

End-to-end deep neural network based speaker and language recognition

Ming Li¹

Data Science Research Center, Duke Kunshan University
Department of Electrical and Computer Engineering, Duke University

Sep 17th 2019

¹Thanks Weicheng Cai, Danwei Cai, Qingjian Lin and Haiwei Wu for their contributions

Table of Contents

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

Table of Contents

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

Problem Formulation

- Speech signal not only contains lexicon information, but also deliver various kinds of **paralinguistic speech attribute information**, such as **speaker**, **language**, gender, age, emotion, channel, voicing, psychological states, etc.

Problem Formulation

- Speech signal not only contains lexicon information, but also deliver various kinds of **paralinguistic speech attribute information**, such as **speaker**, **language**, gender, age, emotion, channel, voicing, psychological states, etc.
- The core technique question behind it is utterance level supervised learning based on text independent or text dependent speech signal with flexible duration



Problem Formulation

- Speech signal not only contains lexicon information, but also deliver various kinds of **paralinguistic speech attribute information**, such as **speaker**, **language**, gender, age, emotion, channel, voicing, psychological states, etc.
- The core technique question behind it is utterance level supervised learning based on text independent or text dependent speech signal with flexible duration
- The traditional framework

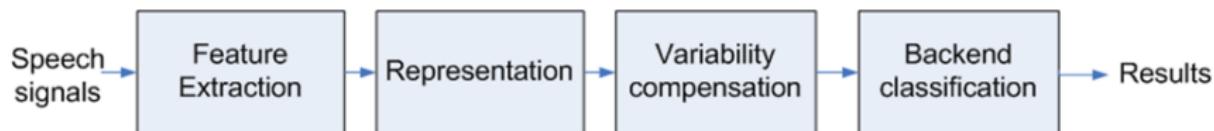


Figure: General framework

Table of Contents

1 Problem Formulation

2 Traditional Framework

- Feature Extraction
- Representation
- Variability Compensation
- Backend Classification

3 End-to-End Deep Neural Network based Framework

- System Pipeline
- Data Preparation
- Network Structure
- Encoding Mechanism
- Loss Function
- Data Augmentation
- Domain Adaptation

4 Robust Modeling of End-to-End methods

- Speech under Far Field and Complex Environment Settings
- Previous Methods on Robust Modeling
- Robust Modeling of End-to-End Methods

5 Other Applications of End-to-End Methods

- Speaker Diarization
- Paralinguistic Speech Attribute Recognition
- Anti-spoofing Countermeasures

Feature Extraction

- MFCC, PLP, SDC [1]², PNCC[2]³, GFCC[3]⁴ , CQCC [4]⁵,etc.

²P. Torres-Carrasquillo et al. "Approaches to language identification using gaussian mixture models and shifted delta cepstral features". In: *Proc. of ICSLP*. 2002, pp. 89–92.

³C. Kim and R. M. Stern. "Power-Normalized Cepstral Coefficients PNCC for Robust Speech Recognition". In: *IEEE Transactions on Audio Speech and Language Processing* 24.7 (2016), pp. 1315–1329.

⁴Shao Yang and De Liang Wang. "Robust speaker identification using auditory features and computational auditory scene analysis". In: *Proc. of ICASSP*. 2008.

⁵Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification". In: *Computer Speech and Language* 45 (2017).



Feature Extraction

- MFCC, PLP, SDC [1], PNCC[2], GFCC[3] , CQCC [4],etc.
- Bottleneck [5]²[6]³, Phoneme Posterior Probability [7]⁴[8]⁵, etc.

²Pavel Matejka et al. "Neural Network Bottleneck Features for Language Identification." In: *Proc. of Odyssey*. 2014.

³Achintya K Sarkar et al. "Combination of cepstral and phonetically discriminative features for speaker verification". In: *IEEE Signal Processing Letters* 21.9 (2014), pp. 1040–1044.

⁴Ming Li and Wenbo Liu. "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features". In: *Proc. of Interspeech*. 2014.

⁵F. Richardson, D. Reynolds, and N. Dehak. "Deep Neural Network Approaches to Speaker and Language Recognition". In: *IEEE Signal Processing Letters* 22.10 (2015), pp. 1671–1675.



Feature Extraction

- MFCC, PLP, SDC [1], PNCC[2], GFCC[3] , CQCC [4],etc.
- Bottleneck [5][6], Phoneme Posterior Probability [7][8], etc.
- LLD/OpenSmile [9]², Speech attributes [10]³, Acoustic-to-articulatory inversion [11]⁴, subglottal[12]⁵, etc.

²Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proc. of ACM Multimedia*. 2010, pp. 1459–1462.

³Hamid Behravan et al. "Introducing attribute features to foreign accent recognition". In: *Proc. of ICASSP*. IEEE. 2014, pp. 5332–5336.

⁴Ming Li et al. "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals". In: *Computer speech & language* 36 (2016), pp. 196–211.

⁵Jinxi Guo et al. "Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features." In: *Proc. of INTERSPEECH*. 2016, pp. 2219–2222.

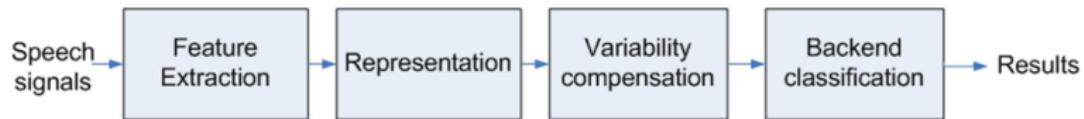
Feature Extraction

- MFCC, PLP, SDC [1], PNCC[2], GFCC[3] , CQCC [4],etc.
- Bottleneck [5][6], Phoneme Posterior Probability [7][8], etc.
- LLD/OpenSmile [9], Speech attributes [10], Acoustic-to-articulatory inversion [11], subglottal[12], etc.
- IMFCC[13]², Modified Group Delay[14]³ , etc.

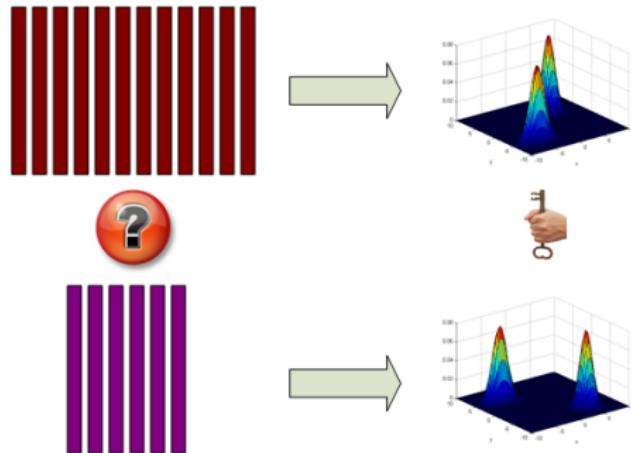
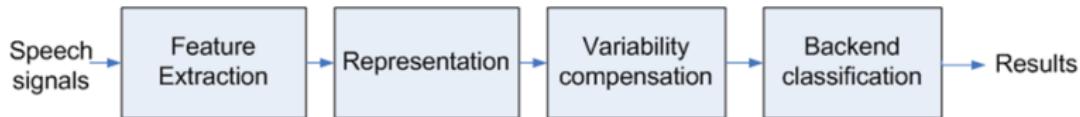
²Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. "A comparison of features for synthetic speech detection". In: (2015).

³Zhiheng Wu, Eng Siong Chng, and Haizhou Li. "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition". In: *Proc. of Interspeech*. 2012.

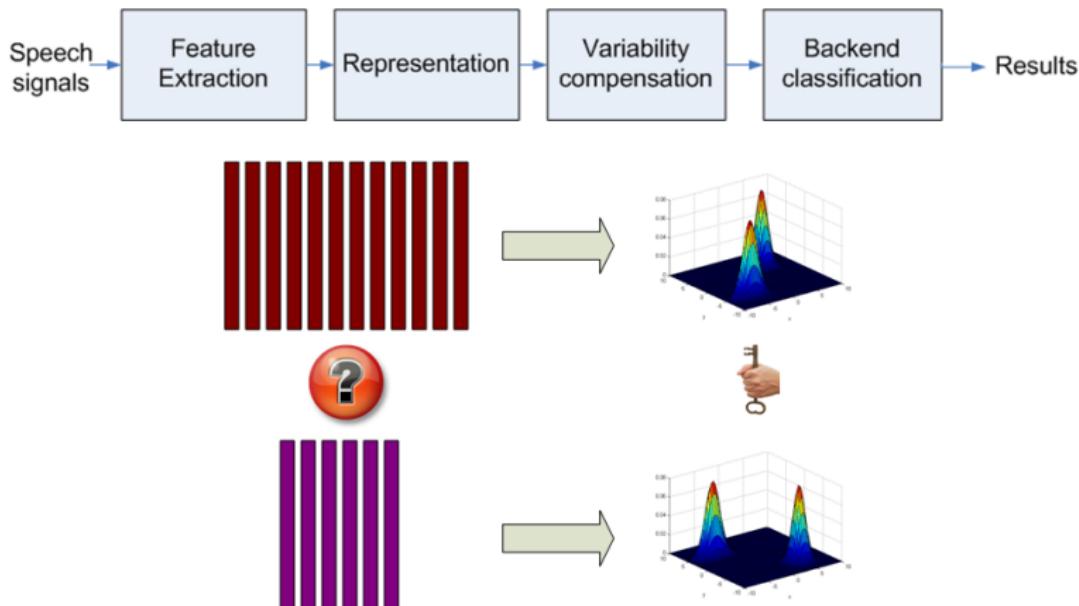
Representation



Representation



Representation



- time **varying** property \Rightarrow short time **frame level** features
- **generative model** for data description \Rightarrow features (**supervectors**) in model parameters' space for classification

Generative model, adaptation, supervectors

- Gaussian Mixture Model (GMM) [15]⁴ serves as the generative model

⁴D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models" In: *Digital Signal Processing*. 2000, 1941.



Generative model, adaptation, supervectors

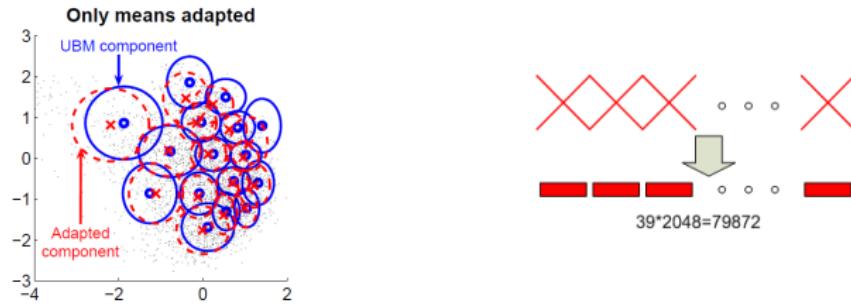
- Gaussian Mixture Model (GMM) [15]⁴ serves as the generative model
- **model adaptation** from universal background model (UBM)

⁴D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models" In: *Digital Signal Processing*. 2000, 1941.



Generative model, adaptation, supervectors

- Gaussian Mixture Model (GMM) [15] serves as the generative model
- **model adaptation** from universal background model (UBM)
 - **MAP adaptation**, concatenating mean vector from all GMM components to get a large dimensional **GMM mean supervector** [16]⁴



⁴W.M Campbell et al. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: Proc. of ICASSP. Vol. 1. 2006, pp. 97–100.

Generative model, adaptation, supervectors

- Gaussian Mixture Model (GMM) [15] serves as the generative model
- **model adaptation** from universal background model (UBM)
 - **MAP adaptation**, concatenating mean vector from all GMM components to get a large dimensional **GMM mean supervector** [16]⁴
 - **Maximum Likelihood Linear Regression (MLLR)** adaptation
the linear regression matrix becomes GMM MLLR supervector [17]⁵

⁴W.M Campbell et al. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: *Proc. of ICASSP*. Vol. 1. 2006, pp. 97–100.

⁵Andreas Stolcke et al. "MLLR transforms as features in speaker recognition". In: *Ninth European Conference on Speech Communication and Technology*. 2005.



Generative model, adaptation, supervectors

- The **statistics vector** for a set of features on UBM



Generative model, adaptation, supervectors

- The **statistics vector** for a set of features on UBM
 - 0^{th} order statistics vector N , centered normalized 1^{st} order statistics vector F

$$N_c = \sum_{t=1}^L P(c|\mathbf{y}_t, \lambda) \quad (1)$$

Cumulated by L frames

$$\tilde{\mathbf{F}}_c = \frac{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)(\mathbf{y}_t - \mu_c)}{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)}. \quad (2)$$



Generative model, adaptation, supervectors

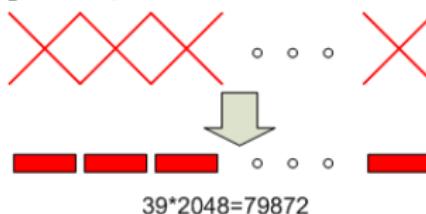
- The **statistics vector** for a set of features on UBM

- 0th order statistics vector N , centered normalized 1st order statistics vector F

$$N_c = \sum_{t=1}^L P(c|\mathbf{y}_t, \lambda) \quad (1)$$

Cumulated by L frames

$$\tilde{\mathbf{F}}_c = \frac{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)(\mathbf{y}_t - \mu_c)}{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)}. \quad (2)$$



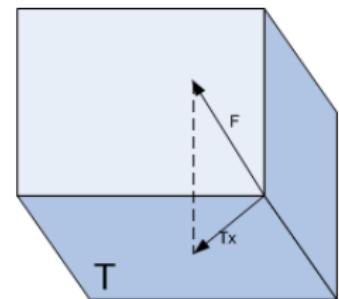
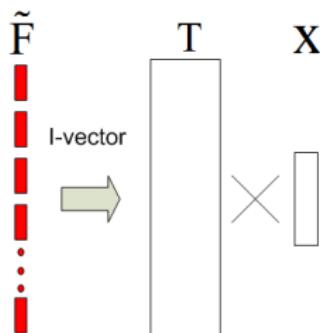
Generative model, adaptation, supervectors

- The **statistics vector** for a set of features on UBM
 - 0^{th} order statistics vector N , centered normalized 1^{st} order statistics vector F
- $$N_c = \sum_{t=1}^L P(c|\mathbf{y}_t, \lambda) \quad (1)$$
Cumulated by L frames
- $$\tilde{\mathbf{F}}_c = \frac{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)(\mathbf{y}_t - \mu_c)}{\sum_{t=1}^L P(c|\mathbf{y}_t, \lambda)}. \quad (2)$$
- **Mapping** from a set of feature vectors to a fixed dimensional supervector



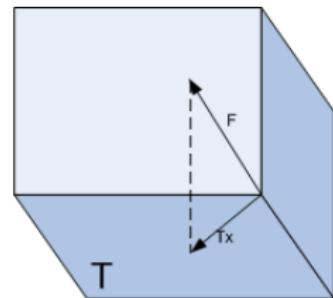
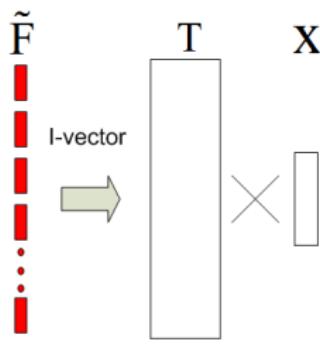
Factor analysis based dimension reduction

- Factor analysis on the concatenated centered normalized 1st order statistics vector or GMM mean supervector



Factor analysis based dimension reduction

- Factor analysis on the concatenated centered normalized 1st order statistics vector or GMM mean supervector
 - total variability i-vector [18]⁶
$$\tilde{F} = Tx \quad (3)$$
T: factor loading matrix, x: i-vector



⁶N. Dehak et al. "Front-end factor analysis for speaker verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

⁷Patrick Kenny et al. "Joint factor analysis versus eigenchannels in speaker recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1435–1447.

Factor analysis based dimension reduction

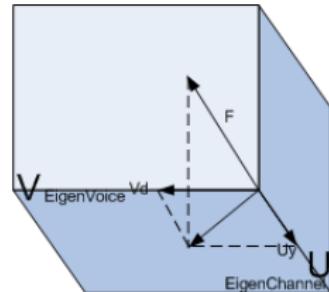
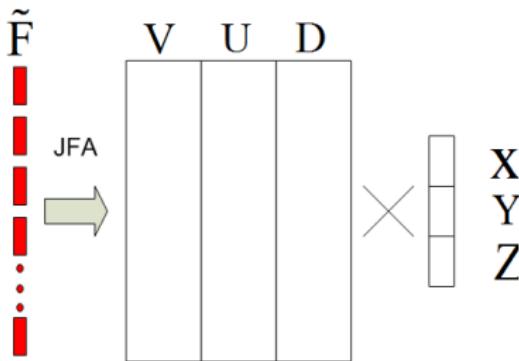
- Factor analysis on the concatenated centered normalized 1st order statistics vector or GMM mean supervector

- total variability i-vector [18]⁶

$$\tilde{\mathbf{F}} = \mathbf{T}\mathbf{x} \quad (3) \quad \mathbf{T}: \text{factor loading matrix}, \mathbf{x}: \text{i-vector}$$

- joint factor analysis (JFA) [19]⁷

$$\tilde{\mathbf{F}} = \mathbf{V}\mathbf{x} + \mathbf{U}\mathbf{y} + \mathbf{D}\mathbf{z} \quad (4) \quad \begin{aligned} \mathbf{V}: & \text{Eigenvoices, } \mathbf{U}: \text{Eigenchannels,} \\ \mathbf{x}: & \text{speaker factor, } \mathbf{y}: \text{channel factor,} \\ \mathbf{D}: & \text{diagonal covariance matrix} \end{aligned}$$



⁶N. Dehak et al. "Front-end factor analysis for speaker verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

⁷Patrick Kenny et al. "Joint factor analysis versus eigenchannels in speaker recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1435–1447.

Variability Compensation

LDA, WCCN [20]⁸, NAP[16]⁹, NDA [21]¹⁰, LSDA [22]¹¹, LFDA [23]¹², etc.

⁸ A.O. Hatch, S. Kajarekar, and A. Stolcke. "Within-class covariance normalization for SVM-based speaker recognition". In: *Proc. of INTERSPEECH*. Vol. 4. 2006, pp. 1471–1474.

⁹ W.M Campbell et al. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: *Proc. of ICASSP*. Vol. 1. 2006, pp. 97–100.

¹⁰ Seyed Omid Sadjadi, Jason Pelecanos, and Weizhong Zhu. "Nearest neighbor discriminant analysis for robust speaker recognition". In: *Proc. of Interspeech*. 2014.

¹¹ Danwei Cai et al. "Locality sensitive discriminant analysis for speaker verification". In: *Proc. of APSIPA ASC*. 2016, pp. 1–5.

¹² Peng Shen et al. "Local fisher discriminant analysis for spoken language identification". In: *Proc. of ICASSP*. 2016, pp. 5825–5829.

Backend Classification

SVM [16]¹³, PLDA [24]¹⁴[25]¹⁵, NN [26]¹⁶[27]¹⁷, Joint Bayesian [28]¹⁸ , Cosine Similarity, etc.

¹³W.M Campbell et al. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: *Proc. of ICASSP*. Vol. 1. 2006, pp. 97–100.

¹⁴S.J.D. Prince and J.H. Elder. "Probabilistic linear discriminant analysis for inferences about identity". In: *Proc. ICCV*. 2017.

¹⁵D. Garcia-Romero and C. Y Espy-Wilson. "Analysis of i-vector Length Normalization in Speaker Recognition Systems." In: *Proc. INTERSPEECH*. 2011, pp. 249–252.

¹⁶Kyu Jeong Han et al. "TRAP language identification system for RATS phase II evaluation". In: *Proc. of Interspeech*. 2013, pp. 1502–1506.

¹⁷Omid Ghahabi et al. "Deep Neural Networks for iVector Language Identification of Short Utterances in Cars". In: *Proc. of Interspeech*. 2016, pp. 367–371.

¹⁸Yiyan Wang, Haotian Xu, and Zhijian Ou. "Joint bayesian gaussian discriminant analysis for speaker verification". In: *Proc. of ICASSP*. IEEE. 2017, pp. 5390–5394.

Table of Contents

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

System Pipeline



- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29]¹⁹, Mel-filterbank energies [30]²⁰, MFCC [31]²¹), or even the raw waveform [32]²².

¹⁹ Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>.

²⁰ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, *arXiv:1705.02304* (2017), *arXiv:1705.02304*. *arXiv:1705.02304 [cs.CL]*.

²¹ D. Snyder et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *Proc. IEEE SLT*. 2017.

²² Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *Proc. of SLT*. IEEE. 2018, pp. 1021–1028.

System Pipeline



- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29]¹⁹, Mel-filterbank energies [30]²⁰, MFCC [31]²¹), or even the raw waveform [32]²².
- The local pattern extractor plays a role as an automatic representation learning module. (TDNN/CNN/LSTM/CNN-LSTM/CNN-BLSTM).

¹⁹ Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>.

²⁰ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

²¹ D. Snyder et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *Proc. IEEE SLT*. 2017.

²² Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *Proc. of SLT*. IEEE. 2018, pp. 1021–1028.

System Pipeline



- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29]¹⁹, Mel-filterbank energies [30]²⁰, MFCC [31]²¹), or even the raw waveform [32]²².
- The local pattern extractor plays a role as an automatic representation learning module. (TDNN/CNN/LSTM/CNN-LSTM/CNN-BLSTM).
- The encoding layer encodes the variable-length sequence into a fixed-dimensional utterance-level representation. (Recurrent encoding / Pooling)

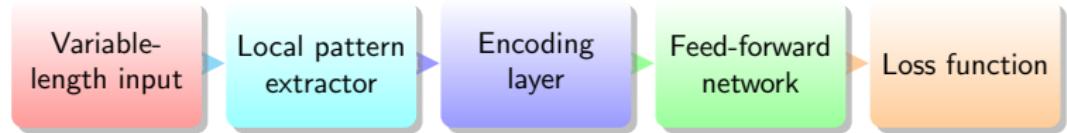
¹⁹ Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>.

²⁰ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, *arXiv:1705.02304* (2017), *arXiv:1705.02304*. *arXiv:1705.02304 [cs.CL]*.

²¹ D. Snyder et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *Proc. IEEE SLT*. 2017.

²² Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *Proc. of SLT*. IEEE. 2018, pp. 1021–1028.

System Pipeline



- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29]¹⁹, Mel-filterbank energies [30]²⁰, MFCC [31]²¹), or even the raw waveform [32]²².
- The local pattern extractor plays a role as an automatic representation learning module. (TDNN/CNN/LSTM/CNN-LSTM/CNN-BLSTM).
- The encoding layer encodes the variable-length sequence into a fixed-dimensional utterance-level representation. (Recurrent encoding / Pooling)
- All the network components are jointly optimized with a global loss function.
(Forward + Backward + Stochastic gradient descent)

¹⁹ Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>.

²⁰ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

²¹ D. Snyder et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *Proc. IEEE SLT*. 2017.

²² Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *Proc. of SLT*. IEEE. 2018, pp. 1021–1028.

System Pipeline



Task

- Language identification or paralinguistic speech attributes detection(Closed-set)
Network output → Utterance-level posteriors

System Pipeline



Task

- Language identification or paralinguistic speech attributes detection(Closed-set)
Network output → Utterance-level posteriors
- Speaker Verification (Open-set)
Utterance-level speaker embedding + Cosine / PLDA → Pairwise scores

Data preparation

Traditional workflow

- Off-the-shelf full-length utterance

Network workflow



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently

Network workflow



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .

Network workflow



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .
- Arbitrary duration audio waveform → variable-length feature sequence → utterane-level fixed-dimensional embedding (e.g. i-vector).

Network workflow



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .
- Arbitrary duration audio waveform → variable-length feature sequence → utterane-level fixed-dimensional embedding (e.g. i-vector).

Network workflow

- Well-prepared mini-batch tensor block in the training stage.



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .
- Arbitrary duration audio waveform → variable-length feature sequence → utterane-level fixed-dimensional embedding (e.g. i-vector).

Network workflow

- Well-prepared mini-batch tensor block in the training stage.
- Several utterances are grouped together → Multi-dimensinal array



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .
- Arbitrary duration audio waveform → variable-length feature sequence → utterane-level fixed-dimensional embedding (e.g. i-vector).

Network workflow

- Well-prepared mini-batch tensor block in the training stage.
- Several utterances are grouped together → Multi-dimensinal array
- The parameters are updated for each batch of data



Data preparation

Traditional workflow

- Off-the-shelf full-length utterance
- Each utterance is performed independently
- The parameters are updated after seeing all the (or sampled) utterances .
- Arbitrary duration audio waveform → variable-length feature sequence → utterance-level fixed-dimensional embedding (e.g. i-vector).

Network workflow

- Well-prepared mini-batch tensor block in the training stage.
- Several utterances are grouped together → Multi-dimensinal array
- The parameters are updated for each batch of data
- In the testing stage, arbitrary duration audio waveform → variable-length feature sequence → utterance-level fixed-dimensional embedding (e.g. x-vector).



DNN data preparation

D-vector [33]²³[34]²⁴[35]²⁵

- Raw feature sequences are broken into multiple small fixed-length data chunks at the frame level.

²³Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.

²⁴Yuan Liu et al. "Deep feature for text-dependent speaker verification". In: *Speech Communication* 73 (2015), pp. 1–13.

²⁵Lantian Li et al. "Deep speaker vectors for semi text-independent speaker verification". In: *arXiv preprint arXiv:1505.06427* (2015).

DNN data preparation

D-vector [33]²³[34]²⁴[35]²⁵

- Raw feature sequences are broken into multiple small fixed-length data chunks at the frame level.
- The input layer is fed with dozens of frames formed by stacking the currently processed frame and its several left-right context frames.

²³Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.

²⁴Yuan Liu et al. "Deep feature for text-dependent speaker verification". In: *Speech Communication* 73 (2015), pp. 1–13.

²⁵Lantian Li et al. "Deep speaker vectors for semi text-independent speaker verification". In: *arXiv preprint arXiv:1505.06427* (2015).

DNN data preparation

D-vector [33]²³[34]²⁴[35]²⁵

- Raw feature sequences are broken into multiple small fixed-length data chunks at the frame level.
- The input layer is fed with dozens of frames formed by stacking the currently processed frame and its several left-right context frames.
- This data preparation procedure generates a large amount of temporary data chunks.

²³Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.

²⁴Yuan Liu et al. "Deep feature for text-dependent speaker verification". In: *Speech Communication* 73 (2015), pp. 1–13.

²⁵Lantian Li et al. "Deep speaker vectors for semi text-independent speaker verification". In: *arXiv preprint arXiv:1505.06427* (2015).

DNN data preparation

D-vector [33]²³[34]²⁴[35]²⁵

- Raw feature sequences are broken into multiple small fixed-length data chunks at the frame level.
- The input layer is fed with dozens of frames formed by stacking the currently processed frame and its several left-right context frames.
- This data preparation procedure generates a large amount of temporary data chunks.
- In the testing stage, it is also necessary to break the testing segments into a bunch of fixed-length frames.

²³Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.

²⁴Yuan Liu et al. "Deep feature for text-dependent speaker verification". In: *Speech Communication* 73 (2015), pp. 1–13.

²⁵Lantian Li et al. "Deep speaker vectors for semi text-independent speaker verification". In: *arXiv preprint arXiv:1505.06427* (2015).

Data preparation

X-vector [36]²⁶

- Several archive files containing data chunks with different segment lengths and augmentation types are prepared carefully beforehand

²⁶David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

Data preparation

X-vector [36]²⁶

- Several archive files containing data chunks with different segment lengths and augmentation types are prepared carefully beforehand
- The input layer is fed with variable-length segments.

²⁶David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

Data preparation

X-vector [36]²⁶

- Several archive files containing data chunks with different segment lengths and augmentation types are prepared carefully beforehand
- The input layer is fed with variable-length segments.
- This data preparation procedure also generates a large amount of temporary data chunks when data augmentation is performed.

²⁶David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

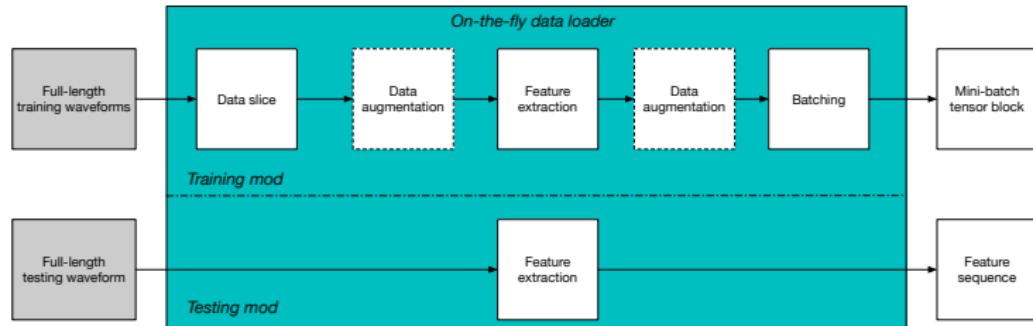
Data preparation

X-vector [36]²⁶

- Several archive files containing data chunks with different segment lengths and augmentation types are prepared carefully beforehand
- The input layer is fed with variable-length segments.
- This data preparation procedure also generates a large amount of temporary data chunks when data augmentation is performed.
- In the testing stage, the full-length utterance-level feature sequence can be directly fed into the network.

²⁶David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

Data preparation

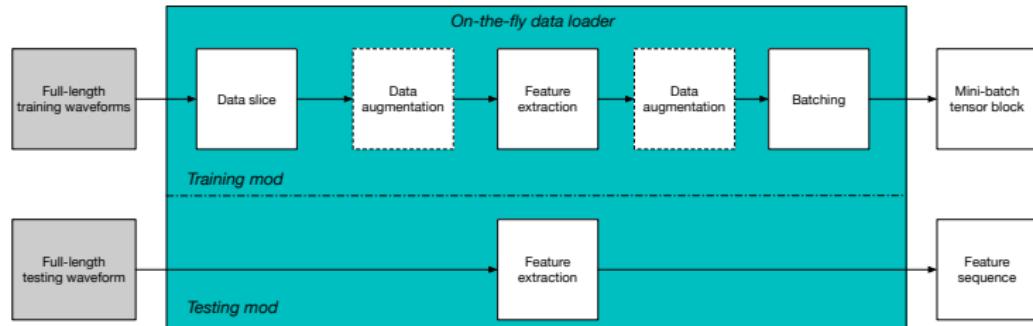


On-the-fly data loader [37]²⁷

- Offline augmentation requires us to generate all the necessary training samples into disk beforehand. On the contrary, a data loader here maintains an online processing work flow to generate training sample on the fly.

²⁷Weicheng Cai et al. "On-the-Fly Data Loader and Utterance-level Aggregation for Speaker and Language Recognition". In: submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing (2019).

Data preparation

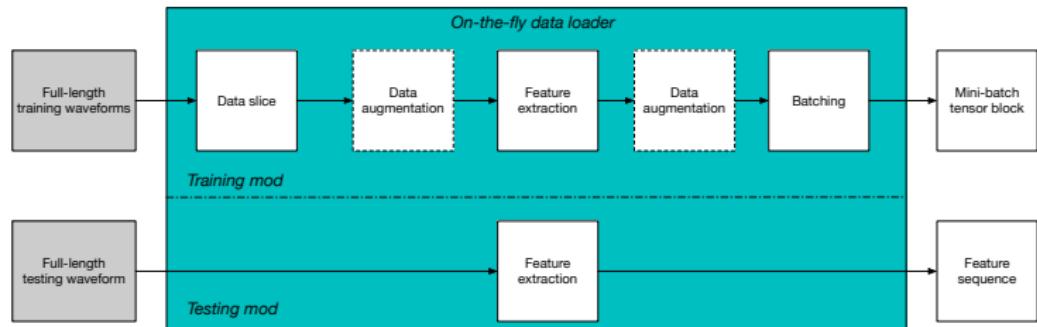


On-the-fly data loader [37]²⁷

- Offline augmentation requires us to generate all the necessary training samples into disk beforehand. On the contrary, a data loader here maintains an online processing work flow to generate training sample on the fly.
- Multiple real-time operations within the data loader: the data slice, the data transformation (including feature extraction and data augmentation), and the data batching operation.

²⁷Weicheng Cai et al. "On-the-Fly Data Loader and Utterance-level Aggregation for Speaker and Language Recognition". In: submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing (2019).

Data preparation

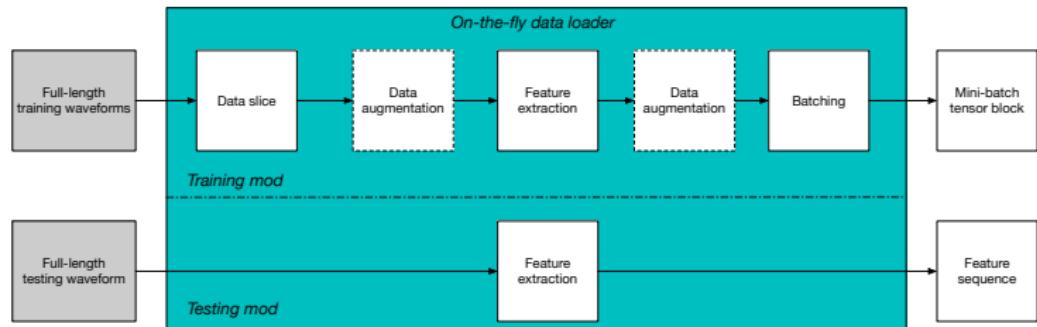


On-the-fly data loader [37]²⁷

- This design principle allows us to perform the batch-wise random perturbation, such as variable-length data slice and online data augmentation efficiently. All the operations are eagerly executed on the fly, and the training samples are generated in the memory just before feeding it into the DNNs.

²⁷Weicheng Cai et al. "On-the-Fly Data Loader and Utterance-level Aggregation for Speaker and Language Recognition". In: submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing (2019).

Data preparation



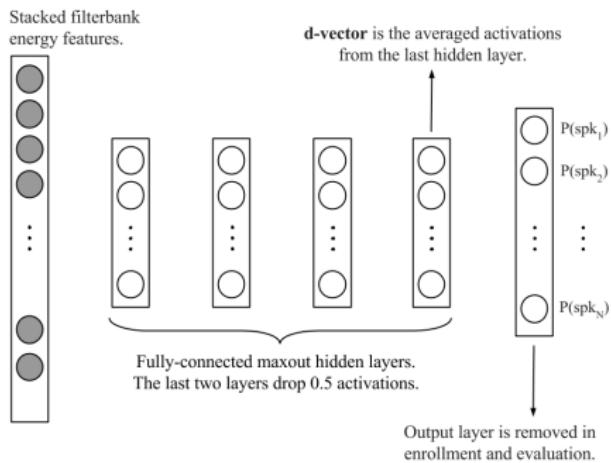
On-the-fly data loader [37]²⁷

- This design principle allows us to perform the batch-wise random perturbation, such as variable-length data slice and online data augmentation efficiently. All the operations are eagerly executed on the fly, and the training samples are generated in the memory just before feeding it into the DNNs.
- Since we maintain the dataflow from the raw waveform to the DNN output, it also promotes model inference and deployment ease. After the DNN has been trained, the data loader can simply tune into the “testing” mode by setting the batch size to one and removing the data slice, data augmentation and data batching modules.

²⁷Weicheng Cai et al. “On-the-Fly Data Loader and Utterance-level Aggregation for Speaker and Language Recognition”. In: submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing (2019).

Network Structure

Feed-forward DNN(FF-DNN)

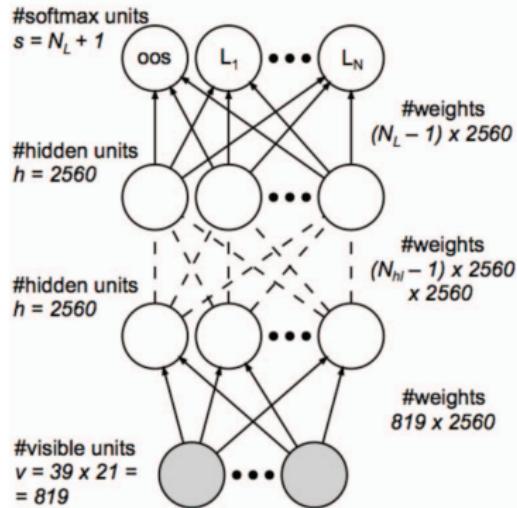


- D-vector for SV [33]²⁸

²⁸Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: Proc. of ICASSP. 2014, pp. 4080–4084.

Network Structure

Feed-forward DNN(FF-DNN)



- FF-DNN for LID [38]²⁹

²⁹I. Lopez-Moreno et al. "Automatic language identification using deep neural networks". In: Proc. of ICASSP. 2014, pp. 5337–5341.

Network Structure

Feed-forward DNN(FF-DNN)

- Text-dependent ("Ok google")



Network Structure

Feed-forward DNN(FF-DNN)

- Text-dependent ("Ok google")
- Short duration ($\leq 3\text{s}$ test segment)

Network Structure

Feed-forward DNN(FF-DNN)

- Text-dependent ("Ok google")
- Short duration ($\leq 3\text{s}$ test segment)
- Fixed-length flattened input (Stacked frames)

Network Structure

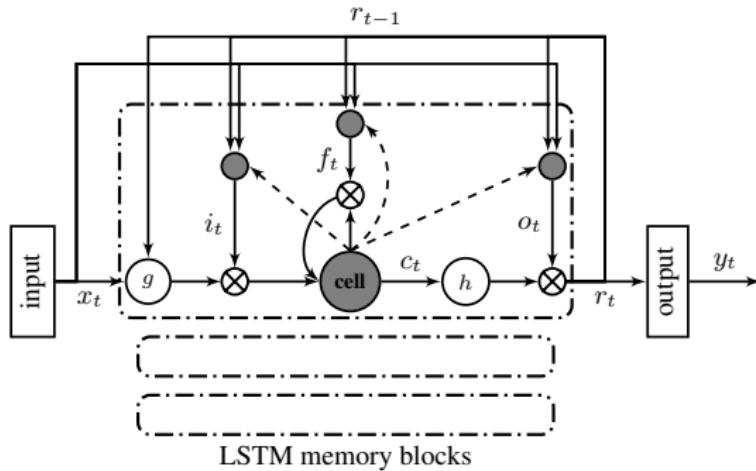
Feed-forward DNN(FF-DNN)

- Text-dependent ("Ok google")
- Short duration ($\leq 3\text{s}$ test segment)
- Fixed-length flattened input (Stacked frames)
- Fram-level + Post average \rightarrow Utterance-level



Network Structure

RNN/LSTM

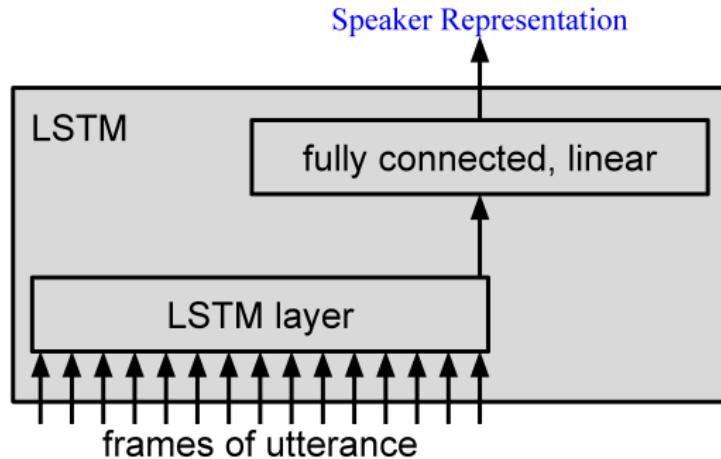


- LSTM for LID [39]³⁰

³⁰J. Gonzalez-Dominguez et al. "Automatic language identification using long short-term memory recurrent neural networks". In: Proc. INTERSPEECH, pp. 2155–2159.

Network Structure

RNN/LSTM

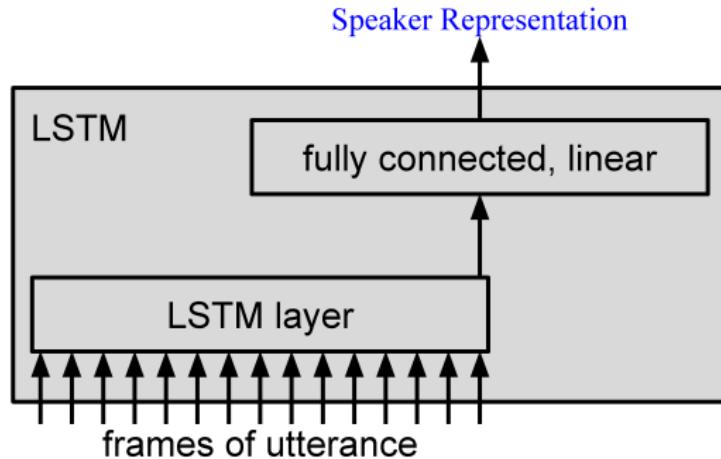


- LSTM for SV [40]³⁰

³⁰ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP 2016*.

Network Structure

RNN/LSTM

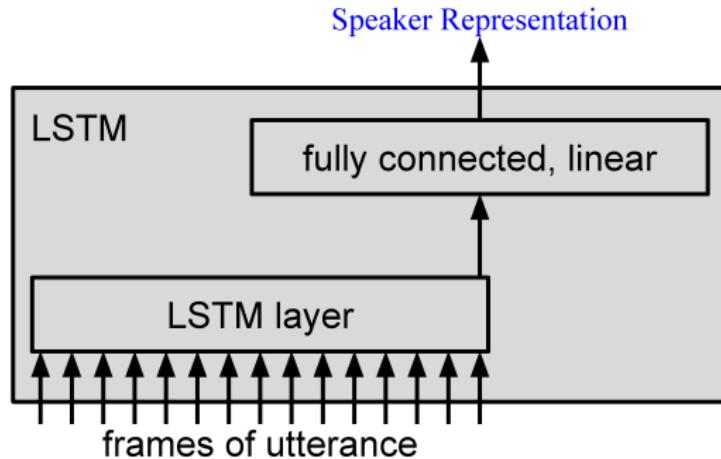


- LSTM for SV [40]³⁰
- Adopt the last several output units of LSTM

³⁰ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP 2016*.

Network Structure

RNN/LSTM

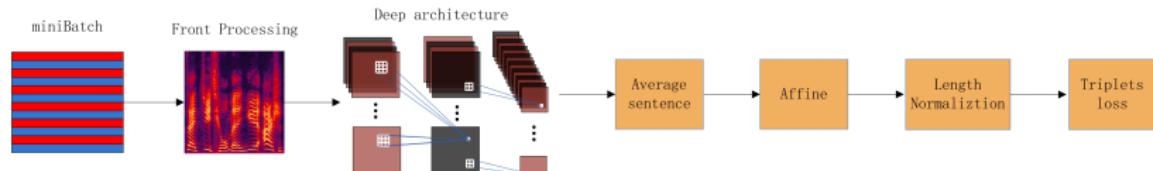


- LSTM for SV [40]³⁰
- Adopt the last several output units of LSTM
- Short duration (\leq 3s test segment)

³⁰ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP 2016*.

Network Structure

CNN

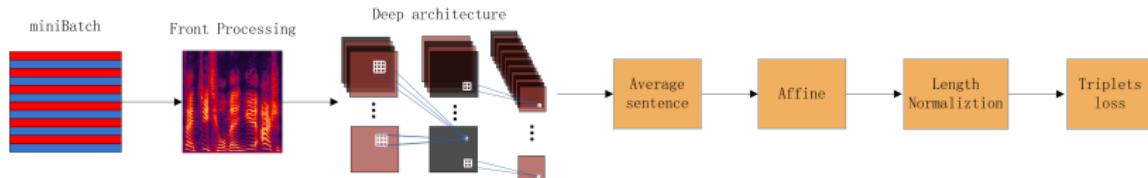


- CNN: Deep Speaker [30]³¹

³¹Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: arXiv e-prints, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

Network Structure

CNN



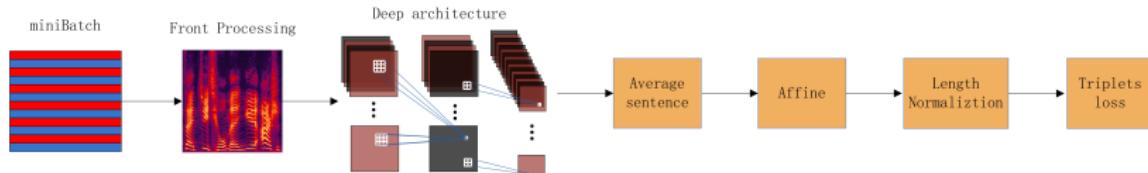
- CNN: Deep Speaker [30]³¹
- Anti-spoofing [41]³²

³¹Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

³²Weicheng Cai et al. "Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion". In: *Proc. of Interspeech* 2017, pp. 217–221.

Network Structure

CNN



- CNN: Deep Speaker [30]³¹
- Anti-spoofing [41]³²
- Speaker and language recognition [42]³³[43]³⁴

³¹Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

³²Weicheng Cai et al. "Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion". In: *Proc. of Interspeech. 2017*, pp. 17–21.

³³Weicheng Cai, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Speaker Odyssey. 2018*, pp. 74–81.

³⁴Chunlei Zhang, Kazuhito Koishida, and John H. L. Hansen. "Text-independent Speaker Verification Based on Triplet Convolutional Neural Network Embedding". In: *IEEE/ACM Transactions on Audio Speech & Language Processing* 26.9 (2018), pp. 1633–1644.

Network Structure

TDNN

Layer	Layer context	Total context	Input x output
frame1	$[t - 2, t + 2]$	5	120x512
frame2	$\{t - 2, t, t + 2\}$	9	1536x512
frame3	$\{t - 3, t, t + 3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	$1500T \times 3000$
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512x N

- x-vector $[36]^{35}$

³⁵ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: Proc. of ICASSP. IEEE, 2018, pp. 5329–5333.

Encoding Mechanism

Conventional approaches

- Average: An utterance-level embedding is derived by averaging the frame-level DNN hidden layer output. (D-vector)



Encoding Mechanism

Conventional approaches

- Average: An utterance-level embedding is derived by averaging the frame-level DNN hidden layer output. (D-vector)
- Average: An utterance-level scores is derived by averaging the frame-level DNN output posteriors.



Encoding Mechanism

Conventional approaches

- Average: An utterance-level embedding is derived by averaging the frame-level DNN hidden layer output. (D-vector)
- Average: An utterance-level scores is derived by averaging the frame-level DNN output posteriors.
- Voting: An utterance-level results is derived by voting the frame-level DNN predictions.



Encoding Mechanism

Encoding layer

- Recurrent layer (Context-dependent)



Encoding Mechanism

Encoding layer

- Recurrent layer (Context-dependent)
 - LSTM/GRU encoding[39]³⁶

³⁶ J. Gonzalez-Dominguez et al. "Automatic language identification using long short-term memory recurrent neural networks". In: *Proc. INTERSPEECH*, pp. 2155–2159.

Encoding Mechanism

Encoding layer

- Recurrent layer (Context-dependent)
 - LSTM/GRU encoding[39]³⁶
 - LSTM/GRU + Attention [44]³⁷

³⁶ J. Gonzalez-Dominguez et al. "Automatic language identification using long short-term memory recurrent neural networks". In: *Proc. INTERSPEECH*, pp. 2155–2159.

³⁷ Wang Geng et al. "End-to-End Language Identification Using Attention-Based Recurrent Neural Networks." In: *Proc. INTERSPEECH*. 2016, pp. 2944–2948.

Encoding Mechanism

Encoding layer

- Recurrent layer (Context-dependent)
 - LSTM/GRU encoding [39]³⁶
 - LSTM/GRU + Attention [44]³⁷
 - Bi-LSTM + Attention [45]³⁸

³⁶ J. Gonzalez-Dominguez et al. "Automatic language identification using long short-term memory recurrent neural networks". In: *Proc. INTERSPEECH*, pp. 2155–2159.

³⁷ Wang Geng et al. "End-to-End Language Identification Using Attention-Based Recurrent Neural Networks." In: *Proc. INTERSPEECH*. 2016, pp. 2944–2948.

³⁸ W. Cai et al. "Utterance-level end-to-end language identification using attention-based CNN-BLSTM". In: *Proc. ICASSP*. 2019.

Encoding Mechanism

- Pooling layer (Context-independent)

Encoding Mechanism

- Pooling layer (Context-independent)
 - Temporal pooling (mean) [30]³⁹

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

Encoding Mechanism

- Pooling layer (Context-independent)
 - Temporal pooling (mean) [30]³⁹
 - Statistics pooling (mean + std) [36]⁴⁰

³⁹Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

Encoding Mechanism

- Pooling layer (Context-independent)

- Temporal pooling (mean) [30]³⁹
- Statistics pooling (mean + std) [36]⁴⁰
- Bilinear pooling [46]⁴¹

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹ J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

Encoding Mechanism

- Pooling layer (Context-independent)

- Temporal pooling (mean) [30]³⁹
- Statistics pooling (mean + std) [36]⁴⁰
- Bilinear pooling [46]⁴¹
- Self-attentive pooling (mean) [47]⁴²

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹ J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

⁴² G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification" In: *Proc. Interspeech*. 2017, pp. 1517–1521.



Encoding Mechanism

- Pooling layer (Context-independent)

- Temporal pooling (mean) [30]³⁹
- Statistics pooling (mean + std) [36]⁴⁰
- Bilinear pooling [46]⁴¹
- Self-attentive pooling (mean) [47]⁴²
- Attentive statistics pooling (mean + std) [48]⁴³ [49]⁴⁴

³⁹Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

⁴²G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech*. 2017, pp. 1517–1521.

⁴³Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech*. 2018, pp. 2252–2256.

⁴⁴Yingke Zhu et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3573–3577.

Encoding Mechanism

- Pooling layer (Context-independent)

- Temporal pooling (mean) [30]³⁹
- Statistics pooling (mean + std) [36]⁴⁰
- Bilinear pooling [46]⁴¹
- Self-attentive pooling (mean) [47]⁴²
- Attentive statistics pooling (mean + std) [48]⁴³ [49]⁴⁴
- Multi-head attentive pooling [50]⁴⁵

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹ J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

⁴² G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech*. 2017, pp. 1517–1521.

⁴³ Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech*. 2018, pp. 2252–2256.

⁴⁴ Yingke Zhu et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3573–3577.

⁴⁵ Yi Liu et al. "Exploring a Unified Attention-Based Pooling Framework for Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 200–204.



Encoding Mechanism

- Pooling layer (Context-independent)

- Temporal pooling (mean) [30]³⁹
- Statistics pooling (mean + std) [36]⁴⁰
- Bilinear pooling [46]⁴¹
- Self-attentive pooling (mean) [47]⁴²
- Attentive statistics pooling (mean + std) [48]⁴³ [49]⁴⁴
- Multi-head attentive pooling [50]⁴⁵
- Learnable dictionary encoding [51]⁴⁶

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹ J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

⁴² G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech*. 2017, pp. 1517–1521.

⁴³ Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech*. 2018, pp. 2252–2256.

⁴⁴ Yingke Zhu et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3573–3577.

⁴⁵ Yi Liu et al. "Exploring a Unified Attention-Based Pooling Framework for Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 200–204.

⁴⁶ W. Cai et al. "A novel learnable dictionary encoding layer for end-to-end language identification". In: *Proc. ICASSP*. 2018, pp. 5189–5193.

Encoding Mechanism

- Pooling layer (Context-independent)
 - Temporal pooling (mean) [30]³⁹
 - Statistics pooling (mean + std) [36]⁴⁰
 - Bilinear pooling [46]⁴¹
 - Self-attentive pooling (mean) [47]⁴²
 - Attentive statistics pooling (mean + std) [48]⁴³ [49]⁴⁴
 - Multi-head attentive pooling [50]⁴⁵
 - Learnable dictionary encoding [51]⁴⁶
 - NetFV/NetVLAD/Ghost VLAD [52]⁴⁷ [53]⁴⁸

³⁹ Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].

⁴⁰ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁴¹ J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.

⁴² G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech*. 2017, pp. 1517–1521.

⁴³ Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech*. 2018, pp. 2252–2256.

⁴⁴ Yingke Zhu et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3573–3577.

⁴⁵ Yi Liu et al. "Exploring a Unified Attention-Based Pooling Framework for Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 200–204.

⁴⁶ W. Cai et al. "A novel learnable dictionary encoding layer for end-to-end language identification". In: *Proc. ICASSP*. 2018, pp. 5189–5193.

⁴⁷ J. Chen et al. "End-to-end Language Identification using NetFV and NetVLAD". In: *Proc. ISCSLP*. 2018, pp. 5189–5193.



Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)

Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)
- Contrastive/Triplet loss [54]⁴⁹ [55]⁵⁰

⁴⁹ Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.

⁵⁰ Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.



Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)
- Contrastive/Triplet loss [54]⁴⁹ [55]⁵⁰
- End-to-End loss [40]⁵¹ [56]⁵²

⁴⁹ Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.

⁵⁰ Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.

⁵¹ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2016.

⁵² Li Wan et al. "Generalized end-to-end loss for speaker verification". In: *Proc. of ICASSP*. 2018, pp. 4879–4883.



Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)
- Contrastive/Triplet loss [54]⁴⁹ [55]⁵⁰
- End-to-End loss [40]⁵¹ [56]⁵²
- Center loss [42]⁵³

⁴⁹ Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.

⁵⁰ Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.

⁵¹ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2016.

⁵² Li Wan et al. "Generalized end-to-end loss for speaker verification". In: *Proc. of ICASSP*. 2018, pp. 4879–4883.

⁵³ Weicheng Cai, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Speaker Odyssey*. 2018, pp. 74–81.



Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)
- Contrastive/Triplet loss [54]⁴⁹ [55]⁵⁰
- End-to-End loss [40]⁵¹ [56]⁵²
- Center loss [42]⁵³
- Angular softmax loss [57]⁵⁴ [42][58]⁵⁵

⁴⁹ Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.

⁵⁰ Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.

⁵¹ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2016.

⁵² Li Wan et al. "Generalized end-to-end loss for speaker verification". In: *Proc. of ICASSP*. 2018, pp. 4879–4883.

⁵³ Weicheng Cai, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Speaker Odyssey*. 2018, pp. 74–81.

⁵⁴ W. Liu et al. "Sphereface: Deep hypersphere embedding for face recognition". In: *Proc. CVPR*. Vol. 1. 2017.

⁵⁵ Zili Huang, Shuai Wang, and Kai Yu. "Angular Softmax for Short-Duration Text-independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3623–3627.



Loss Function

- Standard cross-entropy loss with softmax function (softmax loss)
- Contrastive/Triplet loss [54]⁴⁹ [55]⁵⁰
- End-to-End loss [40]⁵¹ [56]⁵²
- Center loss [42]⁵³
- Angular softmax loss [57]⁵⁴ [42][58]⁵⁵
- Additive margin loss [33]⁵⁶

⁴⁹ Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.

⁵⁰ Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.

⁵¹ Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2016.

⁵² Li Wan et al. "Generalized end-to-end loss for speaker verification". In: *Proc. of ICASSP*. 2018, pp. 4879–4883.

⁵³ Weicheng Cai, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Speaker Odyssey*. 2018, pp. 74–81.

⁵⁴ W. Liu et al. "Sphereface: Deep hypersphere embedding for face recognition". In: *Proc. CVPR*. Vol. 1. 2017.

⁵⁵ Zili Huang, Shuai Wang, and Kai Yu. "Angular Softmax for Short-Duration Text-independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3623–3627.

⁵⁶ Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.

Data Augmentation

- Add noise, music, babble, reverberation [36]⁵⁷

⁵⁷ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁵⁸ Suwon Shon, Ahmed Ali, and James Glass. "Convolutional neural networks and language embeddings for end-to-end dialect recognition". In: *arXiv preprint arXiv:1803.04567* (2018).

⁵⁹ Yexin Yang et al. "Generative Adversarial Networks based X-vector Augmentation for Robust Probabilistic Linear Discriminant Analysis in Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 205–209.



Data Augmentation

- Add noise, music, babble, reverberation [36]⁵⁷
- Speed perturbation [59]⁵⁸

⁵⁷ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁵⁸ Suwon Shon, Ahmed Ali, and James Glass. "Convolutional neural networks and language embeddings for end-to-end dialect recognition". In: *arXiv preprint arXiv:1803.04567* (2018).

⁵⁹ Yexin Yang et al. "Generative Adversarial Networks based X-vector Augmentation for Robust Probabilistic Linear Discriminant Analysis in Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 205–209.



Data Augmentation

- Add noise, music, babble, reverberation [36]⁵⁷
- Speed perturbation [59]⁵⁸
- Generative adversarial network (GAN) [60]⁵⁹

⁵⁷ David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

⁵⁸ Suwon Shon, Ahmed Ali, and James Glass. "Convolutional neural networks and language embeddings for end-to-end dialect recognition". In: *arXiv preprint arXiv:1803.04567* (2018).

⁵⁹ Yexin Yang et al. "Generative Adversarial Networks based X-vector Augmentation for Robust Probabilistic Linear Discriminant Analysis in Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 205–209.



Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted



Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted



Domain Adapation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰

⁶⁰ Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰
- Maximum mean disprepancy [62]⁶¹

⁶⁰Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

⁶¹Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.



Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰
- Maximum mean discrepancy [62]⁶¹
- Autoencoder based domain adaptation (AEDA) [63]⁶²

⁶⁰ Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

⁶¹ Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.

⁶² Suwon Shon et al. "Autoencoder based domain adaptation for speaker recognition under insufficient channel information". In: *arXiv preprint arXiv:1708.01227* (2017).



Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰
- Maximum mean discrepancy [62]⁶¹
- Autoencoder based domain adaptation (AEDA) [63]⁶²
- Domain adversarial training (DAT) [64]⁶³

⁶⁰ Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

⁶¹ Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.

⁶² Suwon Shon et al. "Autoencoder based domain adaptation for speaker recognition under insufficient channel information". In: *arXiv preprint arXiv:1708.01227* (2017).

⁶³ Qing Wang et al. "Unsupervised domain adaptation via domain adversarial training for speaker recognition". In: *Proc. of ICASSP*. 2018, pp. 4889–4893.



Domain Adapatation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰
- Maximum mean discrepancy [62]⁶¹
- Autoencoder based domain adaptation (AEDA) [63]⁶²
- Domain adversarial training (DAT) [64]⁶³
- CORAL [65]⁶⁴

⁶⁰ Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

⁶¹ Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.

⁶² Suwon Shon et al. "Autoencoder based domain adaptation for speaker recognition under insufficient channel information". In: *arXiv preprint arXiv:1708.01227* (2017).

⁶³ Qing Wang et al. "Unsupervised domain adaptation via domain adversarial training for speaker recognition". In: *Proc. of ICASSP*. 2018, pp. 4889–4893.

⁶⁴ Md Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny. "Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation." In: *Proc. of Odyssey*, pp. 176–180.



Domain Adaptation

Traditional domain adaptation is suitable for both the i-vector and deep speaker embedding, performed after the speaker embedding is extracted

- AHC clustering + PLDA adaptation [61]⁶⁰
- Maximum mean discrepancy [62]⁶¹
- Autoencoder based domain adaptation (AEDA) [63]⁶²
- Domain adversarial training (DAT) [64]⁶³
- CORAL [65]⁶⁴
- CORAL+ [66]⁶⁵

⁶⁰Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.

⁶¹Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.

⁶²Suwon Shon et al. "Autoencoder based domain adaptation for speaker recognition under insufficient channel information". In: *arXiv preprint arXiv:1708.01227* (2017).

⁶³Qing Wang et al. "Unsupervised domain adaptation via domain adversarial training for speaker recognition". In: *Proc. of ICASSP*. 2018, pp. 4889–4893.

⁶⁴Md Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny. "Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation." In: *Proc. of Odyssey*, pp. 176–180.

⁶⁵Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka. "The CORAL+ algorithm for unsupervised domain adaptation of PLDA". In: *Proc. of ICASSP*. 2019, pp. 5821–5825.



Domain Adaptation

End-to-End Domain adaptation

- End-to-end adversarial training [67]⁶⁶

⁶⁶ G. Bhattacharya, J. Alam, and P. Kenny. "Adapting End-to-end Neural Speaker Verification to New Languages and Recording Conditions with Adversarial Training". In: *Proc. of ICASSP. 2019*, pp. 6041–6045.

⁶⁷ J. Zhou et al. "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding". In: *Proc. of ICASSP. 2019*, pp. 6196–6200.

Domain Adaptation

End-to-End Domain adaptation

- End-to-end adversarial training [67]⁶⁶
- Generative adversarial network (GAN) [68]⁶⁷

⁶⁶ G. Bhattacharya, J. Alam, and P. Kenny. "Adapting End-to-end Neural Speaker Verification to New Languages and Recording Conditions with Adversarial Training". In: *Proc. of ICASSP. 2019*, pp. 6041–6045.

⁶⁷ Gautam Bhattacharya et al. "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification". In: *Proc. of ICASSP. 2019*, pp. 6226–6230.

⁶⁸ J. Zhou et al. "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding". In: *Proc. of ICASSP. 2019*, pp. 6196–6200.

Domain Adaptation

End-to-End Domain adaptation

- End-to-end adversarial training [67]⁶⁶
- Generative adversarial network (GAN) [68]⁶⁷
- Multi-task adversarial network [69]⁶⁸

⁶⁶ G. Bhattacharya, J. Alam, and P. Kenny. "Adapting End-to-end Neural Speaker Verification to New Languages and Recording Conditions with Adversarial Training". In: *Proc. of ICASSP. 2019*, pp. 6041–6045.

⁶⁷ Gautam Bhattacharya et al. "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification". In: *Proc. of ICASSP. 2019*, pp. 6226–6230.

⁶⁸ J. Zhou et al. "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding". In: *Proc. of ICASSP. 2019*, pp. 6196–6200.

Table of Contents

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

Speech under Far Field and Complex Environment Settings

- Long range fading
- Room reverberation
 - Early reverberation (reflections within 50 to 100 ms): may improve the received speech quality
 - Late reverberation: smearing spectral-temporal structures, amplifying the low-frequency energy, and flattening the formant transitions, etc
- Complex environmental noises
 - fill in regions with low speech energy in the time-frequency plane and blur the spectral details



Previous Methods on Robust Modeling

- Signal level
 - Dereverberation: linear prediction inverse modulation transfer function filter [70]⁶⁹, weighted prediction error (WPE) [71]⁷⁰

⁶⁹B. J. Borgstrom and A. McCree. "The Linear Prediction Inverse Modulation Transfer Function (IP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition". In: *Proc. ICASSP*. 2012, pp. 4065–4068.

⁷⁰L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

Previous Methods on Robust Modeling

- Signal level
 - Dereverberation: linear prediction inverse modulation transfer function filter [70]⁶⁹, weighted prediction error (WPE) [71]⁷⁰
 - DNN based denoising methods for single-channel speech enhancement [72]⁷¹ [73]⁷² [74]⁷³, [75]⁷⁴

⁶⁹B. J. Borgstrom and A. McCree. "The Linear Prediction Inverse Modulation Transfer Function (IP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition". In: *Proc. ICASSP*. 2012, pp. 4065–4068.

⁷⁰L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷¹X. Zhao, Y. Wang, and D. Wang. "Robust Speaker Identification in Noisy and Reverberant Conditions". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 836–845.

⁷²M. Kolboek, Z. Tan, and J. Jensen. "Speech Enhancement Using Long Short-Term Memory based Recurrent Neural Networks for Noise Robust Speaker Verification". In: *Proc. of SLT*. 2016, pp. 305–311.

⁷³Z. Oo et al. "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification". In: *Proc. of INTERSPEECH*. 2016, pp. 2204–2208.

⁷⁴S. E. Eskimez et al. "Front-End Speech Enhancement for Commercial Speaker Verification Systems". In: *Speech Communication* 99 (2018), pp. 101–113.

Previous Methods on Robust Modeling

- Signal level

- Dereverberation: linear prediction inverse modulation transfer function filter [70]⁶⁹, weighted prediction error (WPE) [71]⁷⁰
- DNN based denoising methods for single-channel speech enhancement [72]⁷¹ [73]⁷² [74]⁷³, [75]⁷⁴
- Beamforming for multi-channel speech enhancement [71]⁷⁵

⁶⁹ B. J. Borgstrom and A. McCree. "The Linear Prediction Inverse Modulation Transfer Function (IP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition". In: *Proc. ICASSP*. 2012, pp. 4065–4068.

⁷⁰ L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷¹ X. Zhao, Y. Wang, and D. Wang. "Robust Speaker Identification in Noisy and Reverberant Conditions". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 836–845.

⁷² M. Kolboek, Z. Tan, and J. Jensen. "Speech Enhancement Using Long Short-Term Memory based Recurrent Neural Networks for Noise Robust Speaker Verification". In: *Proc. of SLT*. 2016, pp. 305–311.

⁷³ Z. Oo et al. "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification". In: *Proc. of INTERSPEECH*. 2016, pp. 2204–2208.

⁷⁴ S. E. Eskimez et al. "Front-End Speech Enhancement for Commercial Speaker Verification Systems". In: *Speech Communication* 99 (2018), pp. 101–113.

⁷⁵ L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.



Previous Methods on Robust Modeling

- Feature level
 - Sub-band Hilbert envelopes based features [76]⁷⁶, [77]⁷⁷

⁷⁶T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.

⁷⁷L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

Previous Methods on Robust Modeling

- Feature level
 - Sub-band Hilbert envelopes based features [76]⁷⁶, [77]⁷⁷
 - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78]⁷⁸

⁷⁶T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.

⁷⁷L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷⁸Q. Jin et al. "Speaker Identification with Distant Microphone Speech". In: *Proc. of ICASSP*. 2010, pp. 4518–4521.

Previous Methods on Robust Modeling

- Feature level

- Sub-band Hilbert envelopes based features [76]⁷⁶, [77]⁷⁷
- Warped minimum variance distortionless response (MVDR) cepstral coefficients [78]⁷⁸
- Blind spectral weighting (BSW) based features [79]⁷⁹

⁷⁶T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.

⁷⁷L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷⁸Q. Jin et al. "Speaker Identification with Distant Microphone Speech". In: *Proc. of ICASSP*. 2010, pp. 4518–4521.

⁷⁹S. O. Sadjadi and J. H. L. Hansen. "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.5 (2014), pp. 937–945.



Previous Methods on Robust Modeling

- Feature level

- Sub-band Hilbert envelopes based features [76]⁷⁶, [77]⁷⁷
- Warped minimum variance distortionless response (MVDR) cepstral coefficients [78]⁷⁸
- Blind spectral weighting (BSW) based features [79]⁷⁹
- Power-normalized cepstral coefficients (PNCC) [80]⁸⁰[81]⁸¹

⁷⁶T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.

⁷⁷L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷⁸Q. Jin et al. "Speaker Identification with Distant Microphone Speech". In: *Proc. of ICASSP*. 2010, pp. 4518–4521.

⁷⁹S. O. Sadjadi and J. H. L. Hansen. "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.5 (2014), pp. 937–945.

⁸⁰Chanwoo Kim and Richard M Stern. "Power-Normalized Cepstral Coefcients (PNCC) for Robust Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24.7 (2016), pp. 1315–1329.

⁸¹D. Cai et al. "The DKU-SMIIP System for the Speaker Recognition Task of the VOiCES from a Distance Challenge". In: *Proc. of INTERSPEECH*. 2019.



Previous Methods on Robust Modeling

- Feature level
 - Sub-band Hilbert envelopes based features [76]⁷⁶, [77]⁷⁷
 - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78]⁷⁸
 - Blind spectral weighting (BSW) based features [79]⁷⁹
 - Power-normalized cepstral coefficients (PNCC) [80]⁸⁰[81]⁸¹
 - DNN bottleneck features [82]⁸², etc.

⁷⁶T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.

⁷⁷L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP*. 2018, pp. 5254–5258.

⁷⁸Q. Jin et al. "Speaker Identification with Distant Microphone Speech". In: *Proc. of ICASSP*. 2010, pp. 4518–4521.

⁷⁹S. O. Sadjadi and J. H. L. Hansen. "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.5 (2014), pp. 937–945.

⁸⁰Chanwoo Kim and Richard M Stern. "Power-Normalized Cepstral Coefcients (PNCC) for Robust Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24.7 (2016), pp. 1315–1329.

⁸¹D. Cai et al. "The DKU-SMIIP System for the Speaker Recognition Task of the VOiCES from a Distance Challenge". In: *Proc. of INTERSPEECH*. 2019.

⁸²T. Yamada, L. Wang, and A. Kai. "Improvement of Distant Talking Speaker Identification Using Bottleneck Features of DNN". In: *Proc. of INTERSPEECH*. 2013, pp. 3661–2664.



Previous Methods on Robust Modeling

- Model level
 - Reverberation matching with multi-condition training models within the UBM or i-vector based front-end systems [83]⁸³,[84]⁸⁴
 - Multi-channel i-vector combination [85]⁸⁵
 - Multi-condition training of PLDA models [86]⁸⁶
- Score level
 - Score normalization [83]⁸⁷
 - Multi-channel score fusion [87]⁸⁸,[88]⁸⁹

⁸³I. Peer, B. Rafaely, and Y. Zigel. "Reverberation Matching for Speaker Recognition". In: *Proc. of ICASSP*. 2008, pp. 4829–4832.

⁸⁴A. R Avila et al. "Improving the Performance of Far-Field Speaker Verification Using Multi-Condition Training: The Case of GMM-UBM and i-Vector Systems". In: *Proc. of INTERSPEECH*. 2014, pp. 1096–1100.

⁸⁵A. Bruttì and A. Abad. "Multi-Channel i-vector Combination for Robust Speaker Verification in Multi-Room Domestic Environments". In: *Proc. of Odyssey*. 2016, pp. 252–258.

⁸⁶D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson. "Multicondition Training of Gaussian Plda Models in i-vector Space for Noise and Reverberation Robust Speaker Recognition". In: *Proc. of ICASSP*. 2012, pp. 4257–4260.

⁸⁷I. Peer, B. Rafaely, and Y. Zigel. "Reverberation Matching for Speaker Recognition". In: *Proc. of ICASSP*. 2008, pp. 4829–4832.

⁸⁸Q. Jin, T. Schultz, and A. Waibel. "Far-Field Speaker Recognition". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.7 (2007), pp. 2023–2032.

⁸⁹M. Ji et al. "Text-Independent Speaker Identification using Soft Channel Selection in Home Robot Environments". In: *IEEE Transactions on Consumer Electronics* 54.1 (2008), pp. 140–144.



Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA
 - Retransmitted speech in reverberant environments

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA
 - Retransmitted speech in reverberant environments
 - Speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA
 - Retransmitted speech in reverberant environments
 - Speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems
- Two interesting findings of end-to-end methods for robust modeling [90]⁹¹

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA
 - Retransmitted speech in reverberant environments
 - Speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems
- Two interesting findings of end-to-end methods for robust modeling [90]⁹¹
 - The performance gain achieves by data augmentation in the end-to-end method is larger than in the i-vector framework

⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.



Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89]⁹⁰
 - X-vector + PLDA
 - Retransmitted speech in reverberant environments
 - Speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems
- Two interesting findings of end-to-end methods for robust modeling [90]⁹¹
 - The performance gain achieves by data augmentation in the end-to-end method is larger than in the i-vector framework
 - For end-to-end methods with data augmentation, speech enhancement algorithms may cause mismatch between the training data (clean and augmented data) and the enhanced testing speech.

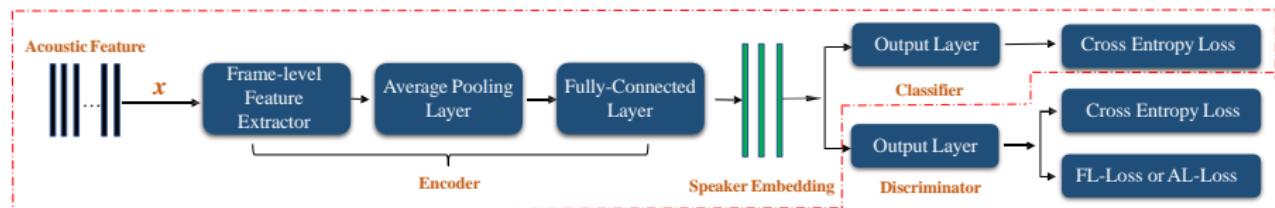
⁹⁰M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.

⁹¹D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.



Robust Modeling of End-to-End Methods

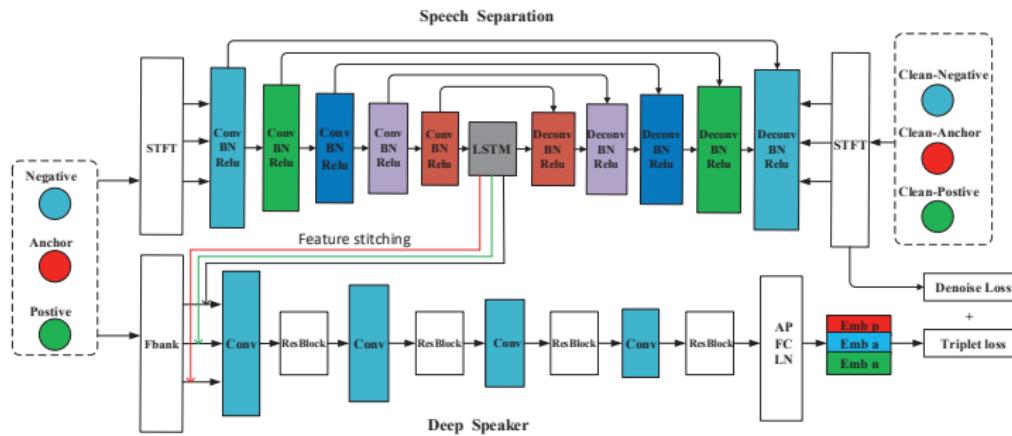
- Multi-task adversarial network for noise-robust speaker embedding [69]⁹²
 - Encoding network for speaker embedding
 - Speaker classifier
 - Noise discriminator
 - Adversarial training by using fix-label loss or anti-label loss (take wrong label with cross entropy) of the noise discriminator
 - Outperform the other methods without adversarial training in noisy environments



⁹²J. Zhou et al. "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding". In: Proc. of ICASSP. 2019, pp. 6196–6200.

Robust Modeling of End-to-End Methods

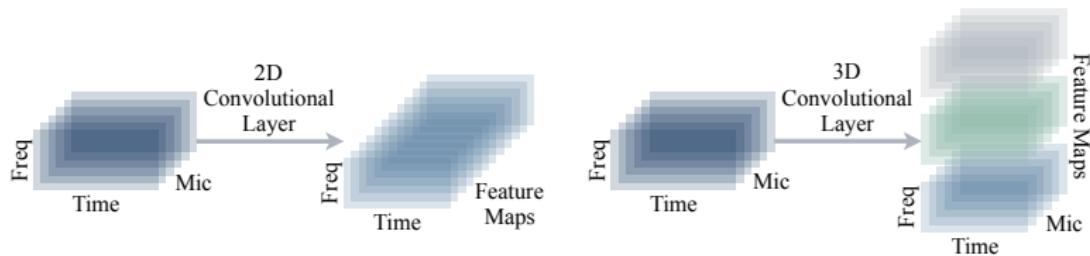
- Joint training of denoising and speaker embedding network[91]⁹³
 - Denoising network
 - extract the target speech from noisy speech
 - extract bottleneck features
 - Speaker embedding network
 - Concatenate bottleneck features with fbank as inputs



⁹³F. Zhao, H. Li, and X. Zhang. "A Robust Text-independent Speaker Verification Method Based on Speech Separation and Deep Speaker". In: Proc. of ICASSP. 2019, pp. 6101–6105.

Robust Modeling of End-to-End Methods

- Multi-channel training framework for speaker recognition under reverberant and noisy environment [90]⁹⁴
 - 3D CNN structure as front-end convolutional network
 - Extract the time-, frequency-, and spatial-information
 - Significantly outperforms the i-vector system with front-end signal enhancement as well as the single-channel robust deep speaker embedding system



⁹⁴D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.

Robust Modeling of End-to-End Methods

- Far-field text-dependent speaker verification [92]⁹⁵
 - Mixed training data with transfer learning
 - Utilize the content and speaker diversity of text-independent data
 - Train model with text-independent data and perform transfer learning with text-dependent data
 - Enrollment data augmentation
 - Enrollment and testing speech can be collected in different environmental settings (e.g. Cell phone enroll, Smart speakers test)
 - Corpus: AISHELL-2019B-eval dataset⁹⁶
 - Open source wake-up words speech database

⁹⁵X. Qin, D Cai, and M. Li. "Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation". In: *Proc. of INTERSPEECH*. 2019.



⁹⁶https://www.aishelltech.com/aishell_2019B_eval

Table of Contents

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

Speaker Diarization

Speaker diarization is a task of “who spoke when” [93]⁹⁷[94]⁹⁸. In general, it consists of four essential submodules:

- ① Voice activity detection (VAD): remove nonspeech from audios.
- ② Speech segmentation: split speech into speaker-homogeneous segments.
- ③ Similarity measurement: compute the speaker similarity of any two segments in the same audio.
- ④ Clustering: cluster segments belonging to the same speaker.

Other submodules like resegmentation and overlap detection are optional.

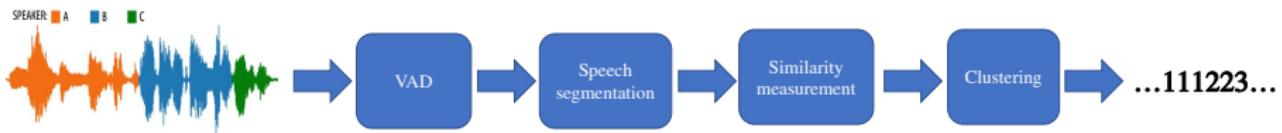


Figure: Essential submodules in diarization.

⁹⁷ S. E. Tranter and D. A. Reynolds. “An Overview of Automatic Speaker Diarization Systems”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5 (2006), pp. 1557–1565. ISSN: 1558-7916.

⁹⁸ X. Anguera et al. “Speaker Diarization: A Review of Recent Research”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 356–370. ISSN: 1558-7916.

Speaker Diarization: VAD

- ① Discriminant classifiers: e.g. in [95]⁹⁹ and [96]¹⁰⁰, linear discriminant analysis (LDA) and support vector machine are used for classifying MFCC frames into speech/non-speech.
- ② Recently, DNN-based discriminant classifiers for VAD become popular. In [97]¹⁰¹, the LSTM architecture is employed for sequential modeling of the VAD task and shows state-of-the-art performance.

⁹⁹ Elias Rentzeperis et al. "The 2006 Athens Information Technology Speech activity detection and speaker diarization systems". In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer. 2006, pp. 385–395.

¹⁰⁰ Andrey Temko, Dusan Macho, and Climent Nadeu. "Enhanced SVM training for robust speech activity detection". In: *Proc. of ICASSP*. Vol. 4. IEEE. 2007.

¹⁰¹ Florian Eyben et al. "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies". In: *Proc. of ICASSP*. IEEE. 2013, pp. 483–487.

Speaker Diarization: Speech Segmentation

- ① **SCD:** Speaker changepoint detection (SCD) usually searches for change points first and then splits speech into speaker-homogeneous segments.

The first general approach brought up by [98]¹⁰² is a variation on the Bayesian information criterion (BIC) [99]¹⁰³. This technique applies a sliding window over speech data. It determines whether current windowed speech is better modelled by a single distribution (no change point, H_0) or two different distributions (change point, H_1) by computing BIC scores.

- ② Generalized likelihood ratio (GLR) [100]¹⁰⁴.

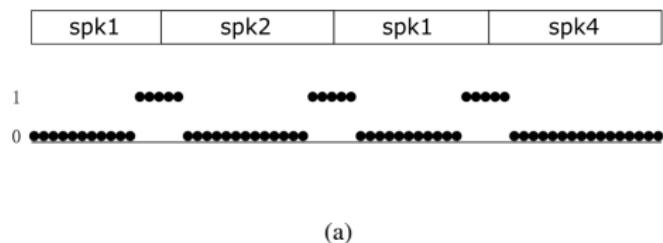
¹⁰² Douglas A Reynolds and P Torres-Carrasquillo. *The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations*. Tech. rep. 2004.

¹⁰³ Scott Chen, Ponani Gopalakrishnan, et al. "Speaker, environment and channel change detection and clustering via the bayesian information criterion". In: *Proc. DARPA broadcast news transcription and understanding workshop*. Vol. 8. Virginia, USA. 1998, pp. 127–132.

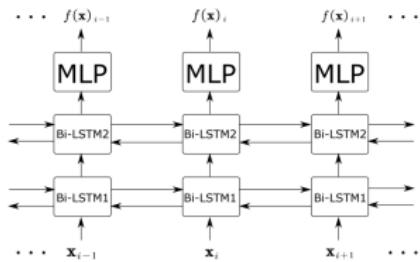
¹⁰⁴ Herbert Gish, M-H Siu, and Robin Rohlicek. "Segregation of speakers for speech recognition and speaker identification". In: *Proc. of ICASSP*. IEEE. 1991, pp. 873–876.

Speaker Diarization: Speech Segmentation

- ② Recently DNN-based SCD models have also been proposed. For example, [101]¹⁰⁵ labels speaker change points and their collars of 0.5s as 1, while the rest as 0, and carries out a 2-layer LSTM training.



(a)



(b)

- ③ However, SCD only provides an initial base segmentation in diarization, which will be clustered and often resegmented later. According to [102]¹⁰⁶ and [103]¹⁰⁷, using a simple initial uniform segmentation instead doesn't significantly degrade the overall diarization performance.

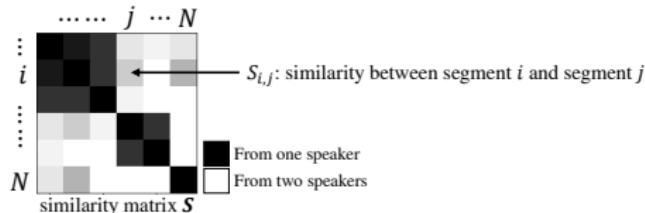
¹⁰⁵ Ruiqing Yin, Herv Bredin, and Claude Barras. "Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks". In: *Proc. Interspeech*. 2017, pp. 3827–3831.

¹⁰⁶ Chuck Wooters et al. "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system". In: *RT-04F Workshop*. Vol. 23. 2004, p. 23.

¹⁰⁷ Sylvain Meignier et al. "Step-by-step and integrated approaches in broadcast news speaker diarization". In: *Computer Speech & Language* 20.2-3 (2006), pp. 303–330.

Speaker Diarization: Similarity Measurement

- ① BIC and GLR measurement can be applied here [104]¹⁰⁸.
- ② [105]¹⁰⁹ first extracts i-vectors from segments and then measures their speaker similarity using cosine distance or PLDA [24]¹¹⁰. I-vector is later substituted with other speaker embeddings like d-vector [106]¹¹¹ and x-vector [107]¹¹² to improve the precision.
- ③ [108]¹¹³ adopts LSTM to infer the similarity matrix directly.



¹⁰⁸ Elie El Khoury, Christine Senac, and Régine André-Obrecht. "Speaker diarization: towards a more robust and portable system". In: *Proc. of ICASSP*. Vol. 4. IEEE. 2007, pp. IV-489.

¹⁰⁹ G. Sell and D. Garcia-Romero. "Speaker diarization with plda i-vector scoring and unsupervised calibration". In: *IEEE Spoken Language Technology Workshop*. 2014, pp. 413–417.

¹¹⁰ S.J.D. Prince and J.H. Elder. "Probabilistic linear discriminant analysis for inferences about identity". In: *Proc. ICCV*. 2017.

¹¹¹ Q. Wang et al. "Speaker Diarization with LSTM". In: *Proc. of ICASSP*. 2018, pp. 5239–5243.

¹¹² Gregory Sell et al. "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge". In: *Proc. Interspeech*. 2018, pp. 2808–2812.

¹¹³ Qingjian Lin et al. "LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization". In: *Proc. Interspeech*. 2019.

Speaker Diarization: Clustering

The purpose of this stage is to associate or cluster segments from the same speaker together.

- ① **AHC**: Agglomerative hierarchical clustering (AHC), as a widely used clustering algorithm, is presented as a binary-tree building process [109]¹¹⁴.
- ② **Spectralclustering** is employed instead of AHC in [106]¹¹⁵ and [108]¹¹⁶. Spectral clustering is a graph-based clustering algorithm [110]¹¹⁷. Given the similarity matrix \mathbf{S} , it considers $S_{i,j}$ as the weight of the edge between nodes i and j in an undirected graph. By removing weak edges with small weights, spectral clustering divides the original graph into subgraphs.
- ③ **UIS – RNN**: In [111]¹¹⁸, the similarity measurement and clustering steps are replaced by the Unbounded Interleaved-State (UIS) RNN model.

¹¹⁴K. Chidananda Gowda and G. Krishna. "Agglomerative Clustering Using the Concept of Mutual Nearest Neighbourhood". In: *Pattern Recognition* 10 (1978), pp. 105–112.

¹¹⁵Q. Wang et al. "Speaker Diarization with LSTM". In: *Proc. of ICASSP*. 2018, pp. 5239–5243.

¹¹⁶Qingjian Lin et al. "LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization". In: *Proc. Interspeech*. 2019.

¹¹⁷Ulrike von Luxburg. "A Tutorial on Spectral Clustering". In: *Statistics and Computing* 17 (2007), pp. 395–416.

¹¹⁸Aonan Zhang et al. "Fully Supervised Speaker Diarization". In: *Proc. of ICASSP*. 2019.

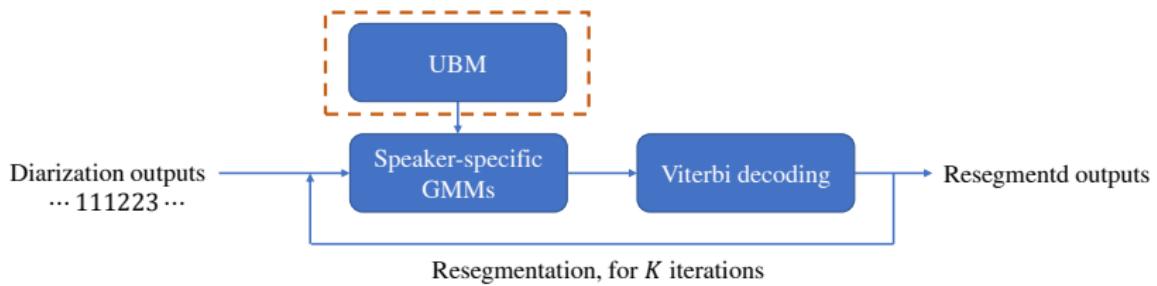


Speaker Diarization: Others

① **Resegmentation:** Re-segmentation is an optional submodule of diarization, aiming at refining the original segment boundaries and filling in short segments that may have been removed for more robust processing in the clustering stage.

Traditionally, a Viterbi decoder with or without iteration is employed. First, speaker-specific GMMs are trained according to diarization outputs, and then data frames are realigned to GMMs with the maximum posterior probabilities.

An improved version is the VB resegmentation [112]¹¹⁹[113]¹²⁰. It builds the speaker-specific GMMs by adapting limited data frames of target speakers to UBM, which enhances the robustness.



¹¹⁹Xianhong Chen et al. "VB-HMM Speaker Diarization with Enhanced and Refined Segment Representation." In: *Proc. of Odyssey*. 2018, pp. 134–139.

¹²⁰Mireia Diez, Lukáš Burget, and Pavel Matejka. "Speaker Diarization based on Bayesian HMM with Eigenvoice Priors." In: *Proc. of Odyssey*. 2018.

Speaker Diarization: Others

- ② **Overlap detection:** Overlap errors account for a large percent of DER in diarization tasks, for example, about 10% in DIHARD. However, current techniques are mostly migrated from VAD, such as two-stage HMMs and DNN-based binary classifiers. They are proved not so efficient in this task.
- ③ [114]¹²¹ reported the improvement after overlap detection, from DER 27.85% to 27.44% on the DIHARD2018 dev dataset. In DIHARD2019, our team also carried out experiments and got similar results.
- ④ Therefore, overlap detection might become one of the most challenging and attractive research directions in speaker diarization.

¹²¹Ondrej Novotný et al. "BUT system for DIHARD speech diarization challenge 2018". In (2018).

Paralinguistic Speech Attribute Recognition

① Background

② Features:

- ① Frame-level features [115, 116, 117]
- ② Utterance-level features [115, 116, 117]

③ Network Structure:

- ① Frame-level DNN structure [118, 119, 120, 121]
- ② Convolutional Network [122, 123, 124]
- ③ Recurrent Network [125, 126, 127]
- ④ Convolutional Recurrent Neural Network [128, 129, 130]

④ Back-end Classifier [126, 131, 124]



Paralinguistic Speech Attribute Recognition

Background

Paralinguistic speech attribute recognition is a task to classify the attributes in speech signals automatically [115]¹²²[116]¹²³ [117]¹²⁴[132]¹²⁵. Since 2009, the Interspeech Computational Paralinguistics ChallengE (ComParE) is held every year to explore the technologies of this area. Topics in recent years include

- ① ComParE2017: Addressee, Cold and Snoring
- ② ComParE2018: Atypical and Self-Assessed Affect, Crying and Heart Beats
- ③ ComParE2019: Styrian Dialects, Continuous Sleepiness, Baby Sounds and Orca Activity

Traditional systems include two steps, utterance-level feature extraction and back-end classifiers training. And Recently, with the development of deep learning algorithm, many end-to-end solutions are proposed.

¹²² Björn W Schuller et al. "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity". In: *Proc. of INTERSPEECH*. 2019.

¹²³ Björn Schuller et al. "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats". In: *Proc. of Interspeech*. 2018, pp. 122–126. URL: <http://dx.doi.org/10.21437/Interspeech.2018-51>.

¹²⁴ Björn Schuller et al. "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring". In: *Proc. of Interspeech*. 2017, pp. 3442–3446. URL: <http://dx.doi.org/10.21437/Interspeech.2017-43>.

¹²⁵ Björn Schuller et al. "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism". In: *Proc. of INTERSPEECH*. 2013, pp. 148–152.



Paralinguistic Speech Attribute Recognition

Frame-level features

- ① STFT-spectrogram [133]¹²⁶[128]¹²⁷
- ② Mel-spectrogram [134]¹²⁸
- ③ Constant Q Transform (CQT) [133][135]¹²⁹
- ④ LPCC, MFCC, RASTA-PLP
- ⑤ etc

¹²⁶Danwei Cai et al. "End-to-End Deep Learning Framework for Speech Paralinguistics Detection Based on Perception Aware Spectrum". In: *Proc. of INTERSPEECH*. 2017, pp. 3452–3456.

¹²⁷Dengke Tang, Junlin Zeng, and Ming Li. "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals". In: *Proc. of INTERSPEECH*. 2018, pp. 162–166.

¹²⁸Mario Lasseck. "Audio-based bird species identification with deep convolutional neural networks". In: *Working Notes of CLEF 2018* (2018).

¹²⁹Massimiliano Todisco, Hctor Delgado, and Nicholas Evans. "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification". In: *Computer Speech & Language*: 45 (Feb. 2017).



Paralinguistic Speech Attribute Recognition

Utterance-level features [115, 116, 117]

- ① Handcrafted features: OpenSMILE ComParE set [9]¹³⁰
- ② Features extracted from unsupervised models. These features summarize local features descriptors in a vectorial statistic.
 - ① Fisher Encoding: train a GMM model as a visual word dictionary, extract features by storing a statistics of the difference between dictionary elements.
 - ② Bag-of-Audio-Word: quantize based on a codebook, represent audio chunks as histograms of acoustic LLDs.
 - ③ AuDeep feature: obtained from unsupervised representation learning with recurrent sequence to sequence autoencoders.
- ③ Supervised deep neural network based features
 - ① Output posteriors: output probabilities of network.
 - ② Embeddings: extracted from the penultimate layer in the network.

¹³⁰ Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proc. of ACM Multimedia*. 2010, pp. 1459–1462.



Paralinguistic Speech Attribute Recognition

Network Structure DNN

Frame-level DNN is an effective structure in the field of paralinguistic attribute recognition [118]¹³¹[119]¹³²[120]¹³³[121]¹³⁴. The algorithm consists of the following steps.

- ① Extract frame-level features.
- ② Train a frame-level DNN model.
- ③ Obtain frame-level DNN posteriors
- ④ Average the frame-level scores or train an extra classifiers to generate final scores

¹³¹ Gbor Gosztolya, Tams Grsz, and Lszl Tth. "General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats". In: *Proc. of Interspeech*. 2018, pp. 531–535.

¹³² Gbor Gosztolya et al. "DNN-Based Feature Extraction and Classifier Combination for Child-Directed Speech, Cold and Snoring Identification". In: *Proc. of Interspeech*. 2017, pp. 3522–3526.

¹³³ Gbor Gosztolya et al. "Estimating the Sincerity of Apologies in Speech by DNN Rank Learning and Prosodic Analysis". In: *Proc. of Interspeech*. 2016, pp. 2026–2030.

¹³⁴ Yishan Jiao et al. "Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features". In: *Proc. of Interspeech*. 2016, pp. 2388–2392.



Paralinguistic Speech Attribute Recognition

Network Structure CNN

Front-end CNN structures can be considered as a local pattern extractor. The CNN system can directly utilize the output as final decision or train an extra classifier or regressor to generate scores.

- ① Plain CNN [122]¹³⁵
- ② Residual structure [123]¹³⁶
- ③ 1D-CNN structure [124]¹³⁷
- ④ etc.

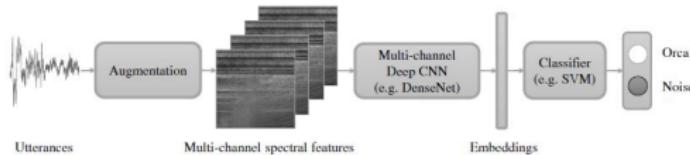


Figure: Structure of the proposed deep embedding system for orca activity detection in [136].

¹³⁵ Johannes Wagner et al. "Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?" In: *Proc. of Interspeech*. 2018, pp. 147–151.

¹³⁶ Q.F. Tan, P.G. Georgiou, S.S. Narayanan, et al. "Enhanced sparse imputation techniques for a robust speech recognition front-end". In: *IEEE Transactions on Audio Speech and Language Processing* 19.8 (2011), p. 2418.

¹³⁷ Ahmed Imtiaz Humayun et al. "An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification". In: *Proc. of Interspeech*. 2018, pp. 127–131.

Paralinguistic Speech Attribute Recognition

Network Structure RNN

Recurrent neural network can take sequential information into consideration.

- ① LSTM Network [125]¹³⁸
- ② Bi-LSTM Network [126]¹³⁹
- ③ Attention structure [126]¹⁴⁰[127]¹⁴¹
- ④ etc.

¹³⁸ Heysem Kaya et al. "LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition". In: *Proc. of Interspeech*. 2018, pp. 521–525.

¹³⁹ Bo-Hao Su et al. "Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features". In: *Proc. of Interspeech*. 2018, pp. 536–540.

¹⁴⁰ Bo-Hao Su et al. "Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features". In: *Proc. of Interspeech*. 2018, pp. 536–540.

¹⁴¹ Cristina Gorrostieta et al. "Attention-based Sequence Classification for Affect Detection". In: *Proc. of Interspeech*. 2018, pp. 506–510.

Paralinguistic Speech Attribute Recognition

Network Structure CRNN

Convolutional Recurrent Neural Network is popular in paralinguistic recognition task recently and in some tasks it achieve the state-of-the-art performance [128]¹⁴²[129]¹⁴³[130]¹⁴⁴. The RNN structure includes GRU, LSTM, BLSTM.

¹⁴²Dengke Tang, Junlin Zeng, and Ming Li. "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals". In: *Proc. of INTERSPEECH*. 2018, pp. 162–166.

¹⁴³Ming Li et al. "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder". In: *Computer Speech & Language* 56 (2019), pp. 80–94.

¹⁴⁴Danqing Luo, Yuexian Zou, and Dongyan Huang. "Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition". In: *Proc. of Interspeech*. 2018, pp. 152–156.



Paralinguistic Speech Attribute Recognition

Back-end Classifier

Due to the lack of large scale training data in this task, back-end classifiers such as SVM [126]¹⁴⁵[131]¹⁴⁶, LDA and MLP [124]¹⁴⁷ are still employed in many situations. Back-end classifiers are commonly applied on

- ① CNN/RNN/CRNN embeddings extracted from the penultimate layer of the network
- ② DNN posteriors directly obtained from networks' output
- ③ CNN embeddings concatenated with handcrafted features

¹⁴⁵ Bo-Hao Su et al. "Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features". In: *Proc. of Interspeech*. 2018, pp. 536–540.

¹⁴⁶ Shahin Amiriparian et al. "Snore Sound Classification Using Image-Based Deep Spectrum Features". In: *Proc. of INTERSPEECH*. 2017, pp. 3512–3516.

¹⁴⁷ Ahmed Imtiaz Humayun et al. "An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification". In: *Proc. of Interspeech*. 2018, pp. 127–131.



Anti-spoofing countermeasures

- Severe vulnerability of state-of-the-art ASV system under a diverse range of intentional fraudulent.
- Physical access scenario: replayed recording
- Logical access scenario: synthesised audio, e.g. text-to-speech and voice conversion
- Anti-spoofing: Develop countermeasure system to distinguish between the bona fide and the spoof audio.

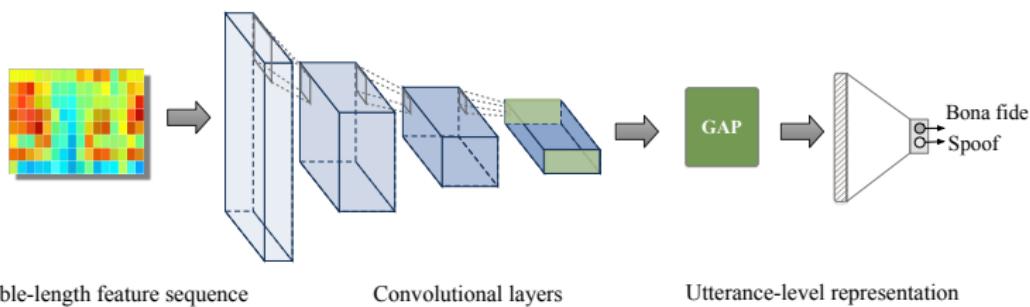
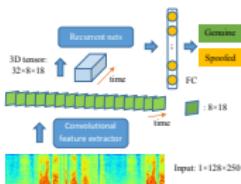
Anti-spoofing countermeasures

- Input: Variable-length audio waveform
- Output: Utterance-level attribute (bona fide or spoof)
- The same processing pipeline as speaker and language recognition
- But we fed different types of features to the network



Anti-spoofing countermeasures

Network structure: CNN or CNN + RNN architecture [137]¹⁴⁸ [138]¹⁴⁹ [139]¹⁵⁰



¹⁴⁸ Chunlei Zhang, Chengzhu Yu, and John HL Hansen. "An investigation of deep-learning frameworks for speaker verification antispoofing". In: *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2017), pp. 684–694.

¹⁴⁹ Galina Lavrentyeva et al. "Audio Replay Attack Detection with Deep Learning Frameworks." In: *Proc. of Interspeech*. 2017, pp. 82–86.

¹⁵⁰ Francis Tom, Mohit Jain, and Prasenjit Dey. "End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention." In: *Proc. of Interspeech*. 2018, pp. 681–685.

Anti-spoofing countermeasures

Feature representation

- Phase information, e.g. Modified group delay feature (MODGDF)
- High frequency information (CQCC/LFCC/IMFCC)
- High resolution representation (STFT gram, Group delay gram)



Summary

- 1 Problem Formulation
- 2 Traditional Framework
 - Feature Extraction
 - Representation
 - Variability Compensation
 - Backend Classification
- 3 End-to-End Deep Neural Network based Framework
 - System Pipeline
 - Data Preparation
 - Network Structure
 - Encoding Mechanism
 - Loss Function
 - Data Augmentation
 - Domain Adaptation
- 4 Robust Modeling of End-to-End methods
 - Speech under Far Field and Complex Environment Settings
 - Previous Methods on Robust Modeling
 - Robust Modeling of End-to-End Methods
- 5 Other Applications of End-to-End Methods
 - Speaker Diarization
 - Paralinguistic Speech Attribute Recognition
 - Anti-spoofing Countermeasures

Thank you very much!

Email: ming.li369@duke.edu

Website: <https://scholars.duke.edu/person/MingLi>

Slide Download Link: <https://sites.duke.edu/dkusmiip>



References I

- [1] P. Torres-Carrasquillo et al. "Approaches to language identification using gaussian mixture models and shifted delta cepstral features". In: *Proc. of ICSLP*. 2002, pp. 89–92.
- [2] C. Kim and R. M. Stern. "Power-Normalized Cepstral Coefficients PNCC for Robust Speech Recognition". In: *IEEE Transactions on Audio Speech and Language Processing* 24.7 (2016), pp. 1315–1329.
- [3] Shao Yang and De Liang Wang. "Robust speaker identification using auditory features and computational auditory scene analysis". In: *Proc. of ICASSP*. 2008.
- [4] Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification". In: *Computer Speech and Language* 45 (2017).
- [5] Pavel Matejka et al. "Neural Network Bottleneck Features for Language Identification." In: *Proc. of Odyssey*. 2014.
- [6] Achintya K Sarkar et al. "Combination of cepstral and phonetically discriminative features for speaker verification". In: *IEEE Signal Processing Letters* 21.9 (2014), pp. 1040–1044.
- [7] Ming Li and Wenbo Liu. "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features". In: *Proc. of Interspeech*. 2014.
- [8] F. Richardson, D. Reynolds, and N. Dehak. "Deep Neural Network Approaches to Speaker and Language Recognition". In: *IEEE Signal Processing Letters* 22.10 (2015), pp. 1671–1675.
- [9] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proc. of ACM Multimedia*. 2010, pp. 1459–1462.
- [10] Hamid Behravan et al. "Introducing attribute features to foreign accent recognition". In: *Proc. of ICASSP*. IEEE. 2014, pp. 5332–5336.
- [11] Ming Li et al. "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals". In: *Computer speech & language* 36 (2016), pp. 196–211.



References II

- [12] Jinxi Guo et al. "Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features." In: *Proc. of INTERSPEECH*. 2016, pp. 2219–2222.
- [13] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. "A comparison of features for synthetic speech detection". In: (2015).
- [14] Zhizheng Wu, Eng Siong Chng, and Haizhou Li. "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition". In: *Proc. of Interspeech*. 2012.
- [15] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models". In: *Digital Signal Processing*. 2000, 1941.
- [16] W.M Campbell et al. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: *Proc. of ICASSP*. Vol. 1. 2006, pp. 97–100.
- [17] Andreas Stolcke et al. "MLLR transforms as features in speaker recognition". In: *Ninth European Conference on Speech Communication and Technology*. 2005.
- [18] N. Dehak et al. "Front-end factor analysis for speaker verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.
- [19] Patrick Kenny et al. "Joint factor analysis versus eigenchannels in speaker recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1435–1447.
- [20] A.O. Hatch, S. Kajarekar, and A. Stolcke. "Within-class covariance normalization for SVM-based speaker recognition". In: *Proc. of INTERSPEECH*. Vol. 4. 2006, pp. 1471–1474.
- [21] Seyed Omid Sadjadi, Jason Pelecanos, and Weizhong Zhu. "Nearest neighbor discriminant analysis for robust speaker recognition". In: *Proc. of Interspeech*. 2014.
- [22] Danwei Cai et al. "Locality sensitive discriminant analysis for speaker verification". In: *Proc. of APSIPA ASC*. 2016, pp. 1–5.



References III

- [23] Peng Shen et al. "Local fisher discriminant analysis for spoken language identification". In: *Proc. of ICASSP*. 2016, pp. 5825–5829.
- [24] S.J.D. Prince and J.H. Elder. "Probabilistic linear discriminant analysis for inferences about identity". In: *Proc. ICCV*. 2017.
- [25] D. Garcia-Romero and C. Y Espy-Wilson. "Analysis of i-vector Length Normalization in Speaker Recognition Systems." In: *Proc. INTERSPEECH*. 2011, pp. 249–252.
- [26] Kyu Jeong Han et al. "TRAP language identification system for RATS phase II evaluation". In: *Proc. of Interspeech*. 2013, pp. 1502–1506.
- [27] Omid Ghahabi et al. "Deep Neural Networks for iVector Language Identification of Short Utterances in Cars". In: *Proc. of Interspeech*. 2016, pp. 367–371.
- [28] Yiyan Wang, Haotian Xu, and Zhijian Ou. "Joint bayesian gaussian discriminant analysis for speaker verification". In: *Proc. of ICASSP*. IEEE. 2017, pp. 5390–5394.
- [29] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>.
- [30] Chao Li et al. "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *arXiv e-prints*, arXiv:1705.02304 (2017), arXiv:1705.02304. arXiv:1705.02304 [cs.CL].
- [31] D. Snyder et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *Proc. IEEE SLT*. 2017.
- [32] Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *Proc. of SLT*. IEEE. 2018, pp. 1021–1028.
- [33] Ehsan Variani et al. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2014, pp. 4080–4084.



References IV

- [34] Yuan Liu et al. "Deep feature for text-dependent speaker verification". In: *Speech Communication* 73 (2015), pp. 1–13.
- [35] Lantian Li et al. "Deep speaker vectors for semi text-independent speaker verification". In: *arXiv preprint arXiv:1505.06427* (2015).
- [36] David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.
- [37] Weicheng Cai et al. "On-the-Fly Data Loader and Utterance-level Aggregation for Speaker and Language Recognition". In: *submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing* (2019).
- [38] I. Lopez-Moreno et al. "Automatic language identification using deep neural networks". In: *Proc. of ICASSP*. 2014, pp. 5337–5341.
- [39] J. Gonzalez-Dominguez et al. "Automatic language identification using long short-term memory recurrent neural networks". In: *Proc. INTERSPEECH*, pp. 2155–2159.
- [40] Georg Heigold et al. "End-to-End Text-Dependent Speaker Verification". In: *Proc. of ICASSP*. 2016.
- [41] Weicheng Cai et al. "Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion". In: *Proc. of Interspeech*. 2017, pp. 17–21.
- [42] Weicheng Cai, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Speaker Odyssey*. 2018, pp. 74–81.
- [43] Chunlei Zhang, Kazuhito Koishida, and John H. L. Hansen. "Text-independent Speaker Verification Based on Triplet Convolutional Neural Network Embedding". In: *IEEE/ACM Transactions on Audio Speech & Language Processing* 26.9 (2018), pp. 1633–1644.
- [44] Wang Geng et al. "End-to-End Language Identification Using Attention-Based Recurrent Neural Networks." In: *Proc. INTERSPEECH*. 2016, pp. 2944–2948.

References V

- [45] W. Cai et al. "Utterance-level end-to-end language identification using attention-based CNN-BLSTM". In: *Proc. ICASSP*. 2019.
- [46] J. Ma et al. "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling". In: *Proc. of INTERSPEECH*, pp. 2571–2575.
- [47] G. Bhattacharya, J. Alam, and P. Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech*. 2017, pp. 1517–1521.
- [48] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech*. 2018, pp. 2252–2256.
- [49] Yingke Zhu et al. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3573–3577.
- [50] Yi Liu et al. "Exploring a Unified Attention-Based Pooling Framework for Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 200–204.
- [51] W. Cai et al. "A novel learnable dictionary encoding layer for end-to-end language identification". In: *Proc. ICASSP*. 2018, pp. 5189–5193.
- [52] J. Chen et al. "End-to-end Language Identification using NetFV and NetVLAD". In: *Proc. ISCSLP*. 2018.
- [53] Weidi Xie et al. "Utterance-level Aggregation for Speaker Recognition in the Wild". In: *Proc. of ICASSP*. 2019, pp. 5791–5795.
- [54] Chunlei Zhang and Kazuhito Koishida. "End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances". In: *Proc. Interspeech*. 2017, pp. 1487–1491.
- [55] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. INTERSPEECH*. 2018, pp. 1086–1090.
- [56] Li Wan et al. "Generalized end-to-end loss for speaker verification". In: *Proc. of ICASSP*. 2018, pp. 4870–4883.



References VI

- [57] W. Liu et al. "Sphereface: Deep hypersphere embedding for face recognition". In: *Proc. CVPR*. Vol. 1. 2017.
- [58] Zili Huang, Shuai Wang, and Kai Yu. "Angular Softmax for Short-Duration Text-independent Speaker Verification." In: *Proc. of Interspeech*. 2018, pp. 3623–3627.
- [59] Suwon Shon, Ahmed Ali, and James Glass. "Convolutional neural networks and language embeddings for end-to-end dialect recognition". In: *arXiv preprint arXiv:1803.04567* (2018).
- [60] Yexin Yang et al. "Generative Adversarial Networks based X-vector Augmentation for Robust Probabilistic Linear Discriminant Analysis in Speaker Verification". In: *Proc. of ISCSLP*. 2018, pp. 205–209.
- [61] Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". In: *Proc. of Odyssey*. 2014.
- [62] Wei-Wei Lin et al. "Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders." In: *Proc. of Odyssey*. 2018, pp. 162–167.
- [63] Suwon Shon et al. "Autoencoder based domain adaptation for speaker recognition under insufficient channel information". In: *arXiv preprint arXiv:1708.01227* (2017).
- [64] Qing Wang et al. "Unsupervised domain adaptation via domain adversarial training for speaker recognition". In: *Proc. of ICASSP*. 2018, pp. 4889–4893.
- [65] Md Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny. "Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation." In: *Proc. of Odyssey*, pp. 176–180.
- [66] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka. "The CORAL+ algorithm for unsupervised domain adaptation of PLDA". In: *Proc. of ICASSP*. 2019, pp. 5821–5825.
- [67] G. Bhattacharya, J. Alam, and P. Kenny. "Adapting End-to-end Neural Speaker Verification to New Languages and Recording Conditions with Adversarial Training". In: *Proc. of ICASSP*. 2019, pp. 6041–6045.
- [68] Gautam Bhattacharya et al. "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification". In: *Proc. of ICASSP*. 2019, pp. 6226–6230.



References VII

- [69] J. Zhou et al. "Training Multi-task Adversarial Network for Extracting Noise-robust Speaker Embedding". In: *Proc. of ICASSP. 2019*, pp. 6196–6200.
- [70] B. J. Borgstrom and A. McCree. "The Linear Prediction Inverse Modulation Transfer Function (IP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition". In: *Proc. ICASSP. 2012*, pp. 4065–4068.
- [71] L. Mosner et al. "Dereverberation and Beamforming in Far-Field Speaker Recognition". In: *Proc. ICASSP. 2018*, pp. 5254–5258.
- [72] X. Zhao, Y. Wang, and D. Wang. "Robust Speaker Identification in Noisy and Reverberant Conditions". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 836–845.
- [73] M. Kolboek, Z. Tan, and J. Jensen. "Speech Enhancement Using Long Short-Term Memory based Recurrent Neural Networks for Noise Robust Speaker Verification". In: *Proc. of SLT. 2016*, pp. 305–311.
- [74] Z. Oo et al. "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification". In: *Proc. of INTERSPEECH. 2016*, pp. 2204–2208.
- [75] S. E. Eskimez et al. "Front-End Speech Enhancement for Commercial Speaker Verification Systems". In: *Speech Communication* 99 (2018), pp. 101–113.
- [76] T. H. Falk and W. Chan. "Modulation Spectral Features for Robust Far-Field Speaker Identification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 90–100.
- [77] S. O. Sadjadi and J. H. L. Hansen. "Hilbert Envelope Based Features for Robust Speaker Identification Under Reverberant Mismatched Conditions". In: *Proc. of ICASSP. 2011*, pp. 5448–5451.
- [78] Q. Jin et al. "Speaker Identification with Distant Microphone Speech". In: *Proc. of ICASSP. 2010*, pp. 4518–4521.
- [79] S. O. Sadjadi and J. H. L. Hansen. "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.5 (2014), pp. 937–945.



References VIII

- [80] Chanwoo Kim and Richard M Stern. "Power-Normalized Cepstral Coefcients (PNCC) for Robust Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24.7 (2016), pp. 1315–1329.
- [81] D. Cai et al. "The DKU-SMIIP System for the Speaker Recognition Task of the VOiCES from a Distance Challenge". In: *Proc. of INTERSPEECH*. 2019.
- [82] T. Yamada, L. Wang, and A. Kai. "Improvement of Distant Talking Speaker Identification Using Bottleneck Features of DNN". In: *Proc. of INTERSPEECH*. 2013, pp. 3661–2664.
- [83] I. Peer, B. Rafaely, and Y. Zigel. "Reverberation Matching for Speaker Recognition". In: *Proc. of ICASSP*. 2008, pp. 4829–4832.
- [84] A. R Avila et al. "Improving the Performance of Far-Field Speaker Verification Using Multi-Condition Training: The Case of GMM-UBM and i-Vector Systems". In: *Proc. of INTERSPEECH*. 2014, pp. 1096–1100.
- [85] A. Bruttì and A. Abad. "Multi-Channel i-vector Combination for Robust Speaker Verification in Multi-Room Domestic Environments". In: *Proc. of Odyssey*. 2016, pp. 252–258.
- [86] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson. "Multicondition Training of Gaussian Plda Models in i-vector Space for Noise and Reverberation Robust Speaker Recognition". In: *Proc. of ICASSP*. 2012, pp. 4257–4260.
- [87] Q. Jin, T. Schultz, and A. Waibel. "Far-Field Speaker Recognition". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.7 (2007), pp. 2023–2032.
- [88] M. Ji et al. "Text-Independent Speaker Identification using Soft Channel Selection in Home Robot Environments". In: *IEEE Transactions on Consumer Electronics* 54.1 (2008), pp. 140–144.
- [89] M. K. Nandwana et al. "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings". In: *Proc. of INTERSPEECH*. 2018, pp. 1106–1110.
- [90] D. Cai, X. Qin, and M. Li. "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment". In: *Proc. of INTERSPEECH*. 2019.



References IX

- [91] F. Zhao, H. Li, and X. Zhang. "A Robust Text-independent Speaker Verification Method Based on Speech Separation and Deep Speaker". In: *Proc. of ICASSP*. 2019, pp. 6101–6105.
- [92] X. Qin, D Cai, and M. Li. "Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation". In: *Proc. of INTERSPEECH*. 2019.
- [93] S. E. Tranter and D. A. Reynolds. "An Overview of Automatic Speaker Diarization Systems". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5 (2006), pp. 1557–1565. ISSN: 1558-7916.
- [94] X. Anguera et al. "Speaker Diarization: A Review of Recent Research". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 356–370. ISSN: 1558-7916.
- [95] Elias Rentzeperis et al. "The 2006 Athens Information Technology Speech activity detection and speaker diarization systems". In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer. 2006, pp. 385–395.
- [96] Andrey Temko, Dusan Macho, and Climent Nadeu. "Enhanced SVM training for robust speech activity detection". In: *Proc. of ICASSP*. Vol. 4. IEEE. 2007.
- [97] Florian Eyben et al. "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies". In: *Proc. of ICASSP*. IEEE. 2013, pp. 483–487.
- [98] Douglas A Reynolds and P Torres-Carrasquillo. *The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations*. Tech. rep. 2004.
- [99] Scott Chen, Ponani Gopalakrishnan, et al. "Speaker, environment and channel change detection and clustering via the bayesian information criterions". In: *Proc. DARPA broadcast news transcription and understanding workshop*. Vol. 8. Virginia, USA. 1998, pp. 127–132.
- [100] Herbert Gish, M-H Siu, and Robin Rohlicek. "Segregation of speakers for speech recognition and speaker identification". In: *Proc. of ICASSP*. IEEE. 1991, pp. 873–876.



References X

- [101] Ruiqing Yin, Herv Bredin, and Claude Barras. "Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks". In: *Proc. Interspeech*. 2017, pp. 3827–3831.
- [102] Chuck Wooters et al. "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system". In: *RT-04F Workshop*. Vol. 23. 2004, p. 23.
- [103] Sylvain Meignier et al. "Step-by-step and integrated approaches in broadcast news speaker diarization". In: *Computer Speech & Language* 20.2-3 (2006), pp. 303–330.
- [104] Elie El Khoury, Christine Senac, and Régine André-Obrecht. "Speaker diarization: towards a more robust and portable system". In: *Proc. of ICASSP*. Vol. 4. IEEE. 2007, pp. IV-489.
- [105] G. Sell and D. Garcia-Romero. "Speaker diarization with plda i-vector scoring and unsupervised calibration". In: *IEEE Spoken Language Technology Workshop*. 2014, pp. 413–417.
- [106] Q. Wang et al. "Speaker Diarization with LSTM". In: *Proc. of ICASSP*. 2018, pp. 5239–5243.
- [107] Gregory Sell et al. "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge". In: *Proc. Interspeech*. 2018, pp. 2808–2812.
- [108] Qingjian Lin et al. "LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization". In: *Proc. Interspeech*. 2019.
- [109] K. Chidananda Gowda and G. Krishna. "Agglomerative Clustering Using the Concept of Mutual Nearest Neighbourhood". In: *Pattern Recognition* 10 (1978), pp. 105–112.
- [110] Ulrike von Luxburg. "A Tutorial on Spectral Clustering". In: *Statistics and Computing* 17 (2007), pp. 395–416.
- [111] Aonan Zhang et al. "Fully Supervised Speaker Diarization". In: *Proc. of ICASSP*. 2019.
- [112] Xianhong Chen et al. "VB-HMM Speaker Diarization with Enhanced and Refined Segment Representation." In: *Proc. of Odyssey*. 2018, pp. 134–139.



References XI

- [113] Mireia Diez, Lukás Burget, and Pavel Matejka. "Speaker Diarization based on Bayesian HMM with Eigenvoice Priors." In: *Proc. of Odyssey*. 2018.
- [114] Ondrej Novotný et al. "BUT system for DIHARD speech diarization challenge 2018". In: (2018).
- [115] Björn W Schuller et al. "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity". In: *Proc. of INTERSPEECH*. 2019.
- [116] Björn Schuller et al. "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats". In: *Proc. of Interspeech*. 2018, pp. 122–126. URL: <http://dx.doi.org/10.21437/Interspeech.2018-51>.
- [117] Björn Schuller et al. "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring". In: *Proc. of Interspeech*. 2017, pp. 3442–3446. URL: <http://dx.doi.org/10.21437/Interspeech.2017-43>.
- [118] Gbor Gosztolya, Tams Grsz, and Lszl Tth. "General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats". In: *Proc. of Interspeech*. 2018, pp. 531–535.
- [119] Gbor Gosztolya et al. "DNN-Based Feature Extraction and Classifier Combination for Child-Directed Speech, Cold and Snoring Identification". In: *Proc. of Interspeech*. 2017, pp. 3522–3526.
- [120] Gbor Gosztolya et al. "Estimating the Sincerity of Apologies in Speech by DNN Rank Learning and Prosodic Analysis". In: *Proc. of Interspeech*. 2016, pp. 2026–2030.
- [121] Yishan Jiao et al. "Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features". In: *Proc. of Interspeech*. 2016, pp. 2388–2392.
- [122] Johannes Wagner et al. "Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?" In: *Proc. of Interspeech*. 2018, pp. 147–151.
- [123] Q.F. Tan, P.G. Georgiou, S.S. Narayanan, et al. "Enhanced sparse imputation techniques for a robust speech recognition front-end". In: *IEEE Transactions on Audio Speech and Language Processing* 19.8 (2011), p. 2418



References XII

- [124] Ahmed Imtiaz Humayun et al. "An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification". In: *Proc. of Interspeech*. 2018, pp. 127–131.
- [125] Heysem Kaya et al. "LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition". In: *Proc. of Interspeech*. 2018, pp. 521–525.
- [126] Bo-Hao Su et al. "Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features". In: *Proc. of Interspeech*. 2018, pp. 536–540.
- [127] Cristina Gorrostieta et al. "Attention-based Sequence Classification for Affect Detection". In: *Proc. of Interspeech*. 2018, pp. 506–510.
- [128] Dengke Tang, Junlin Zeng, and Ming Li. "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals". In: *Proc. of INTERSPEECH*. 2018, pp. 162–166.
- [129] Ming Li et al. "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder". In: *Computer Speech & Language* 56 (2019), pp. 80–94.
- [130] Danqing Luo, Yuejian Zou, and Dongyan Huang. "Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition". In: *Proc. of Interspeech*. 2018, pp. 152–156.
- [131] Shahin Amiriparian et al. "Snore Sound Classification Using Image-Based Deep Spectrum Features". In: *Proc. of INTERSPEECH*. 2017, pp. 3512–3516.
- [132] Björn Schuller et al. "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism". In: *Proc. of INTERSPEECH*. 2013, pp. 148–152.
- [133] Danwei Cai et al. "End-to-End Deep Learning Framework for Speech Paralinguistics Detection Based on Perception Aware Spectrum". In: *Proc. of INTERSPEECH*. 2017, pp. 3452–3456.
- [134] Mario Lasseck. "Audio-based bird species identification with deep convolutional neural networks". In: *Working Notes of CLEF 2018* (2018).



References XIII

- [135] Massimiliano Todisco, Hctor Delgado, and Nicholas Evans. "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification". In: *Computer Speech & Language* 45 (Feb. 2017).
- [136] Weiqing Wang Haiwei Wu and Ming Li. "The DKU-LENOVO Systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge". In: *Proc. of INTERSPEECH*. 2019.
- [137] Chunlei Zhang, Chengzhu Yu, and John HL Hansen. "An investigation of deep-learning frameworks for speaker verification antispoofting". In: *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2017), pp. 684–694.
- [138] Galina Lavrentyeva et al. "Audio Replay Attack Detection with Deep Learning Frameworks." In: *Proc. of Interspeech*. 2017, pp. 82–86.
- [139] Francis Tom, Mohit Jain, and Prasenjit Dey. "End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention." In: *Proc. of Interspeech*. 2018, pp. 681–685.