and evaluation metrics

Copyright 2021 NTT CORPORATION, BUT

Speech enhancement

- Estimate clean signal by removing interferences
  - Background noise
    → Noise reduction
  - Interfering speakers
    - Speech separation
    - TSE

- Introduce neural-based single-microphone approaches (except in Part 3).

Copyright 2021 NTT CORPORATION, BUT

Noise reduction

Copyright 2021 NTT CORPORATION, BUT

Noise reduction (*Single channel*)

- Remove (non-speech) noise from noisy speech
- A lot of research on single channel noise reduction
  - Spectral subtraction [Boll+79]
  - Wiener filter [Lim+79]
  - Time-frequency (TF) masking [Wang+06]
  - NMF [Virtanen+07]
  - **Neural network-based enhancement**
    - Frequency-domain approaches,
      - Regression-based approach [Xu+15]
      - **Mask-based approach** [Narayanan+13, Weninger+14]
    - Time-domain approaches [Pascual+17, Macartney+18]

Copyright 2021 NTT CORPORATION, BUT

Mask-based approach [Narayanan+13, Weninger+14]

- Learn a TF mask that indicates where speech is dominant over noise

$$\mathbf{M} = \begin{cases} 1 & \text{if speech} > noise \\ 0 & \text{otherwise} \end{cases}$$

- Speech and noise have different spectral characteristics
  → NN can learn to distinguish both

Typical loss function (MSE)

$$\mathcal{L}(\theta) = \left|\mathbf{X} - \hat{\mathbf{X}}\right|^2$$

Clean speech

Estimated speech

Copyright 2021 NTT CORPORATION, BUT

Separation

Copyright 2021 NTT CORPORATION, BUT

Speech separation

- Separate a speech mixture into all its sources

- Various separation methods
  - Independent Component analysis (ICA) [Hyvarinen+04]
  - Non-negative Matrix Factorization (NMF) [Virtanen+07, Smaragdis'17]
  - Mask-based methods
    - Spatial clustering [Sawada+11, Mandel+10]
    - Neural-network-based approaches
      - Speaker/Gender/Dominance dependent approaches [Weng+15, Wang+16]
      - Deep clustering/Deep attractor network [Hershey+16, Chen+17]
      - **Permutation invariant training** [Yu+17, Kolbæk+17]

Copyright 2021 NTT CORPORATION, BUT

Speech sparseness assumption

- Speech signals are sparse, and rarely overlap each other

→ To extract a speaker from a mixture, sufficient to know which TF bins the speaker is dominant (TF mask) and apply this mask to the mixture [Yilmaz+04]

Copyright 2021 NTT CORPORATION, BUT

Permutation invariant training (PIT)

Kolbæk et al. “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE/ACM Trans. ASLP, 2017.

- Separation network predicts a mask for each source
- Training loss:

$$\mathcal{L} = \left|\mathbf{X}_1 - \hat{\mathbf{X}}_1\right|^2 + \left|\mathbf{X}_2 - \hat{\mathbf{X}}_2\right|^2$$

Source 1 Output 1 Source 2 Output 2

  - Source ambiguity at the output

→ PIT Loss

  - Use the minimum of the losses computed over all source permutations

$$\mathcal{L} = \min\left(\left|\mathbf{X}_1 - \hat{\mathbf{X}}_1\right|^2 + \left|\mathbf{X}_2 - \hat{\mathbf{X}}_2\right|^2; \left|\mathbf{X}_1 - \hat{\mathbf{X}}_2\right|^2 + \left|\mathbf{X}_2 - \hat{\mathbf{X}}_1\right|^2\right)$$

Copyright 2021 NTT CORPORATION, BUT

Demo of PIT

$\hat{\mathbf{X}}_1$

$\hat{\mathbf{X}}_2$

mixture

Copyright 2021 NTT CORPORATION, BUT

Time domain separation: TasNet

Luo et al. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM trans. ASLP, 2019.

- Input and output are time-domain signals
- Replace STFT by learnable encoder/decoder layers
- Implemented with convolutional layers → Conv-TasNet

Training loss :
PIT-loss in the time domain

$$\mathcal{L} = \min \begin{pmatrix} \mathrm{SDR}(\mathbf{x}_1, \hat{\mathbf{x}}_1) + \mathrm{SDR}(\mathbf{x}_2, \hat{\mathbf{x}}_2) \\ \mathrm{SDR}(\mathbf{x}_1, \hat{\mathbf{x}}_2) + \mathrm{SDR}(\mathbf{x}_2, \hat{\mathbf{x}}_1) \end{pmatrix}$$

SDR = Signal to distortion ratio

$$\mathrm{SDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log \frac{\|\mathbf{x}\|^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2}$$

Copyright 2021 NTT CORPORATION, BUT

Various configurations

- Various architectures
  - Fully-Connected (FC), BLSTM, CNNs
  - Temporal convolution (TCN)/Conv-TasNet/DPRNN
  - Transformer

- Various losses
  - MSE, CE of masks
  - Time-domain loss such as SDR/SNR or SiSDR/SiSNR

TSE can borrow architectures/losses developed for separation

Conv-TasNet:

Copyright 2021 NTT CORPORATION, BUT

Separation performance

- WSJ0-2mix benchmark [Hershey+16]
  - 2-speaker mixtures from WSJ
  - Widely used benchmark for single-channel neural speech separation

		SDR [dB] (Higher the better)
PIT (BLSTM)	[Kolbæk+17]	10.0
TasNet (Conv)	[Luo+19]	15.6
TasNet (DPRNN)	[Luo+20]	19.0
Wavesplit	*[Zeghidour+20]*	*20.4*
SepFormer	*[Subakan+20]*	*22.3*

https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix

Copyright 2021 NTT CORPORATION, BUT

TSE by separation + identification

- Speaker ambiguity at the output of separation
  - Which output corresponds to the target speaker?

  → Requires *speaker identification* post-processing

- Drawbacks of this scheme

  ☹ Cannot exploit speaker clue to improve separation

  ☹ Not jointly optimized (*accumulate errors of each module*)

  ☹ Requires knowing or estimating the number of speakers

Copyright 2021 NTT CORPORATION, BUT

TSE vs separation

TSE uses a single model

- Realize separation and identification at once
- Potentially more challenging problem but,
  - ☺ Can exploit speaker clue for “extraction”
  - ☺ Jointly optimized → better performance

(WSJ0 2mix)

SDR[dB]

Network configuration	Separation (w/ oracle identification)	TSE
BLSTM (Freq.)	9.2	**9.7**
Conv-TasNet (Time)	16.3	**17.2**

Copyright 2021 NTT CORPORATION, BUT

TSE vs noise reduction

Similar to noise reduction: 1 output

☺ No speaker ambiguity

☺ No need to estimate number of speakers in the mixture

Target speaker clues

	#Speakers	Target ambiguity	Speaker clue
Noise reduction	1	No	No
Separation	> 1 (known number)	Yes ☹	No
TSE	**> 1 (unknown number)**	**No ☺**	**Needed**

Copyright 2021 NTT CORPORATION, BUT

Introduction

- Cocktail party problem and target speech extraction
- Notations
- Relation between TSE and speech enhancement tasks
- **Datasets, toolkits & evaluation metrics**

Copyright 2021 NTT CORPORATION, BUT

Datasets (Publicly available)

There is no common dataset for TSE…

but, we can use most datasets used for separation

- Audio only
  - Simulated mixtures
    - WSJ 2mix [Hershey+16]
    - WHAM [Wichern+19]
    - Librimix [Cosentino+20]
  - Real (or rerecorded) mixtures
    - LibriCSS [Chen+20]
    - CHiME 5/6 [Barker+18]
    - AMI [Carletta+05]
    - DNS challenge Task 2[Reddy+21]
- Audio-visual
  - Lipreading sentences (LRS) [Chung+17]
  - Grid corpus [Cooke+06]
  - AVSpeech [Ephrat+18]

Simulated mixture generation

- Add clean waveforms with gains to vary the Signal to interference ratio (SIR)
- *Eventually add noise and reverberation*

For TSE

- Randomly select another utterance as enrollment

Copyright 2021 NTT CORPORATION, BUT

Toolkits

- No official implementation of most speech extraction approaches but can be easily implemented by modifying separation recipes
  - See discussion end of part 2
  - https://github.com/BUTSpeechFIT/speakerbeam
- Useful toolkits
  - Asteroid (Enhancement, pyTorch)
    - https://github.com/asteroid-team/asteroid
  - ESPnet (E2E ASR, Enhancement, pyTorch)
    - https://github.com/espnet/espnet
  - Kaldi (ASR, i-vector, x-vector, C++)
    - https://github.com/kaldi-asr/kaldi
  - Padertorch (Enhancement, pyTorch)
    - https://github.com/fgnt/padertorch
  - SpeechBrain (ASR, Enhancement, pyTorch)
    - https://github.com/speechbrain/speechbrain

Copyright 2021 NTT CORPORATION, BUT

Evaluation metrics

- Speech quality (higher the better)
  - Signal-to-Distortion Ratio (SDR)/SNR Signal-to-Noise Ratio (SNR)
    - $\mathrm{SDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log \frac{\|\mathbf{x}\|^2}{\|\mathbf{x}-\hat{\mathbf{x}}\|^2}$
  - Scale-invariant SDR (SiSDR)/SiSNR*
    - $\mathrm{SiSDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log \frac{\|\alpha^* \mathbf{x}\|^2}{\|\alpha^* \mathbf{x}-\hat{\mathbf{x}}\|^2}$, where $\alpha^* = \underset{\alpha}{\operatorname{argmin}} \|\alpha \mathbf{x} - \hat{\mathbf{x}}\|^2 = \frac{\hat{\mathbf{x}}^{\mathrm{T}} \mathbf{x}}{\mathbf{x}^{\mathrm{T}} \mathbf{x}}$
  - BSS eval SDR*
    - $\mathrm{SDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log \frac{\|\boldsymbol{\alpha}^* \mathbf{x}\|^2}{\|\boldsymbol{\alpha}^* \mathbf{x}-\hat{\mathbf{x}}\|^2}$, where $\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{\alpha} * \mathbf{x} - \hat{\mathbf{x}}\|^2$
  - Perceptual Evaluation of Speech Quality (PESQ)
- Speech intelligibility (higher the better)
  - Short-Time Objective Intelligibility (STOI)
- ASR (lower the better)
  - Word error rate (WER)

*SDR/SI-SDR can be computed using *BSS eval* toolkit
But we should disable optimal permutation computation

Copyright 2021 NTT CORPORATION, BUT

References (1/3)

[Afouras+18] Afouras et al. “The conversation: Deep audio-visual speech enhancement,” Interspeech, 2018.

[Barker+18] Barker et al. "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," Interspeech, 2018.

[Boll+79] Boll, S. “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. ASSP, 1979.

[Carletta+05] Carletta et al. “The AMI meeting corpus: A pre-announcement,” Springer, 2005.

[Chen+17] Chen et al. “Deep attractor network for single-microphone speaker separation,” ICASSP, 2017.

[Chen+20] Chen et al., "Continuous speech separation: dataset and analysis," ICASSP, 2020

[Chung+17] Chung et al. "Lip reading sentences in the wild," CVPR, 2017.

[Cooke+06] Cooke et al. "An audio-visual corpus for speech perception and automatic speech recognition," JASA, 2006.

[Cosentino+20] Cosentino et al. "Librimix: An open-source dataset for generalizable speech separation," arXiv, 2020.

[Darwin+00] Darwin et al. “Effectiveness of spatial cues, prosody, and talker characteristics in selective attention,” JASA, 2000

[Ephrat+18] Ephrat et al. “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. on Graphics, 2018.

[Hershey+16] Hershey et al. “Deep clustering: Discriminative embeddings for segmentation and separation,” ICASSP, 2016.

[Hyvarinen+04] Hyvarinen et al. “Independent component analysis,” John Wiley & Sons, 2004.

Copyright 2021 NTT CORPORATION, BUT

References (2/3)

[Kolbæk+17] Kolbæk et al. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM trans. ASLP, 2017.

[Lim+79] Lim et al. “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE, 1979.

[Luo+19] Luo et al. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM trans. ASLP, 2019.

[Luo+20] Luo et al. "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," ICASSP, 2020.

[Macartney+18] Macartney et al. "Improved speech enhancement with the wave-u-net," arXiv, 2018.

[Mandel+10] Mandel et al. “Model-based expectation-maximization source separation and localization,” IEEE Trans. ASLP, 2010.

[Narayanan+13] Narayanan et al. “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” Proc. ICASSP, 2013.

[Owens +18] Owens et al. “Audio-visual scene analysis with self-supervised multisensory features,” CoRR, 2018.

[Pascual+17] Pascual et al. "SEGAN: Speech enhancement generative adversarial network." Interspeech, 2017.

[Reddy+21] Reddy et al. "Icassp 2021 deep noise suppression challenge," ICASSP, 2021.

[Sawada+11] Sawada et al. “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” IEEE Trans. ASLP, 2011.

[Smaragdis’17] Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” IEEE Trans. ASLP, 2007.

Copyright 2021 NTT CORPORATION, BUT

References (3/3)

[Subakan+20] Subakan et al. "Attention is all you need in speech separation." ICASSP, 2021.

[Virtanen'07] Virtanen "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. ASLP, 2007.

[Wang+06] Wang et al. "Computational auditory scene analysis: Principles, algorithms, and applications," Hoboken, NJ: Wiley/IEEE Press, 2006.

[Wang+16] Wang et al., "Unsupervised single channel speech separation via deep neural network for different gender mixtures," APSIPA, 2016.

[Weng+15] Weng et al. "Deep neural networks for single-channel multi-talker speech recognition," IEEE Trans. ASLP, 2015.

[Weninger+14] Weninger et al. "The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement," Proc. REVERB, 2014.

[Wichern+19] Wichern et al. "WHAM!: Extending Speech Separation to Noisy Environments," Interspeech, 2019.

[Xu+15] Xu et al. "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. ASLP, 2015.

[Yilmaz+04] Yilmaz et al. "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. SP, 2004.

[Yu+17] Yu et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," ICASSP, 2017.

[Zeghidour+20] Zeghidour et al. "Wavesplit: End-to-end speech separation by speaker clustering," IEEE/ACM IEEE/ACM trans. ASLP, 2021.

[Zmolikova+17] Zmolikova et al. "Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures," Interspeech, 2017.

Copyright 2021 NTT CORPORATION, BUT

2. Target speech extraction: Audio clue-based approaches

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

Audio-clue based extraction

Reminder of the task

Inputs: mixture, audio clue

Output: speech of target speaker

Audio clue: short *enrollment utterance* of the target speaker (e.g. pre-recorded)

Content of the section

1. Aspects of TSE models
2. Examples of existing models

Copyright 2021 NTT CORPORATION, BUT

Aspects of TSE models

1. How to extract the **speaker embeddings**?
2. How to **inform the neural network** using the speaker embedding?
3. Which **input and output domain** to use?
4. Which **loss functions** to train with?
5. What **neural network architecture** to use?

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

Goal: extract a vector containing information about target speaker necessary to identify them in the mixture

Ways to represent speaker information

a. i-vectors
b. NN-based (e.g. x-vectors)

} Used in speaker verification

c. jointly learned embeddings

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

i-vectors [Dehak+10]

- For a long time SOTA in speaker verification
- Often used for adaptation in ASR
- Contains information about both speaker and channel
- Extraction based on a generative model

1. Training of Universal Background Model

2. Fitting GMM on input utterance constrained by UBM

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

i-vectors [Dehak+10]

- For a long time SOTA in speaker verification
- Often used for adaptation in ASR
- Contains information about both speaker and channel
- Extraction based on a generative model

1. Training of Universal Background Model

2. Fitting GMM on input utterance constrained by UBM

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

i-vectors [Dehak+10]

- For a long time SOTA in speaker verification
- Often used for adaptation in ASR
- Contains information about both speaker and channel
- Extraction based on a generative model

1. Training of Universal Background Model

2. Fitting GMM on input utterance constrained by UBM

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

NN-based (x-vectors, d-vectors) [Snyder+18,Wan+18]

- Neural network-based embeddings
- Hidden representations in NN trained for speaker classification

From [Snyder+17]

☺ Highly speaker discriminative

☺ Pre-trained models on large corpora

☹ Trained for a different task

youtube.com/watch?v=AkCPHw2m6bY

Copyright 2021 NTT CORPORATION, BUT

Speaker embeddings

Jointly learned embeddings [Vesely+16]

- Previous embeddings learned with different objective than target speech extraction
- "Optimal" embeddings – learned jointly with the task
- Shown to learn speaker information

Adjusted to the task

Data with enough speaker variability necessary

From [Zmolikova+19]

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Goal: Use the speaker representation to guide the neural network to extract the target speaker

Ways to inform the neural network

a. Concatenation-based
b. Factorized layer
c. Multiplication-based
d. Attention-based

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Position of the adaptation layer varies, commonly:

- First layer
- One of early layers
- All layers

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Concatenation-based e.g. [Wang+19, Xu+19]

- Concatenating embedding to the input of adaptation layer
- Can be seen as modification of bias
- For fully-connected (FC) layer:

$$\mathbf{h}^{l+1} = \sigma(\mathbf{W}^{l}\mathbf{h}^{l} + \underbrace{\mathbf{W}^{(\mathrm{e})}\mathbf{e} + \mathbf{b}^{l}}_{\text{modified bias}})$$

☺ Simple

☹ Modifying bias only might not be strong enough

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Factorized layer [Zmolikova+17a]

- Adaptation split into several sub-layers
- Speaker embedding used to weigh combination of outputs of sub-layers
- Modifies both weight and bias
- For fully-connected (FC) layer:

$$\mathbf{h}^{l+1} = \sigma\left(\sum_i e_i(\mathbf{W}^{l,i}\,\mathbf{h}^l + \mathbf{b}^{l,i})\right)$$

☺ Strong influence of embedding

☹ Memory-expensive

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Multiplication-based [Delcroix+19a]

- Activations multiplied with the speaker embedding
- Can be seen as scaling columns of the weight matrix
- For fully-connected (FC) layer:

$$\mathbf{h}^l = \sigma(\mathbf{e} \cdot (\mathbf{W}^l \mathbf{h}^l + \mathbf{b}^l))$$

☺ Simple and strong

☹ Same modification for all time-frames

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Multiplication-based: FiLM [Perez+18]

- Activations multiplied and summed with the speaker embedding
- Can be seen as scaling columns of the weight matrix and adding bias
- For fully-connected (FC) layer:

$$\mathbf{h}^l = \sigma(\mathbf{e}^{(1)} \cdot (\mathbf{W}^l \mathbf{h}^l + \mathbf{b}^l) + \mathbf{e}^{(2)})$$

☺ Simple and strong

☹ Same modification for all time-frames

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Attention-based [Xiao+19, Li+19]

- Previous schemes apply the same transformation to every frame of the mixed speech

- **Attention schemes** use dynamic information in the enrollment to apply different transformation to every frame of the mixed speech

Copyright 2021 NTT CORPORATION, BUT

Informing the neural network

Attention-based [Xiao+19, Li+19]

- Previous schemes apply the same transformation to every frame of the mixed speech

- **Attention schemes** use dynamic information in the enrollment to apply different transformation to every frame of the mixed speech

Not directly applicable with i- or x-vectors

Copyright 2021 NTT CORPORATION, BUT

Input/output domain

How to represent speech signal at input and output of the neural network?

a. Frequency-domain

b. Time-domain

Copyright 2021 NTT CORPORATION, BUT

Input/output domain

a. Frequency-domain

- Log-magnitude or magnitude of Short-time Fourier transform
- Windows typically around 30-100 ms

b. Time-domain [Luo+19]

- Raw signal
- First convolutional layer of NN acts as encoder
- Windows typically around 1-5 ms

From [Luo+19]

Copyright 2021 NTT CORPORATION, BUT

Loss function

How to train parameters of the network?

a. Frequency-domain

b. Time-domain

c. Speaker-id losses

Similar loss functions as for speech separation, but *does not need permutation invariant training*.

Copyright 2021 NTT CORPORATION, BUT

Loss function

Frequency-domain

- Direct estimation of enhanced magnitude

Mean square error:

$$L_{mse}(\theta) = \left\| |\widehat{\boldsymbol{X}}| - |\boldsymbol{X}| \right\|^2$$

- Mask estimation

Mask cross-entropy:

$$L_{ce}(\theta) = M_{t,f} \log\left(\widehat{M}_{t,f}\right) + \left(1 - M_{t,f}\right)\log(1 - \widehat{M}_{t,f})$$

Masked mean square error:

$$L_{mmse}(\theta) = \left\| \widehat{\boldsymbol{M}} \cdot |\boldsymbol{Y}| - |\boldsymbol{X}| \right\|^2$$

Copyright 2021 NTT CORPORATION, BUT

Loss function

Time-domain

Scale invariant signal-to-noise ratio

$$L_{si-s}(\theta) = 10\log_{10}\frac{\|\boldsymbol{x}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2}$$

$$\boldsymbol{x}_{target} = \frac{<\hat{\boldsymbol{x}}, \boldsymbol{x}> \boldsymbol{x}}{\|\boldsymbol{x}\|^2}$$

$$\boldsymbol{e}_{noise} = \hat{\boldsymbol{x}} - \boldsymbol{x}_{target}$$

Copyright 2021 NTT CORPORATION, BUT

Loss function

Speaker-id loss

- When embedding are jointly learned, explicit speaker loss may improve speaker discrimination
- Different loss functions explored
  - Categorical cross-entropy loss [Ge+20,Delcroix+20]
  - Large margin cosine loss [Ji+20]
  - Triplet loss [Ji+20,Mun+20]

Copyright 2021 NTT CORPORATION, BUT

Architectures

What neural network architecture to choose for main and auxiliary network?

a. BLSTM

b. Conv-TasNet

c. DPRNN

Copyright 2021 NTT CORPORATION, BUT

Architectures

(B)LSTM

- Mostly in early works
- Stack of several Long short-term memory blocks
- Not very efficient in modelling long-time dependencies → not often used for time-domain models

From https://colah.github.io/posts/2015-08-Understanding-LSTMs/

FC: fully connected

Copyright 2021 NTT CORPORATION, BUT

Architectures

Conv-TasNet [Luo+19]

- Used mostly with time-domain signals
- Stack of convolutional “repetitions”
- 1 repetition = series of convolutions with increasing dilation
- Receptive field ~1.5 second (depends)

Copyright 2021 NTT CORPORATION, BUT

Architectures

Dual Path RNN [Luo+20]

- Can model long-range dependencies better than Conv-Tasnet
- Alternating inter- and intra-chunk blocks
- Target speaker information applied after intra+inter chunk block or before DPRNN

Copyright 2021 NTT CORPORATION, BUT

Examples of techniques and results

a. SpeakerBeam
b. VoiceFilter
c. Speaker inventory

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam

- Originated in [Zmolikova+17a]
  - Idea of using speaker embedding to extract target speaker's speech
  - Compared to training a separate neural network for each speaker
- Extensive study in [Zmolikova+19]
  - Comparison of different methods for TSE
  - Comparison with speech separation methods (deep clustering, PIT)
  - Integration with multi-channel approaches, ASR training

Speaker embeddings: jointly learned

Informing style: factorized layer, multiplication

Input/output domain: frequency, time

Loss function: masked MSE, SI-SNR

Architecture: BLSTM, Conv-TasNet

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam

Delcroix et al. “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” ICASSP, 2020.

- Time-domain SpeakerBeam
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: SDR

	FF	MM	FM	avg
Mixture	0.17	0.16	0.16	0.16
TasNet (oracle)	*8.68*	*9.75*	*12.14*	*10.84*
TasNet (xvect)	4.59	4.93	11.44	8.35
FD-SpkBeam	5.19	5.32	10.27	7.94
TD-SpkBeam	9.13	9.47	12.77	11.17

FF: female-female MM: male-male FM: female-male

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam

Delcroix et al. “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” ICASSP, 2020.

- Time-domain SpeakerBeam
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: SDR

	FF	MM	FM	avg
Mixture	0.17	0.16	0.16	0.16
TasNet (oracle)	*8.68*	*9.75*	*12.14*	*10.84*
TasNet (xvect)	4.59	4.93	11.44	8.35
FD-SpkBeam	5.19	5.32	10.27	7.94
TD-SpkBeam	9.13	9.47	12.77	11.17

FF: female-female MM: male-male FM: female-male

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam

Delcroix et al. “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” ICASSP, 2020.

- Time-domain SpeakerBeam
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: SDR

	FF	MM	FM	avg
Mixture	0.17	0.16	0.16	0.16
TasNet (oracle)	*8.68*	*9.75*	*12.14*	*10.84*
TasNet (xvect)	4.59	4.93	11.44	8.35
FD-SpkBeam	5.19	5.32	10.27	7.94
TD-SpkBeam	9.13	9.47	12.77	11.17

FF: female-female MM: male-male FM: female-male

Copyright 2021 NTT CORPORATION, BUT

VoiceFilter

Wang et al. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech, 2019.

Copyright 2021 NTT CORPORATION, BUT

VoiceFilter

Wang et al. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech, 2019.

Copyright 2021 NTT CORPORATION, BUT

VoiceFilter

Wang et al. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech, 2019.

Copyright 2021 NTT CORPORATION, BUT

VoiceFilter

Wang et al. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech, 2019.

Architecture:

Layer	Width		Dilation		Filters / Nodes
	time	freq	time	freq	
CNN 1	1	7	1	1	64
CNN 2	7	1	1	1	64
CNN 3	5	5	1	1	64
CNN 4	5	5	2	1	64
CNN 5	5	5	4	1	64
CNN 6	5	5	8	1	64
CNN 7	5	5	16	1	64
CNN 8	1	1	1	1	8
LSTM	-	-	-	-	400
FC 1	-	-	-	-	600
FC 2	-	-	-	-	600

Copyright 2021 NTT CORPORATION, BUT

VoiceFilter

Wang et al. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech, 2019.

- **Training data**
  - Speaker encoder: Google in-house + VoxCeleb + LibriSpeech
  - VoiceFilter: VCTK / LibriSpeech
- **Results** (mixtures from LibriSpeech)

Target speech extraction (SDR)

Target speech recognition (WER)

Copyright 2021 NTT CORPORATION, BUT

Speaker inventory

Xiao et al. “Single-channel speech extraction using speaker inventory and attention network,” ICASSP, 2019.

- Assumes audio data of potential competing speakers can be available
- Informing the network with attention-based mechanism

- **Results** (mixtures from LibriSpeech, SDR)

Copyright 2021 NTT CORPORATION, BUT

Conclusions

- Target speech extraction can be realized by providing an audio clue
- Combination of methods from
  - *Speaker adaptation* (informing the NN)
  - *Speaker verification* (speaker embeddings)
  - *Speech separation/enhancement* (architecture, loss)
- Current challenges
  - Discrimination of similar voices
  - Robustness to noisy/reverberant conditions
- A lot of existing work
  - Deep extractor [Wang+18], SpEx [Ge+20], SpeakerFilter [He+21], ...
  - Compilation of references

	year	speaker embedding	informing style	domain	loss	
Learning speaker representation for neural network based multichannel speaker extraction	2017	jointly learned	factorized layer	frequency	mask cross entropy	B
Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures	2017	one-hot, posteriors	factorized layer	frequency	mask cross entropy	B
Deep Extractor Network for Target Speaker Recovery From Single Channel Speech Mixtures	2018	jointly learned	concatenation	frequency	masked MSE	L
Multi-Channel Overlapped Speech Recognition with Location Guided Speech Extraction Network	2018	N/A (location)	concatenation	frequency	masked MSE	L
Single Channel Target Speaker Extraction and Recognition with Speaker Beam	2018	jointly learned	factorized layer	frequency	mask cross entropy, ASR loss	B
Single-channel Speech Extraction Using Speaker Inventory and Attention Network	2019	jointly learned	attention, concatenation	frequency	masked MSE	B
Dynamic-attention based Encoder-decoder model for Speaker Extraction with Anchor speech	2019	jointly learned	attention, concatenation	frequency	MSE	L
VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking	2019	d-vector	concatenation	frequency	masked MSE, masked L1	C
Dynamic-attention based Encoder-decoder model for Speaker Extraction with Anchor speech	2019	jointly learned	concatenation	frequency	MSE	A
Optimization of Speaker Extraction Neural Network with Magnitude and Temporal Spectrum Approx	2019	jointly learned	concatenation	frequency	masked MSE, phase-sensitive, delta	B
Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information	2019	N/A (location)	concatenation	frequency	masked MSE	L
TEnet: target speaker extraction network with accumulated speaker embedding for automatic spee	2019	x-vector	concatenation	frequency	masked MSE	C
A Study on Online Source Extraction in the Presence of Changing Speaker Positions	2019	jointly learned	factorized layer	frequency	mask cross entropy	L
A Unified Framework for Neural Speech Separation and Extraction	2019	jointly learned	factorized layer, attention	frequency	masked MSE	B
Target Speaker Extraction for Overlapped Multi-Talker Speaker Verification	2019	jointly learned	factorized layer, concatenation	frequency	masked phase-sensitive MSE, delta	B
End-to-End SpeakerBeam for Single Channel Target Speech Recognition	2019	jointly learned	multiplication	frequency	ASR loss	B
Compact Network for Speakerbeam Target Speaker Extraction	2019	jointly learned	multiplication, factorized layer	frequency	masked phase-sensitive MSE	B
SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures	2019	jointly learned, i-vector	multiplication, factorized layer, c	frequency	masked, phase-sensitive MSE, ASR	B
Direction-Aware Speaker Beam for Multi-Channel Speaker Extraction	2019	jointly learned	multiplication, spatial attention	frequency	masked MSE	B
Time-Domain Speaker Extraction Network	2019	i-vector	concatenation	time	SI-SNR	T
Multi-channel block-online source extraction based on utterance adaptation	2019	jointly learned	factorized layer	frequency	mask cross entropy	L
The Sound of My Voice: Speaker Representation Loss for Target Voice Separation	2020	d-vector	concatenation	frequency	masked MSE; speaker loss	C
A Pitch-aware Speaker Extraction Serial Network	2020	jointly learned	concatenation	frequency	masked MSE, phase-sensitive, delta	B
Deep Ad-hoc Beamforming Based on Speaker Extraction for Target-Dependent Speech Separation	2020	jointly learned	concatenation	frequency	masked phase-sensitive MSE, delta	B
Speakerfilter: Deep Learning-Based Target Speaker Extraction Using Anchor Speech	2020	jointly learned	concatenation	frequency	masked-MSE	C

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam implementation

https://github.com/BUTSpeechFIT/speakerbeam

Contains:

- Time-domain SpeakerBeam model
- Alternative adaptation layers (FiLM, concatenation)
- LibriMix dataset preparation for TSE (includes map from mixtures to enrollment utterances)
- LibriMix recipe

Based on Asteroid toolkit

Copyright 2021 NTT CORPORATION, BUT

SpeakerBeam implementation

https://github.com/BUTSpeechFIT/speakerbeam

mix enroll estimated

jupyter SpeakerBeam_demo_notebook Last Checkpoint: 18 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [23]:
```
import matplotlib.pyplot as plt
import numpy as np
from pathlib import Path
import yaml
import torch
import random
import IPython.display as dsp

from models.td_speakerbeam import TimeDomainSpeakerBeam
from datasets.librimix_informed import LibriMixInformed
```

In [10]:
```
confpath = Path('../egs/libri2mix/exp/train_patience4/conf.yml')
modelpath = Path('../egs/libri2mix/exp/train_patience4/checkpoints/epoch=129-step=602289.ckpt')
```

In [12]:
```
with open(confpath) as f:
    conf = yaml.safe_load(f)
```

In [17]:
```
model = TimeDomainSpeakerBeam(**conf["filterbank"],
                              **conf["masknet"],
                              **conf["enroll"])
ckpt = torch.load(modelpath, map_location = torch.device('cpu'))

state_dict = {}
for k in ckpt["state_dict"]:
    state_dict[k.split('.',1)[1]] = ckpt["state_dict"][k]
model.load_state_dict(state_dict)
```

Out[17]: <All keys matched successfully>

In [18]:
```
test_set = LibriMixInformed(
    csv_dir = '../egs/libri2mix/data/wav8k/min/test',
    task = 'sep_noisy',
    sample_rate = 8000
```

Copyright 2021 NTT CORPORATION, BUT

References (1/2)

[Dehak+10] Dehak et al. "Front-end factor analysis for speaker verification," IEEE Trans. ASLP, 2010.

[Delcroix+15]Delcroix et al. "Context adaptive deep neural networks for fast acoustic model adaptation," ICASSP, 2015.

[Delcroix+19a] Delcroix et al. "Compact Network for Speakerbeam Target Speaker Extraction," ICASSP, 2019.

[Delcroix+20] Delcroix et al. "Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam," ICASSP, 2020.

[Ge+20] Ge et al. "SpEx+: A Complete Time Domain Speaker Extraction Network," Interspeech, 2020.

[Gemelo+07] Gemello et al. "Linear hidden transformations for adaptation of hybrid ANN/HMM models," Speech Communication, 2007.

[He+21] He et al. "Speakerfilter-Pro: an improved target speaker extractor combines the time domain and frequency domain," arxiv, 2021.

[Chen+17] Chen et al. "Deep attractor network for single-microphone speaker separation," ICASSP, 2017.

[Ji+20] Ji et al. "Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction," ICASSP, 2020.

[Li+19] Li et al. "Dynamic-attention based Encoder-decoder model for Speaker Extraction with Anchor speech," APSIPA, 2019.

[Luo+19] Luo et al. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM trans. ASLP, 2019.

[Luo+20] Luo et al. "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," ICASSP 2020.

[Mun+20] Mun et al. "The sound of my voice: Speaker representation loss for target voice separation," ICASSP, 2020.

[Perez+18] Perez et al. "Film: Visual reasoning with a general conditioning layer," AAAI, 2018.

Copyright 2021 NTT CORPORATION, BUT

References (2/2)

[Saon+13] Saon et al. "Speaker adaptation of neural network acoustic models using i-vectors," ASRU, 2013.

[Snyder+17] Snyder et al. "Deep neural network embeddings for text-independent speaker verification," Interspeech, 2017.

[Snyder+18] Snyder et al. "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP, 2018.

[Vesely+16] Vesely et al. "Sequence summarizing neural network for speaker adaptation," ICASSP, 2016.

[Wan+18] Wan et al. "Generalized end-to-end loss for speaker verification," ICASSP, 2018.

[Wang+18] Wang et al. "Deep Extractor Network for Target Speaker Recovery from Single Channel Speech Mixtures," Interspeech, 2018.

[Wang+19] Wang et al. "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," Interspeech, 2019.

[Xiao+19] Xiao et al. "Single-channel speech extraction using speaker inventory and attention network," ICASSP, 2019.

[Xu+19] Xu et al. "Optimization of Speaker Extraction Neural Network with Magnitude and Temporal Spectrum Approximation Loss," ICASSP, 2019.

[Zmolikova+17a] Zmolikova et al. "Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures," Interspeech, 2017.

[Zmolikova+17b] Zmolikova et al. "Learning speaker representation for neural network based multichannel speaker extraction," ASRU, 2017.

[Zmolikova+19] Zmolikova et al. "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," IEEE JSTSP, 2019.

Copyright 2021 NTT CORPORATION, BUT

3. Target speech extraction: Multi-channel approaches

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

Multi-channel approaches

Possibility to use spatial position of the sources to better extract the target speech.

Ways to use spatial information

1. Mask-based beamforming
2. Spatial features
3. Fixed beamforming + selection
4. Location-guided methods

Spatial information about target speaker is part of enrollment

Copyright 2021 NTT CORPORATION, BUT

Multi-channel approaches

Possibility to use spatial position of the sources to better extract the target speech.

Ways to use spatial information

1. **Mask-based beamforming**
2. Spatial features
3. Fixed beamforming + selection
4. Location-guided methods

Spatial information about target speaker is part of enrollment

Copyright 2021 NTT CORPORATION, BUT

Neural mask-based beamforming

Proposed by [Heymann+16], [Erdogan+16]

Copyright 2021 NTT CORPORATION, BUT

Neural mask-based beamforming

Commonly used beamformers:

- Minimum variance distortion-less beamformer (MVDR) [Frost'72]
- Multi-channel Wiener filter (MWF) [Doclo+02]

MVDR: $\boldsymbol{w}_f = \dfrac{\boldsymbol{R}_{(v),f}^{-1}\boldsymbol{R}_{(s),f}\boldsymbol{i}}{\mathrm{tr}(\boldsymbol{R}_{(v),f}^{-1}\boldsymbol{R}_{(s),f})}$

[Souden+09]

$\boldsymbol{R}_{(v),f}$ cross-power spectral density of interference

$\boldsymbol{R}_{(s),f}$ cross-power spectral density of target speech

$$\boldsymbol{R}_{(v),f} = \frac{1}{T}\sum_t (1 - M_{t,f})\boldsymbol{y}_{t,f}^H \boldsymbol{y}_{t,f} \qquad \boldsymbol{R}_{(s),f} = \frac{1}{T}\sum_t M_{t,f}\boldsymbol{y}_{t,f}^H \boldsymbol{y}_{t,f}$$

Copyright 2021 NTT CORPORATION, BUT

Neural mask-based beamforming

Neural network input: single-channel signal

Neural network target: ideal binary mask

$$\mathrm{IBM}_s(f,t) = \begin{cases} 1, & |X_s[f,t]| > |X_{j\neq s}[f,t]| \\ 0, & \text{otherwise} \end{cases}$$

Loss function: binary cross-entropy

Alternatives: time-domain versions [Ochiai+20]

differentiation through beamforming [Heymann+17]

Copyright 2021 NTT CORPORATION, BUT

Neural mask-based beamforming

Results of automatic speech recognition on MC-WSJ0-2mix in terms of Word Error rate using multi-channel recordings and beamforming

Copyright 2021 NTT CORPORATION, BUT

Multi-channel approaches

Possibility to use spatial position of the sources to better extract the target speech.

Ways to use spatial information

1. Mask-based beamforming
2. **Spatial features**
3. Fixed beamforming + selection
4. Location-guided networks

Spatial information about target speaker is part of enrollment

Copyright 2021 NTT CORPORATION, BUT

Spatial features

- In mask-based beamforming, NN processes *single-channel* signal

☺ No dependence on microphone array

☹ No usage of spatial information

- Spatial features from multi-channel signal can improve estimated masks

! Network may over-rely on spatial features and deteriorate for small angle cases [Chen+19]

What spatial features to use?

1. Fixed spatial features (IPD, ILD)
2. Learned spatial features

Copyright 2021 NTT CORPORATION, BUT

Spatial features

1. Fixed spatial features

- **interaural phase difference (IPD)** $\mathrm{IPD}(t,f) = \theta_y^1(t,f) - \theta_y^2(t,f)$
- interaural level difference (ILD) $\mathrm{ILD}(t,f) = |Y^1(t,f)| - |Y^2(t,f)|$
- generalized cross-correlation (GCC) $\mathrm{GCC}(t,f,\tau) = \cos(\mathrm{IPD}(t,f) - \frac{2\pi f}{N}\tau)$

From [Wang+18]

Copyright 2021 NTT CORPORATION, BUT

Spatial features

2. Learned spatial features

- Convolutional layer estimating spatial features from pair of channels

- [Zorilă+21] reports substantial SDR improvements to IPD in *anechoic case*, limited improvement for *reverberant* case on MC-WSJ0-2mix

Copyright 2021 NTT CORPORATION, BUT

Multi-channel approaches

Possibility to use spatial position of the sources to better extract the target speech.

Ways to use spatial information

1. Mask-based beamforming
2. Spatial features
3. **Fixed beamforming + selection**
4. Location-guided networks

Spatial information about target speaker is part of enrollment

Copyright 2021 NTT CORPORATION, BUT

Fixed beamformers

Li et al. "Direction-Aware Speaker Beam for Multi-Channel Speaker Extraction," Interspeech, 2019.

1. Apply set of fixed beamformers in a grid
2. Weigh different directions with attention derived from enrollment utterance.
3. Apply single-channel target speech extraction.

Copyright 2021 NTT CORPORATION, BUT

Multi-channel approaches

Possibility to use spatial position of the sources to better extract the target speech.

Ways to use spatial information

1. Mask-based beamforming
2. Spatial features
3. Fixed beamforming + selection
4. **Location-guided networks**

Spatial information about target speaker is part of enrollment

Copyright 2021 NTT CORPORATION, BUT

Location guided

- Assuming enrollment knowledge about the location of target speaker
- Can be in form of
  - direction of arrival (obtained e.g. from visual information)
  - multi-channel enrollment signal recorded in the same spatial position

1. Target location features

2. Spatial pre-processing

Copyright 2021 NTT CORPORATION, BUT

Location guided

1. Target location features

Chen et al. "Multi-channel overlapped speech recognition with location guided speech extraction network," SLT, 2018.
Gu et al. "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," Interspeech, 2019.

Directional power ratio

fixed beamformer in direction θ

$$\mathrm{DPR}_\theta(t,f) = \frac{\left\|\mathbf{w}_\theta^H(f)\mathbf{Y}(t,f)\right\|_2^2}{\sum_k \left\|\mathbf{w}_k^H(f)\mathbf{Y}(t,f)\right\|_2^2}$$

fixed beamformers in a grid

Copyright 2021 NTT CORPORATION, BUT

Location guided

Heitkaemper et al. "A Study on Online Source Extraction in the Presence of Changing Speaker Positions," Springer SLSP, 2019.
Martín-Doñas et al. "Multi-Channel Block-Online Source Extraction Based on Utterance Adaptation," Interspeech 2019.

2. Spatial pre-processing

1. Estimate beamformer from the enrollment utterance (e.g. MVDR).
2. Process both mixture and enrollment using the beamformer.
3. Apply target speech extraction using the beamformed outputs.

Copyright 2021 NTT CORPORATION, BUT

Conclusions

- Spatial information can improve the quality of the extraction
- Location-guided approaches w/ knowledge of target position: *location features / spatial pre-processing*
- Approaches w/o knowledge of target position

Mask-based beamforming — *Fixed beamformers* — *Spatial features*

Independence to microphone positions ⟷ **Utilization of spatial information by NN**

Copyright 2021 NTT CORPORATION, BUT

References (1/2)

[Doclo+02] Doclo et al. "GSVD-based optimal filtering for single and multimicrophone speech enhancement," IEEE transaction SP, 2002.

[Erdogan+16] Erdogan et al. "Improved mvdr beamforming using single-channel mask prediction network," Interspeech, 2016.

[Frost'72] Frost, O.L. "An algorithm for linearly constrained adaptive array processing," IEEE Proceedings, 1972.

[Gu+19] Gu et al. "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," Interspeech, 2019.

[Heitkaemper+19] Heitkaemper et al. "A Study on Online Source Extraction in the Presence of Changing Speaker Positions," Springer SLSP, 2019.

[Heymann+16] Heymann et al. "Neural network based spectral mask estimation for acoustic beamforming," ICASSP, 2016.

[Heymann+17] Heymann et al. "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," ICASSP, 2017.

[Chen+18] Chen et al. "Multi-channel overlapped speech recognition with location guided speech extraction network," SLT, 2018.

Copyright 2021 NTT CORPORATION, BUT

References (2/2)

[Chen+19] Chen et al. "Multi-band pit and model integration for improved multi-channel speech separation," ICASSP, 2019.

[Li+19] Li et al. "Direction-Aware Speaker Beam for Multi-Channel Speaker Extraction," Interspeech, 2019.

[Martín-Doñas+19] Martín-Doñas et al. "Multi-Channel Block-Online Source Extraction Based on Utterance Adaptation," Interspeech, 2019.

[Ochiai+20] Ochiai et al. "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," ICASSP, 2020.

[Souden+09] Souden, "On optimal frequency-domain multichannel linear filtering for noise reduction," IEEE Transactions ASLP, 2009.

[Subramanian+20] Subramanian et al. "Far-Field Location Guided Target Speech Extraction Using End-to-End Speech Recognition Objectives," ICASSP, 2020.

[Wang+18] Wang et al. "Combining spectral and spatial features for deep learning based blind speaker separation," IEEE/ACM Transactions ASLP, 2018.

[Zorilă+21] Zorilă et al. "An Investigation into the Multi-channel Time Domain Speaker Extraction Network," SLT, 2021.

Copyright 2021 NTT CORPORATION, BUT

4. Target speech extraction: Visual/multimodal approaches

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

Target speech extraction (TSE)

- Extract speech of a target speaker in a mixture

Copyright 2021 NTT CORPORATION, BUT

Audio-visual speech processing

- Visual information investigated for many problems of speech processing
  - Lip-reading/audio-visual ASR
  - Audio-visual voice activity detection
  - Source localization
  - Speech enhancement
- We focus on extraction from speech mixtures
  - Early works showed that visual information could help separation [Hershey+01, Casanovas+10]
  - We focus on recent deep learning-based approaches [Michelsanti+21]

Copyright 2021 NTT CORPORATION, BUT

Visual-clue-based TSE

TSE with visual clues (video of the target speaker speaking in the mixture)

- Studies in neuroscience: visual clues help humans focusing their auditory attention on a particular source [Golumbic+13]

☺ Visual information is not affected by acoustic noise

☺ Visual information can help discriminate speakers with similar voice characteristics

Copyright 2021 NTT CORPORATION, BUT

Classification of approaches

- 3 categories of methods using different clues

Copyright 2021 NTT CORPORATION, BUT

Video-clue-based TSE

Copyright 2021 NTT CORPORATION, BUT

Video-clue based approaches

- Similar architecture than audio-clue-based approaches [Gabbay+18]
- Speaker clue is video
  - Time series that is aligned with the audio!

$$\mathbf{c}^{(v)}[t] \; or \; \mathbf{C}_s^{(v)} \in \mathbb{R}^{D^v \times T}$$

Copyright 2021 NTT CORPORATION, BUT

Video encoder

- Key part is the video encoder
  - Video pre-processing
  - Video embedding extraction

Copyright 2021 NTT CORPORATION, BUT

Video pre-processing

1. Find faces in the video
   - Use off-the-shelf face detector/tractor e.g. Viola-Johns [Viola+04]

2. Select the face image of the target speaker
   - Application dependent and usually assumed to be known in presented works

3. Crop the video to include only the face or lip regions

Copyright 2021 NTT CORPORATION, BUT

Video embedding

- Convert the sequence of images $\mathbf{c}_s^{(v)}[t]$ into a sequence of video embedding vectors $\mathbf{e}_s^{(v)}[t] \in \mathbb{R}^{D^v}$
- Embedding vectors should capture information relevant to extract the target speaker, e.g.
  - Mouth opening/closure regions
  - Finer lip movements that could identify phones

- Video embedding
  - Learned jointly with the extraction network from raw videos
  - Derived from intermediate representation

Copyright 2021 NTT CORPORATION, BUT

Jointly learned embeddings from raw videos

Gabbay et al. “Visual speech enhancement,” Interspeech, 2018.
Hou et al. “Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks,” IEEE Trans. ETCI, 2018.

- Train the extraction system end-to-end
  - i.e. learn parameters of video encoder from scratch directly from raw videos (after some normalization)
  - Showed promising initial results for speaker closed systems (i.e. *trained and tested on same speaker*)
  - Hard to extend to speaker open
    - May need large amount of data to learn general and meaningful video embedding

    → Most speaker open approaches use intermediate representations

Copyright 2021 NTT CORPORATION, BUT

Embedding from intermediate representation

- Exploit pre-trained system to generate meaningful embedding
  - Leverage large amount of data from video/image tasks
  - More robust to variations (face, resolution, luminosity, etc.)
- Add transformation layers to map pre-trained embeddings to useful representation for speech extraction
- 2 categories
  - Face landmarks [Hou+16, Morrone+19]
  - Neural embedding pre-trained on different task
    - From face recognition [Ephrat+18]
    - From lip-reading [Afouras+18]
    - From video synchronization [Owens +18]

Copyright 2021 NTT CORPORATION, BUT

Embedding from intermediate representation

- Exploit pre-trained system to generate meaningful embedding
  - Leverage large amount of data from video/image tasks
  - More robust to variations (face, resolution, luminosity, etc.)
- Add transformation layers to map pre-trained embeddings to useful representation for speech extraction
- 2 categories
  - **Face landmarks [Hou+16, Morrone+19]**
  - Neural embedding pre-trained on different task
    - From face recognition [Ephrat+18]
    - From lip-reading [Afouras+18]
    - From video synchronization [Owens +18]

Copyright 2021 NTT CORPORATION, BUT

Face landmark

Hou et al. "Audio-visual speech enhancement using deep neural networks," APSIPA, 2016.
Morrone et al. “Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments,” ICASSP, 2019.

- Face landmarks represent position of key points in a face
- Can use the x-y coordinate of the face landmarks and/or derived motion (1-st order derivative)

☺ Fast to compute and is interpretable

☺ Used for speaker open TSE even with small corpus such as Grid corpus

😐 It remains to be tested for larger tasks

Copyright 2021 NTT CORPORATION, BUT

Embedding from intermediate representation

- Exploit pre-trained system to generate meaningful embedding
  - Leverage large amount of data from video/image tasks
  - More robust to variations (face, resolution, luminosity, etc.)
- Add transformation layers to map pre-trained embeddings to useful representation for speech extraction
- 2 categories
  - Face landmarks [Hou+16, Morrone+19]
  - **Neural embedding pre-trained on different task**
    - From face recognition [Ephrat+18]
    - From lip-reading [Afouras+18]
    - From video synchronization [Owens +18]

Copyright 2021 NTT CORPORATION, BUT

Neural embedding – face recognition

Ephrat et al. “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. Graphics, 2018.

- Embeddings from face recognition NN such as FaceNet [Schroff+15]
  - Representation not directly related to acoustic content
    - But intermediate representation of FaceNet contains information about face expression (mouth)

      → Use intermediate layer to keep face information
  - Trained on static face images with id of person
    - ☹ Does not learn lip movement dynamics
    - ☺ Easy to get large amount of training data

$$\boldsymbol{e}_s^{(v)}[t]$$

https://towardsdatascience.com/a-facenet-style-approach-to-facial-recognition-dc0944efe8d1

Copyright 2021 NTT CORPORATION, BUT

Neural embedding – lip reading

Afouras et al. “The conversation: Deep audio-visual speech enhancement,” Interspeech, 2018.

- Embeddings from lip reading network trained to predict words or phones [Stafylakis+17]
  - Representation directly related to acoustic content
  - Trained on video
    - ☺ Learns lip movements dynamics
  - Requires video and associated transcriptions for training
    - ☹ More involved to get large amount of training data

Word/Phones prediction

$\boldsymbol{e}_s^{(v)}[t]$

Copyright 2021 NTT CORPORATION, BUT

Neural embedding – synchronization

Owens et al. “Audio-visual scene analysis with self-supervised multisensory features,” CoRR, 2018.

- Embeddings derived from NN that predicts synchronization between audio and video [Owens+18]
  - Related to acoustic content
  - Trained on video
    - ☺ Learns lip movements dynamics
  - Self-supervised training (no transcriptions needed)
    - ☺ Easy to get large amount of training data

Audio-video aligned or not

$\boldsymbol{e}_s^{(v)}[t]$

Copyright 2021 NTT CORPORATION, BUT

Fusion/Adaptation layer

- Video embeddings are time-series
  - Different sampling frequency between audio and video
  - → Up-sample video [Ephrat+18] or down-sample audio [Afouras+18]
- Can use similar fusion strategies than for audio clues
  - Concatenation [Afouras+18, Ephrat+18]
  - Multiplication [Ochiai+19]
  - Factorized layer [Gu+20]

Copyright 2021 NTT CORPORATION, BUT

Overall network architecture

- Video encoder
  - Typically implemented with convolutional layers
- Audio encoder/extraction net
  - Similar architecture than for audio-clue based approaches
    - Frequency domain [Afouras+18, Ephrat+18]
    - Time-domain (Conv-TasNet) [Wu+19, Pan+21, Sato+21]

Copyright 2021 NTT CORPORATION, BUT

Generating simulated training data

- Generating Training data
  - Sample videos from target speaker and a different speaker in the database
  - Mix the audio
  - Use video as target speaker clue

Copyright 2021 NTT CORPORATION, BUT

Looking to Listen

Ephrat et al. “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. Graphics, 2018.

Detailed network configuration

- Video embedding w/ FaceNet
- Fusion: Concatenation
- Frequency domain extraction

Good extraction performance

	SDR
Male-Male	9.7
Female-Female	10.6
Male-Female	10.5

Copyright 2021 NTT CORPORATION, BUT

Sound Demo from *Looking to Listen*

Mixture

Extracted left speaker

https://looking-to-listen.github.io/supplemental/index.html

Copyright 2021 NTT CORPORATION, BUT

Characteristics of video-clue-based TSE

- Strengths
  - Performance competitive with audio-based approaches
  - Particularly advantageous when speakers have similar voice characteristics

Double Brady

From [Ephrat+18]

- Issues
  - Synchronization between audio and video [Lee+21]
  - Privacy concerns from the use of video
    - → Still-image-based approach [Chung+20, Qu+20]
  - Occlusion
    - → Multimodal clues [Afouras+19, Ochiai+19]

Copyright 2021 NTT CORPORATION, BUT

Image-clue-based TSE

Copyright 2021 NTT CORPORATION, BUT

Face informed

Chung et al. “FaceFilter: Audio-visual speech separation using still images,” Interspeech, 2020.
Qu et al. “Multimodal target speech separation with voice and face references,” Interspeech, 2020.

- Use still image of target speaker instead of video

  → Similar to audio-clue-based extraction, but with speaker embedding derived from face picture

- Recent research shows that possible to learn joint embeddings between face and voice [Nagrani+18, Kim+18, Oh+19]
  - Identify voice from face with up to 80% accuracy

From [Nagrani+18]

- Use pre-train face encoder
  - Trained for face recognition [Qu+20]
  - Trained to predict audio/face association [Chung+20]

Copyright 2021 NTT CORPORATION, BUT

Results from *FaceFilter*

Chung et al. “FaceFilter: Audio-visual speech separation using still images,” Interspeech, 2020.

- Network configuration

Figure 1: *Description of the proposed audio-visual speech separation network. Blue blocks are trainable neural networks whereas the red blocks are pre-trained in cross-modal biometric task.*

From [Chung+20]

- SDR improvement
  - Mixture generated from LRS2 data

Mixture	SDRi
Seen-Heard Speaker	
Male-Male	2.084 dB
Male-Female	3.039 dB
Female-Male	3.821 dB
Female-Female	1.930 dB
Unseen-Unheard Speaker	
Male-Male	1.256 dB
Male-Female	3.066 dB
Female-Male	3.830 dB
Female-Female	1.198 dB

From [Chung+20]

- Improvement of 1 to 4 dB
- Challenging for same-gender mixtures

Copyright 2021 NTT CORPORATION, BUT

Multimodal clues-based TSE

Copyright 2021 NTT CORPORATION, BUT

Advantage of multimodal clues

- Using multiple clues can help make the system more robust to clue corruptions
  - Occlusion (Visual-clue)
  - Noise in enrollment (Audio-clue)
  - Mixture of similar voices (Audio-clue)
  - Errors in DOA estimation (Locational clue)
- Several clues investigated
  - Video + audio
    - [Afouras+19, Ochiai+19, Luo+19, Sato+21, Pan+21]
  - Face + audio
    - [Qu+20]
  - Video + audio + location
    - [Gu+20]

Copyright 2021 NTT CORPORATION, BUT

Advantage of multimodal clues

- Using multiple clues can help make the system more robust to clue corruptions
  - Occlusion (Visual-clue)
  - Noise in enrollment (Audio-clue)
  - Mixture of similar voices (Audio-clue)
  - Errors in DOA estimation (Locational clue)
- Several clues investigated
  - **Video + audio**
    - [Afouras+19, Ochiai+19, Luo+19, Sato+21, Pan+21]
  - Face + audio
    - [Qu+20]
  - Video + audio + spatial
    - [Gu+20]

Copyright 2021 NTT CORPORATION, BUT

Audio-visual clue-based approach

Afouras et al. “My lips are concealed: Audio-visual speech enhancement through obstructions,” Interspeech, 2019.
Ochiai et al. “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” Interspeech, 2019.

- Combine both audio and video embeddings
- Can use same architecture than audio/visual approaches

Copyright 2021 NTT CORPORATION, BUT

Multimodal clue fusion

Afouras et al. “My lips are concealed: Audio-visual speech enhancement through obstructions,” Interspeech, 2019.

- Concatenate video and audio embeddings [Afouras+19]

Copyright 2021 NTT CORPORATION, BUT

Multimodal clue fusion

Ochiai et al. “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” Interspeech, 2019.
Sato et al. “Multimodal Attention Fusion for Target Speaker Extraction,” SLT 2021.

- Attention fusion [Ochiai+19]
  - “Select” embedding based on clue reliability with weighted sum
  - Compute attention weights over clues, $a^{\psi}[t]$
  - Concatenation is similar to attention fusion with $a^{\psi}[t] = 0.5$
- Guiding attention weights during training helps better clue selection [Sato+21]

$$\mathbf{e}^{AV}[t] = \sum_{\psi \in (A,V)} a^{\psi}[t]\mathbf{e}_s^{\psi}[t]$$

Copyright 2021 NTT CORPORATION, BUT

Results for multimodal SpeakerBeam

Sato et al. “Multimodal Attention Fusion for Target Speaker Extraction,” SLT 2021.

- Simulated mixtures from LRS3

From [Sato+21]

Demo samples are available at
http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html

Copyright 2021 NTT CORPORATION, BUT

Results for multimodal SpeakerBeam

Sato et al. “Multimodal Attention Fusion for Target Speaker Extraction,” SLT 2021.

- Simulated mixtures from LRS3

From [Sato+21]

Demo samples are available at
http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html

Copyright 2021 NTT CORPORATION, BUT

Results for multimodal SpeakerBeam

Sato et al. “Multimodal Attention Fusion for Target Speaker Extraction,” SLT 2021.

- Simulated mixtures from LRS3
- Using multiple clues makes TSE robust to clue corruptions

From [Sato+21]

Demo samples are available at
http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html

Copyright 2021 NTT CORPORATION, BUT

Conclusions

- Extraction possible with video clue
  - What information is needed? (Fine lip movements OR speaker activity)
- Combining speaker clues can make systems robust
  - Audio and visual
  - More clues also investigated
- Some issues
  - Dominance of a modality during training
    - Training on mixture of same speaker to force using the video information[Gabbay+17]
    - Remove modality during training [Ochiai+19]
    - Reconstruct lip image [Hou+18]
- Further reading: extended overview of audiovisual speech enhancement and separation [Michelsanti+21]

TSE with Audio, Visual and Locational clues

From [Gu+20]

Copyright 2021 NTT CORPORATION, BUT

References (1/2)

[Afouras+18] Afouras et al. “The conversation: Deep audio-visual speech enhancement,” Interspeech, 2018.

[Afouras+19] Afouras et al. “My lips are concealed: Audio-visual speech enhancement through obstructions,” Interspeech, 2019.

[Casanovas+10] Casanovas et al. “Blind audiovisual source separation based on sparse redundant representations,” IEEE Trans. Multimedia, 2010.

[Chung+20] Chung et al. “FaceFilter: Audio-visual speech separation using still images,” Interspeech, 2020.

[Ephrat+18] Ephrat et al. “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. Graphics, 2018.

[Gabbay+18] Gabbay et al. “Visual speech enhancement,” Interspeech, 2018.

[Golumbic+13] Golumbic et al. “Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a “Cocktail Party,” The Journal of Neuroscience, 2013.

[Gu+20] Gu et al. “Multi-modal multi-channel target speech separation,” IEEE Journal of Selected Topic in Signal Processing, 2020.

[Hershey+01] Hershey et al. “Audio-visual sound separation via hidden Markov models,” NIPS, 2001.

[Hou+16] Hou et al. "Audio-visual speech enhancement using deep neural networks," APSIPA, 2016.

[Hou+18] Hou et al. “Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks,” IEEE Trans. ETCI, 2018.

[Kim+18] Kim et al. “On learning associations of faces and voices,” ACCV, 2018.

[Lee+21] Lee et al. “Looking Into Your Speech: Learning Cross-Modal Affinity for Audio-Visual Speech Separation,” CVPR, 2021.

Copyright 2021 NTT CORPORATION, BUT

References (1/2)

[Luo+19] Luo et al. “Audio-visual speech separation using i-Vectors,” ICICSP, 2019.

[Michelsanti+21] Michelsanti et al. “An overview of deep-learning-based audiovisual speech enhancement and separation,” IEEE/ACM Trans. ASLP, 2021.

[Morrone+19] Morrone et al. “Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments,” ICASSP, 2019.

[Nagrani+18] Nagrani et al. "Seeing voices and hearing faces: Cross-modal biometric matching," CVPR, 2018.

[Ochiai+19] Ochiai et al. “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” Interspeech, 2019.

[Oh+19] Oh et al. “Speech2Face: Learning the face behind a voice,” in Proc. of CVPR, 2019.

[Owens +18] Owens et al. “Audio-visual scene analysis with self-supervised multisensory features,” CoRR, 2018.

[Pan+21] Pan et al., “Muse: Multi-modal target speaker extraction with visual cues,” ICASSP, 2021.

[Qu+20] Qu et al. “Multimodal target speech separation with voice and face references,” Interspeech, 2020.

[Sato+21] Sato et al. “Multimodal Attention Fusion for Target Speaker Extraction,” SLT 2021.

[Schroff+15] Schroff et al. "Facenet: A unified embedding for face recognition and clustering," CVPR, 2015.

[Stafylakis+17] Stafylakis et al. “Combining Residual Networks with LSTMs for Lipreading,” Interspeech, 2017.

[Viola+04] Viola et al. “Robust Real-Time Face Detection,” IJCV, 2004.

[Wu+19] Wu et al. “Time domain audio visual speech separation,” ASRU, 2019.

Copyright 2021 NTT CORPORATION, BUT

5. Other tasks

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

5.1 Application to diarization

Copyright 2021 NTT CORPORATION, BUT

Diarization task

- Detecting who spoke when
- Could be tackled by target speech extraction + voice activity detection (VAD)

- This is attempting to solve harder task than necessary → instead, directly predict target speaker activity

Copyright 2021 NTT CORPORATION, BUT

Personal VAD

Ding et al. "Personal VAD: Speaker-conditioned voice activity detection," Odyssey, 2020.

Voice activity detection: speech VS silence

Personal voice activity detection: target speech VS non-target speech VS silence

- Motivation: activating personal devices only when target speaker speaks
  lightweight system
- Informed by d-vector of the target speaker

Classification accuracy

Method					Network parameters (million)
	tss	ns	ntss	mean	
VAD+SV	0.777	0.908	0.768	0.801	4.88 (SV) + 0.06 (VAD)
PVAD	0.916	0.883	0.920	0.912	**0.13** (PVAD)

Copyright 2021 NTT CORPORATION, BUT

Target speaker VAD

Medennikov et al. "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," Interspeech, 2020.

- Winner of CHiME-6 challenge (highly distorted and overlapped data)
- Single-speaker model + Multi-speaker post-processing
- i-vector as speaker information
- Fixed number of speakers (extension to unknown number in [He+21])

Results (CHiME-6, DER)

Copyright 2021 NTT CORPORATION, BUT

Audio-visual VAD

Tao et al. "Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection," ACM Multimedia, 2021.

TalkNet

- **Inputs**: cropped face video, audio
- **Outputs**: binary classification of target speaker activity
- **Network**: audio encoder, visual encoder, audio-visual cross attention

Other works at https://github.com/TaoRuijie/TalkNet_ASD/blob/main/awesomeASD.md

Copyright 2021 NTT CORPORATION, BUT

References

[Ding+2020] Ding et al. "Personal VAD: Speaker-conditioned voice activity detection," Odyssey, 2020.

[Medennikov+20] Medennikov et al. "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," Interspeech, 2020.

[Tao+21] Tao et al. "Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection," ACM Multimedia, 2021.

Copyright 2021 NTT CORPORATION, BUT

5.2 Application to ASR

Copyright 2021 NTT CORPORATION, BUT

Target speech processing for ASR

Target Speaker ASR (TS-ASR)

- Transcribe speech of the target speaker
- Use enrollment utterance

Anchored ASR

- Transcribe speech of the speaker that spoke a *wakeup word*
- Use wakeup word to get target speaker clue

Meeting ASR

- Transcribe speech of each speaker
- Use speaker clues for each speaker

Copyright 2021 NTT CORPORATION, BUT

ASR: Hybrid vs End-to-end (E2E)

Hybrid (DNN-HMM) [Hinton+12, Yu+15]

- Estimate word sequence $\hat{\mathbf{W}}$ as

$$\hat{\mathbf{W}} = argmax_{\mathbf{W}}\, p(\mathbf{X}|\mathbf{W})p(\mathbf{W})$$

$p(\mathbf{X}|\mathbf{W})$: Acoustic model (AM), use a DNN to map speech features $\mathbf{X}$ to HMM state posterior

$p(\mathbf{W})$: Language model (LM)

E2E [Graves+14, Chan+15, Bahdanau+16]

- Estimate word sequence $\hat{\mathbf{W}}$ as

$$\hat{\mathbf{W}} = argmax_{\mathbf{W}}\, p(\mathbf{W}|\mathbf{X})$$

Neural network maps directly speech to characters/word sequence

- Encoder-decoder w/ attention, CTC, Transformers, RNN-transducers

Copyright 2021 NTT CORPORATION, BUT

Target speaker ASR (TS-ASR)

Copyright 2021 NTT CORPORATION, BUT

TS-ASR

With explicit (signal) extraction

[Delcroix+18, Zmolikova+18, Delcroix+19, Subramanian+20]

- Cascade connection of TSE and ASR modules

Without explicit extraction

[King+17, Delcroix+18, Kanda+19, Delcroix+19, Denisov+19]

- Single module directly transcribe speech of the target w/o explicit extraction of the extracted speech signal

**Can also be done with video clues [Chao+16, Yu+21, Pan+21], but this is not covered in this tutorial*

Copyright 2021 NTT CORPORATION, BUT

TS-ASR

With explicit (signal) extraction

[Delcroix+18, Zmolikova+18, Delcroix+19, Subramanian+20]

- Cascade connection of TSE and ASR modules

Without explicit extraction

[King+17, Delcroix+18, Kanda+19, Delcroix+19, Denisov+19]

- Single module directly transcribe speech of the target w/o explicit extraction of the extracted speech signal

Can also be done with video clues [Chao+16, Yu+21, Pan+21], but this is not covered in this tutorial

Copyright 2021 NTT CORPORATION, BUT

TS-ASR w/ explicit extraction

Zmolikova et al. “Optimization of speaker-aware multichannel speaker extraction with ASR criterion,” ICASSP, 2018
Delcroix et al. “Single channel target speaker extraction and recognition with speaker beam,” ICASSP, 2018.
Wang et al. “Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition,” Interspeech, 2020.

TSE (1-ch/multi-ch)

[Zmolikova+18, Delcroix+18, Delcroix+19]

- Provides access to the extracted speech signal

TSE in the feature domain

[Wang+20]

- Online processing/Low latency

Copyright 2021 NTT CORPORATION, BUT

TS-ASR w/ explicit extraction

Zmolikova et al. “Optimization of speaker-aware multichannel speaker extraction with ASR criterion,” ICASSP, 2018
Delcroix et al. “Single channel target speaker extraction and recognition with speaker beam,” ICASSP, 2018.
Delcroix et al. “End-to-end SpeakerBeam for single channel target speech recognition,” Interspeech, 2019.

Interconnection with ASR to reduce effect of distortions from TSE front-end

- Training loss of TSE that avoids over-suppression [Wang+20]
- Process only overlapped speech [Wang+20, Sato+21]

- Retraining the ASR on speech processed with TSE front-end [Zmolikova+18]
- Joint-training: All front-end operations are differentiable and can thus be jointly optimized [Delcroix+18, Zmolikova+18]

Copyright 2021 NTT CORPORATION, BUT

TS-ASR

With explicit (signal) extraction

[Delcroix+18, Zmolikova+18, Delcroix+19, Subramanian+20]

- Cascade connection of TSE and ASR modules

Without explicit extraction

[King+17, Delcroix+18, Kanda+19, Delcroix+19, Denisov+19]

- Single module directly transcribes speech of the target w/o explicit extraction of the speech signal

**Can also be done with video clues [Chao+16, Yu+21, Pan+21], but this is not covered in this tutorial*

Copyright 2021 NTT CORPORATION, BUT

TS-ASR w/o explicit extraction

- Using speaker embedding as auxiliary input to AM for speaker adaptation[Saon+13]
  - Applied to single speaker **(Not a mixture)**
  - Typically, i-vectors are often used and can be computed from input speech
  - Enables adapting model to the speaker characteristics

→ Applying similar idea to realize TS-ASR

Copyright 2021 NTT CORPORATION, BUT

TS-ASR w/o explicit extraction

Delcroix et al. “Single channel target speaker extraction and recognition with speaker beam,” ICASSP, 2018.
Denisov et al., “End-to-End Multi-Speaker Speech Recognition using Speaker Embeddings and Transfer Learning,” Interspeech, 2019.
Kanda et al., “Auxiliary Interference Speaker Loss for Target-Speaker Speech Recognition,” Interspeech, 2019.

- Add auxiliary features to recognizer
  - Hybrid ASR
    - Embedding added to the input [King+17]
    - Embedding added within the network [Delcroix+18, Kanda+19]
  - E2E ASR
    - Input of the encoder [Denisov+19]
    - Intermediate layer of the encoder [Delcroix+19]
    - Within attention mechanism [Wang+19]
- Can also add interference speaker recognition branch & loss to improve speaker discrimination [Kanda+19]

Copyright 2021 NTT CORPORATION, BUT

Result for E2E TS-ASR

Delcroix et al. “End-to-end SpeakerBeam for single channel target speech recognition,” Interspeech, 2019.

- E2E ASR of WSJ0-2mix
- All systems jointly-trained from scratch

(1) Encoder-decoder E2E TS-ASR w/o explicit extraction

(2) E2E TS-ASR w/ explicit extraction (SpeakerBeam, Joint training)

Copyright 2021 NTT CORPORATION, BUT

Anchored ASR

Copyright 2021 NTT CORPORATION, BUT

Anchored ASR (1ch)

King et al., " Robust speech recognition via anchor word representations," Interspeech, 2017.
Wang et al. "End-to-end Anchored Speech Recognition," ICASSP, 2019.

- Transcribe speech of the speaker that spoke a wakeup word/anchor
  - Separate wakeup word and query
  - Use wakeup word to get target speaker clue
- Reported up to 35% relative WER reduction when command is corrupted with background and multimedia speech

Copyright 2021 NTT CORPORATION, BUT

Anchored ASR (multi-ch)

Kida et al. "Speaker Selective Beamformer with Keyword Mask Estimation," SLT, 2018.

- Recognize speech coming from the direction from where wakeup word was spoken
  - Wakeup separation use a DNN to compute a TF-mask that extract keyword in a mixture
  - Compute Beamformer filter coefficients from the wakeup speech signal
  - Apply beamforming on the query to extract the target
- Reported up to 27% relative WER reduction

Copyright 2021 NTT CORPORATION, BUT

Anchored ASR (1ch)

King et al., " Robust speech recognition via anchor word representations," Interspeech, 2017.
Wang et al. "End-to-end Anchored Speech Recognition," ICASSP, 2019.

- Transcribe speech of the speaker that spoke a wakeup word/anchor
  - Separate wakeup word and query
  - Use wakeup word to get target speaker clue
- Compute speaker characteristics on wakeup word
  - Opt1: Simply perform cepstral mean subtraction (CMS) using statistics computed from anchor (AMS)
  - Opt2: Use neural net to compute embedding and append to input of AM
  - Also investigated with E2E ASR [Wang+19]
- Reported up to 35% relative WER reduction when command is corrupted with background and multimedia speech

Copyright 2021 NTT CORPORATION, BUT

Anchored ASR (multi-ch)

Kida et al. "Speaker Selective Beamformer with Keyword Mask Estimation," SLT, 2018.

- Recognize speech coming from the direction from where wakeup word was spoken
  - Wakeup separation use a DNN to compute a TF-mask that extract keyword in a mixture
  - Compute Beamformer filter coefficients from the wakeup speech signal
  - Apply beamforming on the query to extract the target
- Reported up to 27% relative WER reduction

Copyright 2021 NTT CORPORATION, BUT

Meeting recognition

Copyright 2021 NTT CORPORATION, BUT

Meeting recognition

- Recognize and identify speech of all participants in a conversation [Waibel+98, Renals+07, Hori+12]

- Speaker clues can help solve the problem with different schemes
  - Clues from enrollment or from mixture
  - Parallel or joint-processing schemes

Copyright 2021 NTT CORPORATION, BUT

Clues from enrollment or from mixture

Clues from enrollment

Collect enrollment for all meeting participants [Zmolikova+21, Kanda+21]

☺ Can link transcriptions with speaker identities

Clues from mixture

Use diarization to estimate speaker activity

- Compute speaker embeddings [Kanda+19]
- Directly use as clue for TSE [Boeddecker+18, Delcroix+21]

☺ Does not require registering participants

Copyright 2021 NTT CORPORATION, BUT

Parallel or joint-processing schemes

Parallel processing scheme
[*Kanda+19,* Zmolikova+21]

Decode for each speaker independently
→ Complex

Joint-processing scheme
[Kanda+20]

Simultaneously process for all speakers
→ Can be more efficient

e.g. Speaker attributed ASR [Kanda+20]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes
- Use a speaker inventory
  - Contains speaker embeddings of meeting participants
  - Can include other speakers or all irrelevant speakers

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes
- Use a speaker inventory
  - Contains speaker embeddings of meeting participants
  - Can include other speakers or all irrelevant speakers
  - Target speaker embedding computed by attention mechanism using inventory and mixture

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes
- Use a speaker inventory
  - Contains speaker embeddings of meeting participants
  - Can include other speakers or all irrelevant speakers
  - Target speaker embedding computed by attention mechanism using inventory and mixture
- Speaker embeddings used
  - Associate ASR results with speakers (Speaker identification)

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Speaker attributed ASR

Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

- E2E system that output ASR results of all speakers in a sequence with tags identifying speaker changes
- Use a speaker inventory
  - Contains speaker embeddings of meeting participants
  - Can include other speakers or all irrelevant speakers
  - Target speaker embedding computed by attention mechanism using inventory and mixture
- Speaker embeddings used
  - Associate ASR results with speakers (Speaker identification)
  - Inform ASR decoder

From [Kanda+21]

Copyright 2021 NTT CORPORATION, BUT

Evaluation on real meetings

N. Kanda et al., “A comparative study of modular and joint approaches for speaker-attributed ASR on monaural long-form audio,” Submitted to ASRU 2021

- Results for AMI meeting recognition task
  - 3 to 5 participants
  - Single distant microphone

Diarization only (M1)

Diarization + Separation

Diarization + SA-ASR

Use speaker embeddings estimated from diarization

Copyright 2021 NTT CORPORATION, BUT

Conclusion

- Using target speaker clues can benefit ASR/diarization for overlapping speech conditions
  - TS-VAD is the state-of-the-art diarization method for CHiME 6
  - TS-ASR/Anchored ASR/SA-ASR are showing promising results
- TSE was also shown to be beneficial for other tasks such as speaker verification [Rao+19, Rikhye+21]
- Various systems configurations for TS-ASR
  - Multi-speaker ASR or TS-ASR
  - w/ or w/o explicit speech extraction
  - Hybrid or E2E
  - Using other modalities

Copyright 2021 NTT CORPORATION, BUT

TS-ASR vs Multi-speaker ASR

- Multi-speaker E2E ASR

  [Settle+18, Chang+18, Chang+19]

  - Includes separation capability into E2E ASR encoder

  ☺ Recognize all speakers in the mixture

  ☹ Difficult to identify which is the target speaker from text output

- For TS-ASR

  ☺ No speaker ambiguity at output

Copyright 2021 NTT CORPORATION, BUT

w/ or w/o explicit extraction?

w/ explicit extraction

w/o explicit extraction

Difficult to conclude what approach is better in terms of performance because of different configurations

- ☹ Potentially larger model
- ☺ Easier to deploy w/ existing ASR system
- ☺ Better interpretability (Can listen to extracted speech)

- ☺ Compact model
- ☹ More difficult to deploy (Need to modify the ASR module)
- ☹ Less interpretability

Copyright 2021 NTT CORPORATION, BUT

Hybrid or E2E

Hybrid systems

[Delcroix+18, Zmolikova+18, Kanda+19]

- Require mixtures and HMM-forced alignment of target speech
- Effectively require parallel mixture/clean speech
  - → Limit for simulated data

E2E systems

[Delcroix+19 , Denisov+19, Shi+21]

- Requires only target transcriptions,
- No need for parallel data
  - → Could train with real mixtures (even for explicit systems)

Copyright 2021 NTT CORPORATION, BUT

Hybrid or E2E

Hybrid systems

[Delcroix+18 , Zmolikova+18, Kanda+19]

- Require mixtures and HMM-forced alignment of target speech
- Effectively require parallel mixture/clean speech
  - → Limit for simulated data

E2E systems

[Delcroix+19 , Denisov+19, Shi+21]

- Requires only target transcriptions,
- No need for parallel data
  - → Could train with real mixtures (even for explicit systems)

Copyright 2021 NTT CORPORATION, BUT

Using other modalities for TS-ASR

- Audio-visual TS-ASR [Chao+16, Pan+21]
  - Use visual clues instead of enrollment
- Audio-visual-Locational TS-ASR [Yu+20]
  - Combine Locational+Visual-based extraction network with ASR module
  - ASR module also use lip movement information as auxiliary information

Copyright 2021 NTT CORPORATION, BUT

References (1/4)

[Bahdanau+16] Bahdanau et al. “End-to-end attention-based large vocabulary speech recognition,” ICASSP, 2016.

[Boeddecker+18] Boeddecker et al. “Front-end processing for the CHiME-5 dinner party scenario,” CHiME, 2018.

[Chan+15] Chan et al. “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” ICASSP, 2015.

[Chang+18] Chang et al. “Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks,” Interspeech, 2018.

[Chang+19] Chang et al., “End-to-end monaural multi-speaker asr system without pretraining,” ICASSP, 2019.

[Chao+16] Chao et al., “Speaker-Targeted Audio-Visual Models for Speech Recognition in Cocktail-Party Environments,” Interspeech, 2016

[Delcroix+18] Delcroix et al. “Single channel target speaker extraction and recognition with speaker beam,” ICASSP, 2018.

[Delcroix+19] Delcroix et al. “End-to-end SpeakerBeam for single channel target speech recognition,” Interspeech, 2019.

[Delcroix+21] Delcroix et al., “Speaker Activity Driven Neural Speech Extraction,” ICASSP, 2021.

[Denisov+19] Denisov et al., “End-to-End Multi-Speaker Speech Recognition using Speaker Embeddings and Transfer Learning,” Interspeech, 2019.

[Graves+14] Graves et al. “Towards end-to-end speech recognition with recurrent neural networks,” ICML, 2014.

[Hinton+12] Hinton et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, 2012.

Copyright 2021 NTT CORPORATION, BUT

References (2/4)

[Hori+12] Hori et al. "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-Directional Camera," in IEEE Trans. ASLP, 2012

[Kanda+19] Kanda et al., “Auxiliary Interference Speaker Loss for Target-Speaker Speech Recognition,” Interspeech, 2019.

[Kanda+19] Kanda et al. “Simultaneous Speech Recognition and Speaker Diarization for Monaural Dialogue Recordings with Target-Speaker Acoustic Models,” ASRU, 2019.

[Kanda+20] Kanda et al., “Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers,” Interspeech, 2020.

[Kanda+21] Kanda et al., “Investigation of end-to-end speaker attributed ASR for continuous multi-talker recordings,” SLT, 2021.

[Kanda+21] Kanda et al., “A comparative study of modular and joint approaches for speaker-attributed ASR on monaural long-form audio,” Submitted to ASRU, 2021.

[Kida+18] Kida et al. "Speaker Selective Beamformer with Keyword Mask Estimation," SLT, 2018.

[King+17] King et al. " Robust speech recognition via anchor word representations,” Interspeech, 2017.

[Medennikov+20] Medennikov et al. “Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario,” Interspeech, 2020.

[Pan+21] Pan et al. "Selective Hearing through Lip-reading," arXiv, 2021.

[Rao+19] Rao et al. “Target Speaker Extraction for Multi-Talker Speaker Verification,” Interspeech, 2019.0

[Renals+07] Renals et al. "Recognition and understanding of meetings the AMI and AMIDA projects", ASRU, 2007.

Copyright 2021 NTT CORPORATION, BUT

References (3/4)

[Rikhye+21] Rikhye et al., “Personalized Keyphrase Detection using Speaker and Environment Information,” arXiv, 2021.

[Saon+13] Saon et al. “Speaker adaptation of neural network acoustic models using i-vectors,” ASRU, 2013.

[Sato+21] Sato et al. “Should We Always Separate?: Switching Between Enhanced and Observed Signals for Overlapping Speech Recognition,” Interspeech, 2021.

[Settle+18] Settle et al. “End-to-end multi-speaker speech recognition,” ICASSP, 2018.

[Shi+21] Shi et al. “Improving RNN Transducer With Target Speaker Extraction and Neural Uncertainty Estimation,” ICASSP, 2021.

[Subramanian+20] Subramanian et al, “Far-Field Location Guided Target Speech Extraction Using End-to-End Speech Recognition Objectives,” ICASSP, 2020.

[Waibel+98] Waibel et al. "Meeting browser: Tracking and summarizing meetings", BNTU Workshop, 1998.

[Wang+19] Wang et al. “End-to-end Anchored Speech Recognition,” ICASSP, 2019.

[Wang+20] Wang et al., “Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition,” Interspeech, 2020.

Copyright 2021 NTT CORPORATION, BUT

References (4/4)

[Yu+15] Yu et al. “Automatic Speech Recognition – A Deep Learning Approach,” Springer, 2015.

[Yu+21] Yu et al., "Audio-Visual Multi-Channel Integration and Recognition of Overlapped Speech," IEEE/ACM Trans. ASLP, 2021.

[Zmolikova+18] Zmolikova et al. “Optimization of speaker-aware multichannel speaker extraction with ASR criterion,” ICASSP, 2018.

[Zmolikova+21] Zmolikova et al. “Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics,” Interspeech, 2021.

Copyright 2021 NTT CORPORATION, BUT

6. Conclusion & Future directions

Slides: https://butspeechfit.github.io/tse_tutorial

Copyright 2021 NTT CORPORATION, BUT

Summary

- Target-speech processing
  - Several ways to realize
    - Clues: Audio, video, Locational
    - Configurations
  - Can be used for various tasks (Enhancement, VAD, ASR, etc)

- Related to speech separation but it can mitigate
  - Dependence on number of sources
  - Permutation ambiguity

  → Practical alternative for many scenarios

  - Can exploit novel approaches developed for separation

  → Lots of opportunities for research

Copyright 2021 NTT CORPORATION, BUT

Receiving increasing attention

Many related papers at this Interspeech:

- 19:00 **Tue-E-V-2-10** 2253 *Should We Always Separate?: Switching Between Enhanced and Observed Signals for Overlapping Speech Recognition*, Sato et al.
- 11:00 **Wed-M-O-3-1** 986 *Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics*, Zmolikova et al.
- 11:00 **Thu-M-V-3-3** 338 *Continuous Speech Separation Using Speaker Inventory for Long Recording*, Han et al.
- 16:00 **Thu-A-V-3-5** 1369 *Few-Shot Learning of New Sound Classes for Target Sound Extraction*, Delcroix et al.
- 16:00 **Thu-A-V-3-7** 1378 *AvaTr: One-Shot Speaker Extraction with Transformers*, Hu et al.
- 16:00 **Thu-A-V-3-11** 2250 *Robust Speaker Extraction Network Based on Iterative Refined Adaptation*, Deng et al.
- 16:00 **Thu-A-V-3-12** 2260 *Neural Speaker Extraction with Speaker-Speech Cross-Attention Network*, Wang et al.
- 13:30 **Tue-A-S&T-1-7** ST07 *Advanced Semi-Blind Speaker Extraction and Tracking Implemented in Experimental Device with Revolving Dense Microphone Array*, Čmejla et al.
- 11:20 **Wed-M-O-3-2** 1939 *Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers*, Morsdorf et al.
- 12:30 **Wed-M-SS-1-4** 298 *Improving Channel Decorrelation for Multi-Channel Target Speech Extraction*, Han et al.

Copyright 2021 NTT CORPORATION, BUT

Extensions

- Integration of separation and TSE
- Target sound extraction
- EEG-based approaches

Copyright 2021 NTT CORPORATION, BUT

Integration of separation and TSE

- Speech separation can benefit from speaker embeddings
  - Improves separation performance
  - Enables tracking sources for long recordings
- Speaker representation learned from
  - Enrollment of all speakers [Wang+19]
  - Mixture [Kinoshita+20, Cong+21, Zeghidour+21]

Wavesplit [Zeghidour+21]:

From [Zeghidour+21]

Copyright 2021 NTT CORPORATION, BUT

Target sound extraction

- Extension to non-speech signals
  - Extract target sound from a mixture of sounds

Target sounds

Dog	×
Woman	×
Man	×
Siren	○
Car	×

16:00 **Thu-A-V-3-5** 1369 *Few-Shot Learning of New Sound Classes for Target Sound Extraction*, Marc Delcroix, Jorge Bennasar Vázquez, Tsubasa Ochiai, Keisuke Kinoshita and Shoko Araki

Copyright 2021 NTT CORPORATION, BUT

EEG-based approaches

- Extract speech given brain signal (EEG etc.) [O'Sullivan+14, Aroudi+20, Gfeller+21]
  - Hearing aid users could attend to desired speaker
  - Challenging problem but attracting much interest

Ceolini et al. "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," 2020

Copyright 2021 NTT CORPORATION, BUT

Future directions

Copyright 2021 NTT CORPORATION, BUT

Future research directions

- Investigate more diverse acoustic conditions
  - Dynamic conditions (Moving speakers, ⋯)
  - Various recordings (rooms, noise types, SNRs, mics, ⋯)
  - More realistic (Spontaneous speech, Unsegmented data, ⋯)
- This may require
  - Adaptation to deployment conditions
  - Exploiting non-parallel data

11:00 **Wed-M-O-3-1** 986 *Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics*, Zmolikova et al.

Copyright 2021 NTT CORPORATION, BUT

Future research directions

- Deeper investigation of target speaker representation

Better Discrimination

- Improved performance for same-gender mixtures
- False positive not sufficiently investigated

More Robust

- Handling different recording conditions (noise, room, speaker condition/health)

Images from:
https://commons.wikimedia.org/wiki/File:Balance,_by_David.svg
https://freesvg.org/sneezing-into-handkerchief

Copyright 2021 NTT CORPORATION, BUT

Future research directions

- Online processing/lightweight processing
  - Towards processing on hearing aids/hearables

- Other clues?

Copyright 2021 NTT CORPORATION, BUT

References (1/2)

[Aroudi+20] Aroudi et al. "Cognitive-driven convolutional beamforming using EEG-based auditory attention decoding." MLSP, 2020.

[Borsdorf+21] Borsdorf et al. “Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers,” Interspeech, 2021.

[Ceolini +20] Ceolini et al. “Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception,” NeuroImage, 2020.

[Čmejla +21] Čmejla et al. “Advanced Semi-Blind Speaker Extraction and Tracking Implemented in Experimental Device with Revolving Dense Microphone Array,” Interspeech, 2021.

[Cong+21] Cong et al. “Continuous Speech Separation Using Speaker Inventory for Long Recording,” Interspeech, 2021

[Delcroix +21] Delcroix et al. “Few-Shot Learning of New Sound Classes for Target Sound Extraction,” Interspeech, 2021.

[Deng+21] Deng et al. “Robust Speaker Extraction Network Based on Iterative Refined Adaptation,” Interspeech, 2021.

[Gfeller+21] Gfeller et al. “One-shot conditional audio filtering of arbitrary sounds,” ICASSP, 2021.

[Han+21] Han et al. “Continuous Speech Separation Using Speaker Inventory for Long Recording,” Interspeech, 2021.

[Han+21] Han et al. “Improving Channel Decorrelation for Multi-Channel Target Speech Extraction,” Interspeech, 2021.

[Hu+21] Hu et al. “AvaTr: One-Shot Speaker Extraction with Transformers,” Interspeech, 2021.

Copyright 2021 NTT CORPORATION, BUT

References (2/2)

[Kinoshita+20] Kinoshita et al., ``Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system'', 2020

[Kong+20] Kong et al. “Source separation with weakly labelled data: An approach to computational auditory scene analysis,” ICASSP, 2020.

[Lee+19] Lee et al. “Audio query-based music source separation,” ISMIR, 2019.

[O’Sullivan+14] O’Sullivan et al. “Attentional selection in a cocktail party environment can be decoded from single-trial EEG,” Cerebral Cortex, 2014.

[Ochiai+20] Ochiai et al. “Listen to What You Want: Neural Network-Based Universal Sound Selector” Interspeech, 2020.

[Sato+21] Sato et al. “Should We Always Separate?: Switching Between Enhanced and Observed Signals for Overlapping Speech Recognition,” Interspeech, 2021.

[Wang+19] Wang et al. “Speech separation using speaker inventory,” ASRU, 2019.

[Wang+21] Wang et al. “Neural Speaker Extraction with Speaker-Speech Cross-Attention Network,” Interspeech, 2021.

[Zeghidour+21] Zeghidour et al. "Wavesplit: End-to-end speech separation by speaker clustering." IEEE/ACM Trans. ASLP, 2021.

[Zmolikova +21] Zmolikova et al. “Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics,” Interspeech, 2021.

Copyright 2021 NTT CORPORATION, BUT

Thank you!
Questions!?

Slides: https://butspeechfit.github.io/tse_tutorial
Code: https://github.com/BUTSpeechFIT/speakerbeam

Copyright 2021 NTT CORPORATION, BUT