End-to-end deep neural network based speaker recognition

Ming Li

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

Oct 26th 2019

Table of Contents

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Table of Contents

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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

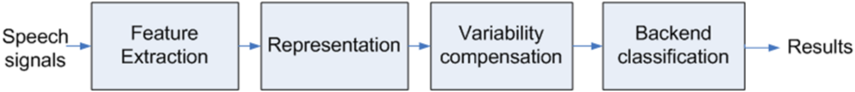


Figure: General framework

Table of Contents

Feature Extraction

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

[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).

Feature Extraction

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

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

[2] 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.

[3] 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.

[4] 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.

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][1], Speech attributes [10][2], Acoustic-to-articulatory inversion [11][3], subglottal[12][4], etc.

---

[1] 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.

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

[3] 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.

[4] 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.

Feature Extraction

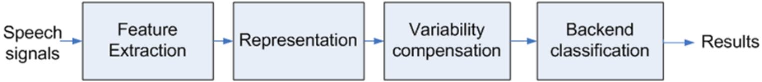
- MFCC, PLP, SDC [1], PNCC[2], GFCC[3] , CQCC [4],etc.
- Bottleneck [5][6], Phoneme Posterior Probability [7][8], etc.
- LLD/OpenSmile [9], Speech attributes [10], Acoustic-to-articulatory inversion [11], subglottal[12], etc.
- IMFCC[13][1], Modified Group Delay[14][2], etc.

---

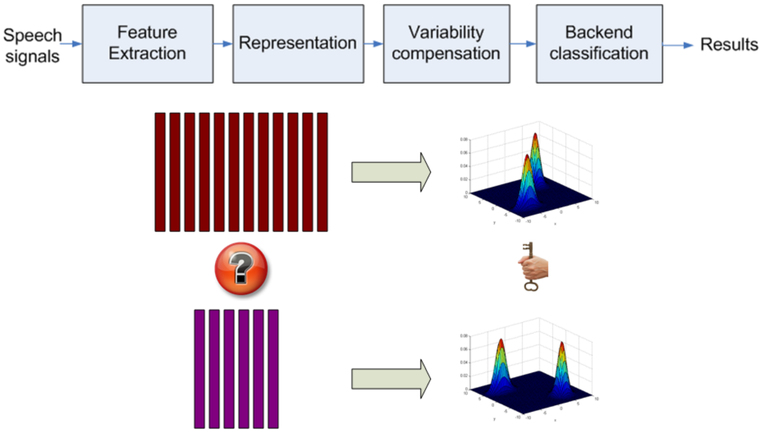
[1] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. "A comparison of features for synthetic speech detection". In: (2015).

[2] 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.

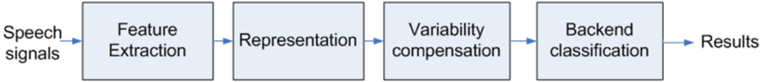
Representation

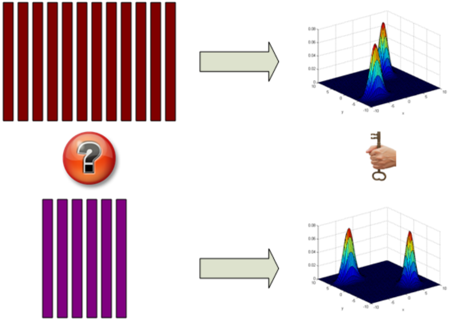


Representation



Representation





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

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Generative model, adaptation, supervectors

- Gaussian Mixture Model (GMM) [15][3] serves as the generative model

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

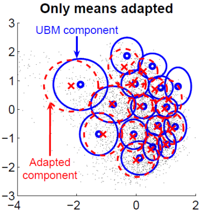
Generative model, adaptation, supervectors

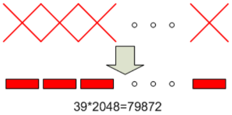
- Gaussian Mixture Model (GMM) [15][3] serves as the generative model
- model adaptation from universal background model (UBM)

[3] D.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][3]





[3] 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.

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][3]
    - Maximum Likelihood Linear Regression (MLLR) adaptation the linear regression matrix becomes GMM MLLR supervector [17][4]

[3] 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.

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

- 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

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda) \qquad (1)$$

$$\tilde{\mathbf{F}}_\mathbf{c} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \boldsymbol{\mu}_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \qquad (2)$$

Cumulated by L frames

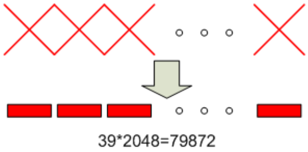
Generative model, adaptation, supervectors

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

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda) \qquad (1)$$

Cumulated by L frames

$$\tilde{\mathbf{F}}_\mathbf{c} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \boldsymbol{\mu}_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \qquad (2)$$



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

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda) \qquad (1)$$

Cumulated by L frames

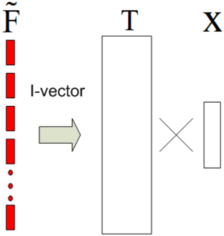
$$\tilde{\mathbf{F}}_{\mathbf{c}} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \boldsymbol{\mu}_{\mathbf{c}})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \qquad (2)$$

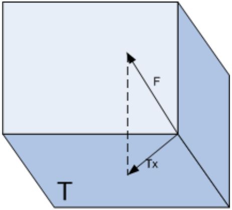
- Mapping from a set of feature vectors to a fixed dimensional supervector

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Factor analysis based dimension reduction

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



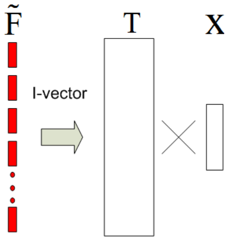


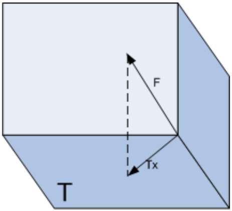
Factor analysis based dimension reduction

- Factor analysis on the concatenated centered normalized 1^{st} order statistics vector or GMM mean supervector
    - total variability i-vector [18][5]

$$\tilde{\mathbf{F}} = \mathbf{T}\mathbf{x} \qquad (3)$$

$\mathbf{T}$: factor loading matrix, $\mathbf{x}$: i-vector





[5] 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.

[6] 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.

Factor analysis based dimension reduction

- Factor analysis on the concatenated centered normalized 1^{st} order statistics vector or GMM mean supervector
  - total variability i-vector [18][5]

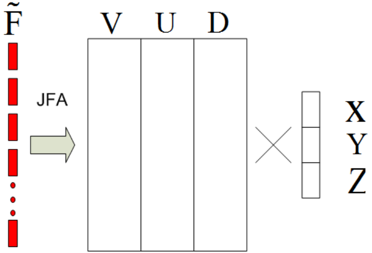
$$\tilde{\mathbf{F}} = \mathbf{T}\mathbf{x} \qquad (3)$$

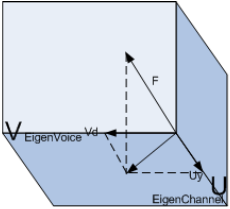
T: factor loading matrix, **x**: i-vector

  - joint factor analysis (JFA) [19][6]

$$\tilde{\mathbf{F}} = \mathbf{V}\mathbf{x} + \mathbf{U}\mathbf{y} + \mathbf{D}\mathbf{z} \qquad (4)$$

V: Eigenvoices, **U**: Eigenchannels, **x**: speaker factor, **y**: channel factor, **D**: diagonal covariance matrix





[5] 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.

[6] 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.

Variability Compensation

LDA, WCCN [20][7], NAP[16][8], NDA [21][9], LSDA [22][10], LFDA [23][11], etc.

[7] 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.

[8] 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.

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

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

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

Backend Classification

SVM [16][12], PLDA [24][13][25][14], NN [26][15][27][16], Joint Bayesian [28][17] , Cosine Similarity, etc.

[12]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.

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

[14]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.

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

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

[17]Yiyan 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

System Pipeline

Variable-length input → Local pattern extractor → Encoding layer → Feed-forward network → Loss function

- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29][18], Mel-filterbank energies [30][19], MFCC [31][20]), or even the raw waveform [32][21].

[18] 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/.

[19] 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].

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

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

System Pipeline

Variable-length input → Local pattern extractor → Encoding layer → Feed-forward network → Loss function

- Speech signal is naturally with arbitrary duration. The input can be a hand-crafted short-term spectral feature (STFT spectrogram [29][18], Mel-filterbank energies [30][19], MFCC [31][20]), or even the raw waveform [32][21].
- The local pattern extractor plays a role as an automatic representation learning module. (TDNN/CNN/LSTM/CNN-LSTM/CNN-BLSTM).

---

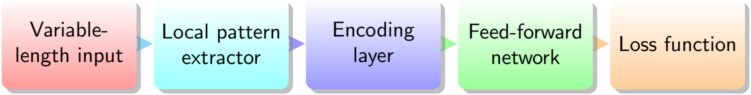
[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: http://www.robots.ox.ac.uk/~vgg/data/voxceleb/.

[19] 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].

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

[21] 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][18], Mel-filterbank energies [30][19], MFCC [31][20]), or even the raw waveform [32][21].
- 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)

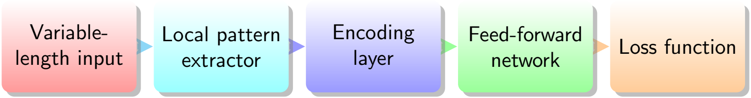
[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset". In: *arXiv preprint arXiv:1706.08612* (2017). URL: http://www.robots.ox.ac.uk/~vgg/data/voxceleb/.

[19] 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].

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

[21] 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][18], Mel-filterbank energies [30][19], MFCC [31][20]), or even the raw waveform [32][21].
- 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)

[18] 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/.

[19] 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].

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

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

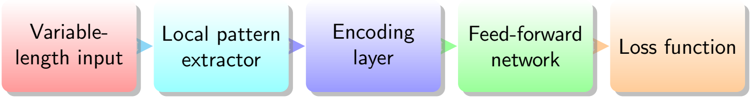
System Pipeline

Variable-length input → Local pattern extractor → Encoding layer → Feed-forward network → Loss function

Task

- Language identification or paralinguistic speech attributes detection(Closed-set)
  Network outopput $\rightarrow$ Utterance-level posteriors

System Pipeline



Task

- Language identification or paralinguistic speech attributes detection(Closed-set)
  Network outoput $\rightarrow$ Utterance-level posteriors
- Speaker Verification (Open-set)
  Utterance-level speaker embedding + Cosine / PLDA $\rightarrow$ 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 $\rightarrow$ variable-length feature sequence $\rightarrow$ 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 $\rightarrow$ variable-length feature sequence $\rightarrow$ 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 $\rightarrow$ variable-length feature sequence $\rightarrow$ 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 $\rightarrow$ 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 $\rightarrow$ variable-length feature sequence $\rightarrow$ 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 $\rightarrow$ 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 → 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
- 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][22][34][23][35][24]

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

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

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

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

D-vector [33][22][34][23][35][24]

- 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.

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

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

[24] Lantian 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][22][34][23][35][24]

- 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.

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

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

[24] Lantian 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][22][34][23][35][24]

- 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.

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

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

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

Data preparation

X-vector [36][25]

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

[25] David 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][25]

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

[25] David 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][25]

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

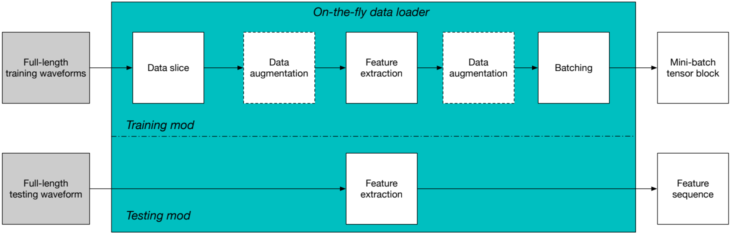
[25] David Snyder et al. "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5329–5333.

Data preparation

X-vector [36][25]

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

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

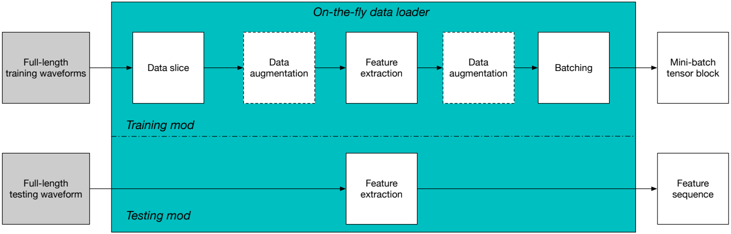


On-the-fly data loader [37][26]

- 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.

[26] 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).

Data preparation

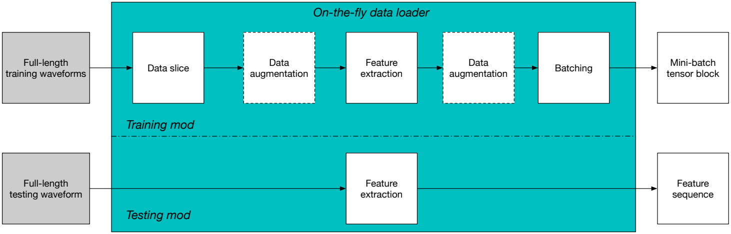


On-the-fly data loader [37][26]

- 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.

[26] 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).

Data preparation

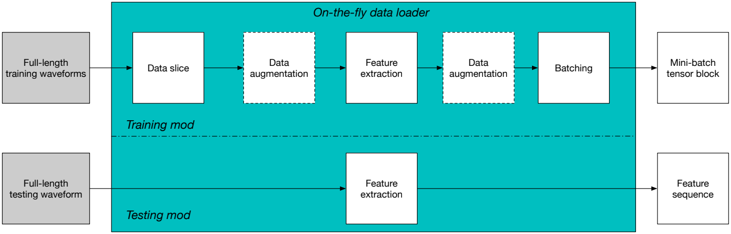


On-the-fly data loader [37][26]

- 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.

[26] 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).

Data preparation



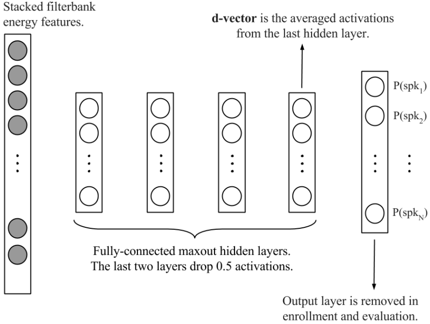
On-the-fly data loader [37][26]

- 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.

[26] 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).

Network Structure

Feed-forward DNN(FF-DNN)

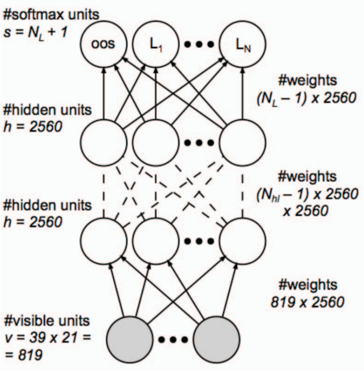


- D-vector for SV [33][27]

[27] Ehsan 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][28]

[28] I. 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 ($\leq$ 3s test segment)

Network Structure

Feed-forward DNN(FF-DNN)

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

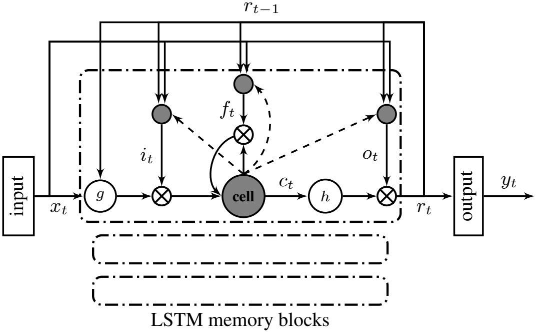
Network Structure

Feed-forward DNN(FF-DNN)

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

Network Structure

RNN/LSTM

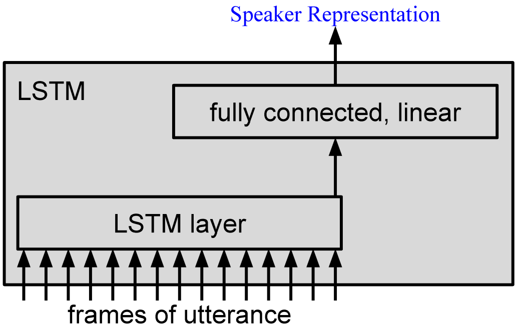


- LSTM for LID [39][29]

[29] J. 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

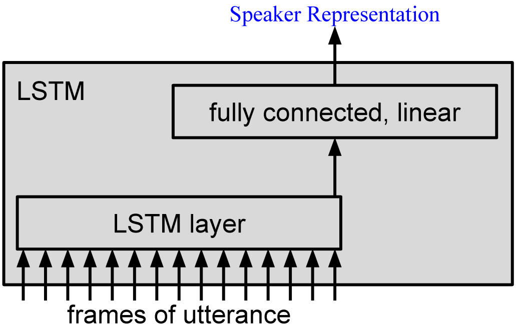


- LSTM for SV [40][29]

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

Network Structure

RNN/LSTM

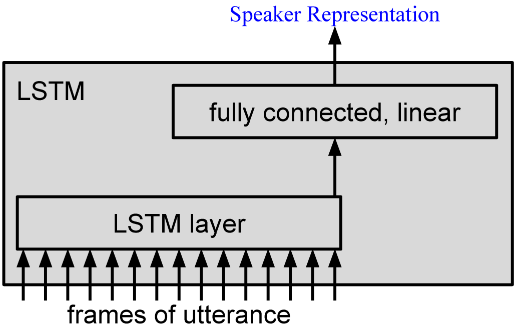


- LSTM for SV [40][29]
- Adopt the last several output units of LSTM

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

Network Structure

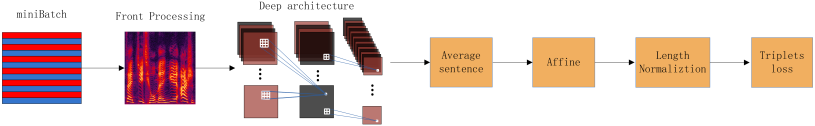
RNN/LSTM



- LSTM for SV [40][29]
- Adopt the last several output units of LSTM
- Short duration ($\leq$ 3s test segment)

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

CNN

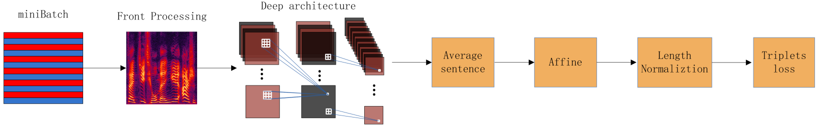


- CNN: Deep Speaker [30][30]

[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].

Network Structure

CNN

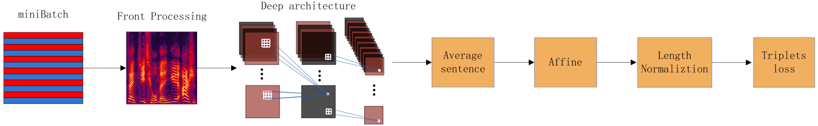


- CNN: Deep Speaker [30][30]
- Anti-spoofing [41][31]

[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] 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.

CNN



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

---

[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] 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.

[32] 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.

[33] 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.

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	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

- x-vector [36][34]

[34] 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][35]

[35] 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][35]
  - LSTM/GRU + Attention [44][36]

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

[36] 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][35]
  - LSTM/GRU + Attention [44][36]
  - Bi-LSTM + Attention [45][37]

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

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

[37] 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][38]

[38] 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][38]
  - Statistics pooling (mean + std) [36][39]

---

[38] 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].

[39] David 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]

[38] 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].

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

[40] 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]
  - Self-attentive pooling (mean) [47][41]

[38] 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].

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

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

[41] 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]
  - Self-attentive pooling (mean) [47][41]
  - Attentive statistics pooling (mean + std) [48][42] [49][43]

---

[38] 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].

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

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

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

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

[43] Yingke 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]
  - Self-attentive pooling (mean) [47][41]
  - Attentive statistics pooling (mean + std) [48][42] [49][43]
  - Multi-head attentive pooling [50][44]

[38] 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].

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

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

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

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

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

[44] 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]
  - Self-attentive pooling (mean) [47][41]
  - Attentive statistics pooling (mean + std) [48][42] [49][43]
  - Multi-head attentive pooling [50][44]
  - Learnable dictionary encoding [51][45]

---

[38] 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].

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

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

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

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

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

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

[45] 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][38]
  - Statistics pooling (mean + std) [36][39]
  - Bilinear pooling [46][40]
  - Self-attentive pooling (mean) [47][41]
  - Attentive statistics pooling (mean + std) [48][42] [49][43]
  - Multi-head attentive pooling [50][44]
  - Learnable dictionary encoding [51][45]
  - NetFV/NetVLAD/Ghost VLAD [52][46] [53][47]

---

[38] 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].

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

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

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

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

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

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

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

[46] J. Chen et al. "End-to-end Language Identification using NetFV and NetVLAD". In: *Proc. ISCSLP*. 2018.

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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][48] [55][49]

[48] 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.

[49] 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][48] [55][49]
- End-to-End loss [40][50] [56][51]

[48] 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.

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

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

[51] 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][48] [55][49]
- End-to-End loss [40][50] [56][51]
- Center loss [42][52]

[48] 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.

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

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

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

[52] 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][48] [55][49]
- End-to-End loss [40][50] [56][51]
- Center loss [42][52]
- Angular softmax loss [57][53] [42][58][54]

---

[48] 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.

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

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

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

[52] 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.

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

[54] 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][48] [55][49]
- End-to-End loss [40][50] [56][51]
- Center loss [42][52]
- Angular softmax loss [57][53] [42][58][54]
- Additive margin loss [33][55]

---

[48] 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.

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

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

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

[52] 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.

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

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

[55] 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][56]

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

[57] 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).

[58] 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][56]
- Speed perturbation [59][57]

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

[57] 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).

[58] 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][56]
- Speed perturbation [59][57]
- Generative adversarial network (GAN) [60][58]

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

[57] 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).

[58] 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 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][59]

[59]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][59]
- Maximum mean disprepancy [62][60]

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

[60] Wei-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][59]
- Maximum mean disprepancy [62][60]
- Autoencoder based domain adaptation (AEDA) [63][61]

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

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

[61] 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][59]
- Maximum mean disprepancy [62][60]
- Autoencoder based domain adaptation (AEDA) [63][61]
- Domain adversarial training (DAT) [64][62]

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

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

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

[62] 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][59]
- Maximum mean disprepancy [62][60]
- Autoencoder based domain adaptation (AEDA) [63][61]
- Domain adversarial training (DAT) [64][62]
- CORAL [65][63]

---

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

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

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

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

[63] 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 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][59]
- Maximum mean disprepancy [62][60]
- Autoencoder based domain adaptation (AEDA) [63][61]
- Domain adversarial training (DAT) [64][62]
- CORAL [65][63]
- CORAL+ [66][64]

---

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

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

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

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

[63] 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.

[64] 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.

Domain Adaptation

End-to-End Domain adaptation

- End-to-end adversarial training [67][65]

[65] 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.

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

End-to-End Domain adaptation

- End-to-end adversarial training [67][65]
- Generative adversarial network (GAN) [68][66]

[65] 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.

[66] 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.

[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][65]
- Generative adversarial network (GAN) [68][66]
- Multi-task adversarial network [69][67]

[65] 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.

[66] 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.

[67] 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

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][68], weighted prediction error (WPE) [71][69]

[68] 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.

[69] L. 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][68], weighted prediction error (WPE) [71][69]
  - DNN based denoising methods for single-channel speech enhancement [72][70] [73][71] [74][72],[75][73]

[68] 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.

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

[70] 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.

[71] 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.

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

[73] S. 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][68], weighted prediction error (WPE) [71][69]
  - DNN based denoising methods for single-channel speech enhancement [72][70] [73][71] [74][72],[75][73]
  - Beamforming for multi-channel speech enhancement [71][74]

---

[68] 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.

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

[70] 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.

[71] 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.

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

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

[74] 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][75],[77][76]

[75] 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.

[76] 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][75],[77][76]
  - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78][77]

[75] 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.

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

[77] Q. 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][75],[77][76]
  - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78][77]
  - Blind spectral weighting (BSW) based features [79][78]

---

[75] 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.

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

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

[78] 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.

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Previous Methods on Robust Modeling

- Feature level
  - Sub-band Hilbert envelopes based features [76][75],[77][76]
  - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78][77]
  - Blind spectral weighting (BSW) based features [79][78]
  - Power-normalized cepstral coefficients (PNCC) [80][79][81][80]

---

[75] 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.

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

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

[78] 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.

[79] 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.

[80] 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.

Previous Methods on Robust Modeling

- Feature level
  - Sub-band Hilbert envelopes based features [76][75],[77][76]
  - Warped minimum variance distortionless response (MVDR) cepstral coefficients [78][77]
  - Blind spectral weighting (BSW) based features [79][78]
  - Power-normalized cepstral coefficients (PNCC) [80][79][81][80]
  - DNN bottleneck features [82][81], etc.

---

[75] 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.

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

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

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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][82],[84][83]
  - Multi-channel i-vector combination [85][84]
  - Multi-condition training of PLDA models [86][85]
- Score level
  - Score normalization [83][86]
  - Multi-channel score fusion [87][87],[88][88]

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

[83] 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.

[84] A. Brutti 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.

[85] 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.

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

[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.

Robust Modeling of End-to-End Methods

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

[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.

Robust Modeling of End-to-End Methods

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

[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.

Robust Modeling of End-to-End Methods

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

[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.

Robust Modeling of End-to-End Methods

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

[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.

Robust Modeling of End-to-End Methods

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

[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.

Robust Modeling of End-to-End Methods

- DNN speaker embedding under far-field and noisy environment [89][89]
  - 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][90]
  - The performance gain achieves by data augmentation in the end-to-end method is lager than in the i-vector framework

[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.

Robust Modeling of End-to-End Methods

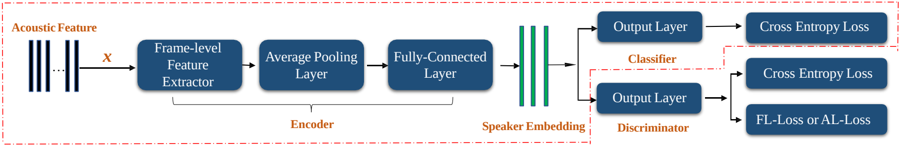
- DNN speaker embedding under far-field and noisy environment [89][89]
  - X-vector + PLDA
  - Retransmitted speech in reverberant environments
  - Speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems
- Two interesting findings of end-to-end methods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

[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.

Robust Modeling of End-to-End Methods

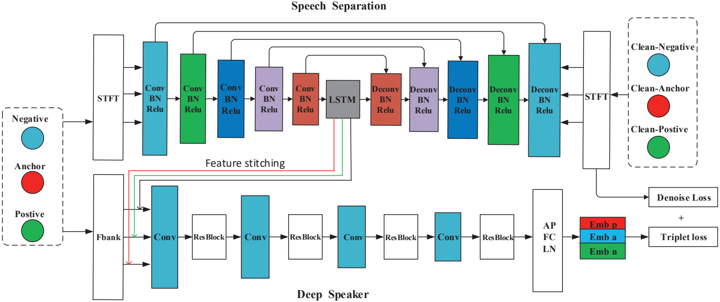
- Multi-task adversarial network for noise-robust speaker embedding [69][91]
    - Encoding network for speaker embedding
    - Speaker classifier
    - Noise discriminator
    - Adversarial training by using fix-label loss or anti-label loss (take wrong label with cross entropy) of the noise discriminator
    - Outperform the other methods without adversarial training in noisy environments



[91] J. 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

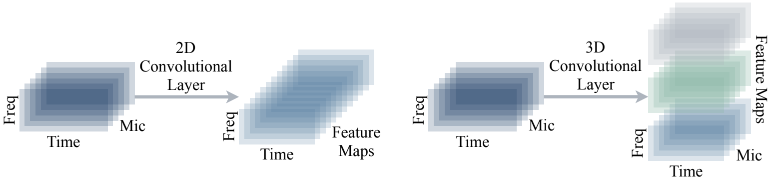
- Joint training of denoising and speaker embedding network[91][92]
  - Denoising network
    - extract the target speech from noisy speech
    - extract bottleneck features
  - Speaker embedding network
    - Concatenate bottleneck features with fbank as inputs



[92] 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.

Robust Modeling of End-to-End Methods

- Multi-channel training framework for speaker recognition under reverberant and noisy environment [90][93]
  - 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



[93] 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.

Robust Modeling of End-to-End Methods

- Far-field text-dependent speaker verification [92][94]
  - 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 dataset [95]
    - Open source wake-up words speech database

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

[95] https://www.aishelltech.com/aishell_2019B_eval

HI-MIA Database Paper

Xiaoyi Qin, Hui Bu, Ming Li, "HI-MIA : A Far-field Text-dependent Speaker Verification Database and the Baselines", submitted to ICASSP 2020.

Ongoing Research Topics in DKU-SMIIP Lab

Robust Speaker Verification (far-field, multi-channel, noisy, etc.)
Robust Speaker Diarization (single channel, multi-channel, far-field, noisy, etc.)
Speech Separation/Enhancement (speech/music, supervised voice-filter, etc.)
Multi-Speaker TTS, Voice Conversion
Replay Detection Anti-Spoofing Database
Paralinguistic Speech Attribute Recognition
Acoustic Scene Analysis/Environmental Sound Classification

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

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Thank you very much!

Email: ming.li369@duke.edu
Website: `https://scholars.duke.edu/person/MingLi`
Slide Download Link: `https://sites.duke.edu/dkusmiip`

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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.

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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. 4879–4883.

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.

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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.

[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.

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

References VIII

[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. Brutti 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.

[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.

References IX

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