Combining Deep Embeddings of Acoustic and Articulatory Features for Speaker Identification

Qian-Bei Hong[1], Chung-Hsien Wu[1,2], Hsin-Min Wang[1], and Chien-Lin Huang[3]

[1]Graduate Program of Multimedia Systems and Intelligent Computing,
National Cheng Kung University and Academia Sinica, Tainan, Taiwan
[2]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
[3]PingAn AI Lab, Palo Alto, CA 94306, USA

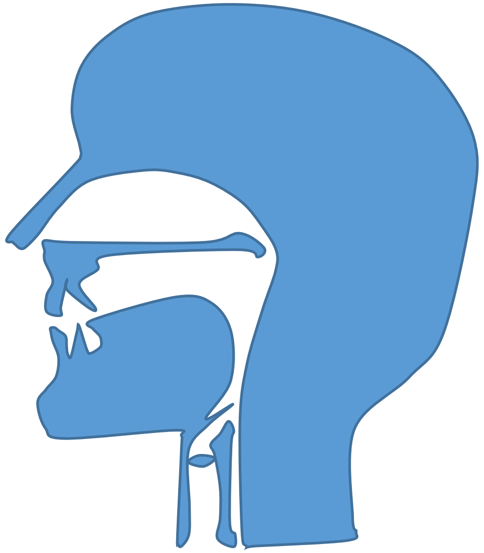
Outline

- Introduction
- The Proposed Speaker Identification System
  - Speaker Embedding Extraction
  - Articulatory Feature Extraction
  - Enrolled Speaker Classifier
- Experimental Results
- Conclusions

Introduction

- Articulatory feature (AF) is an important representation of phonological properties during speech production.
  - AFs have been successfully used as features in speech recognition.
    - Concatenating the acoustic features and AF information to improve speech recognition performance.
  - It is rarely investigated in speaker recognition.

Category	Attribute
Manner	Approximant, Fricative, Nasal, Stop, Vocalic
Place	Anterior, Back, Continuant, Coronal, Dental, High, Labial, Low, Mid, Retroflex, Round, Tense, Velar, Voiced
Silence	Silence

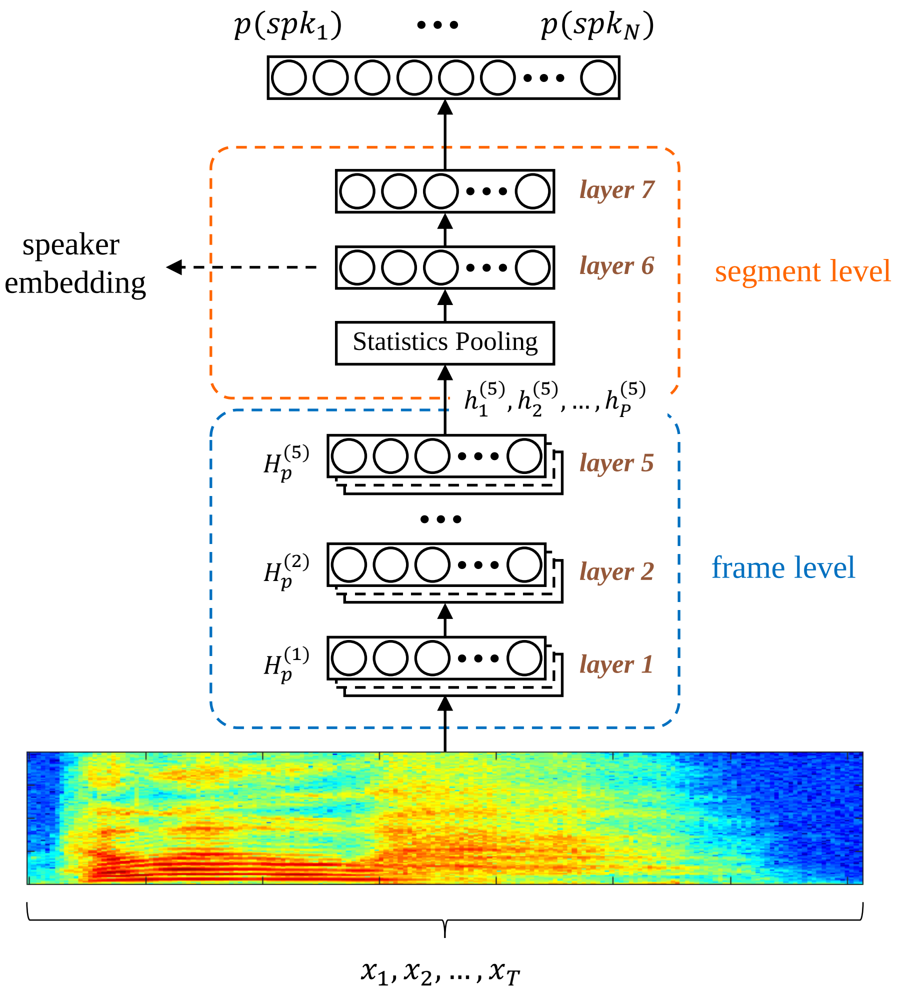


Introduction

- Traditional text-independent speaker recognition systems
  - Gaussian mixture model-universal background model (GMM-UBM) and i-vector
- In recent years, deep neural network (DNN)-based models for speaker recognition have become more and more popular.
  - d-vector
  - x-vector

Introduction

- The x-vector embedding with PLDA classifier is the SOTA system for speaker verification.

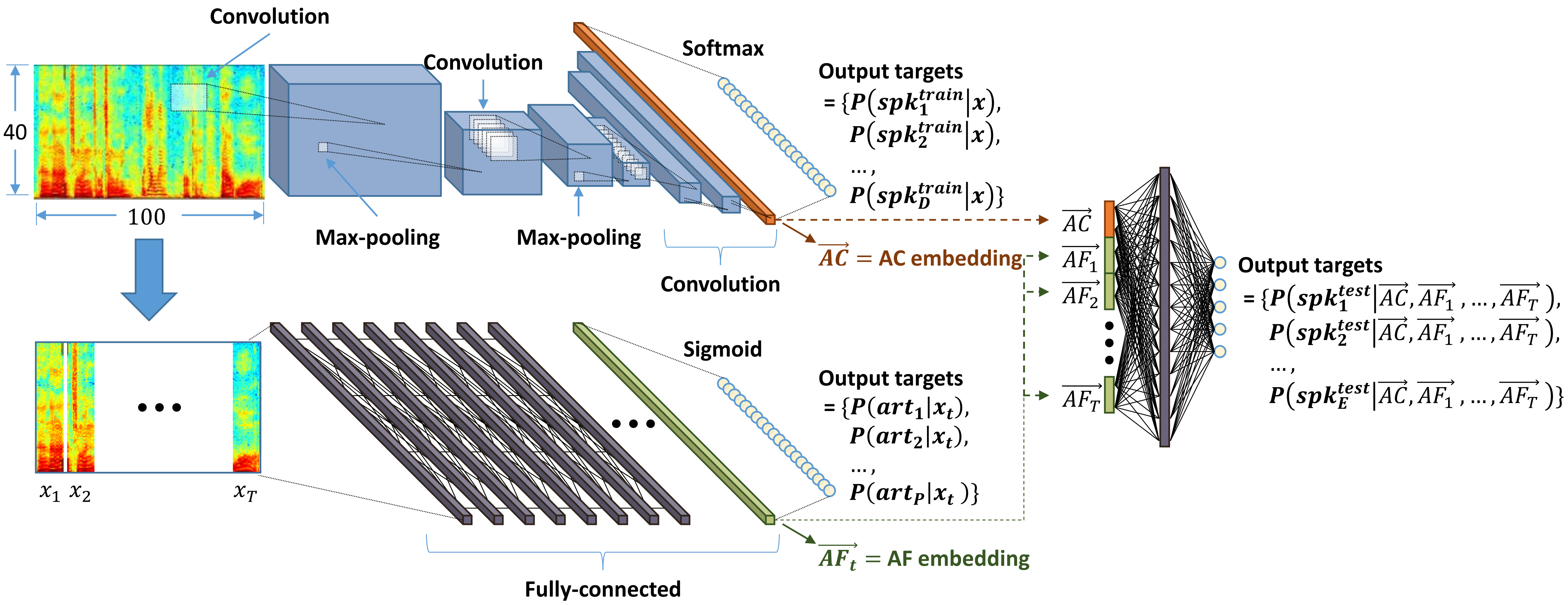


Goal

- This paper integrated speaker embedding with AF embedding for speaker identification.
  - Adding the AFs is helpful for presenting the personal pronunciation attributes to improve speaker identification performance.

- Using CNN-based model to extract the feature embedding
  - To achieve the better performance than traditional feature extraction models.

The Proposed Speaker Identification System



Speaker Embedding Extraction

- A CNN model is trained to produce speaker embedding.

Layer	Layer Type	Kernel Size	Depth	Stride	Data Size
-	input	-	-	-	[100,40,1]
1	convolution	[1,5] [9,1]	16 32	[1,1] [2,1]	[46,36,32]
2	max-pooling	[2,2]	-	[2,2]	[23,18,32]
3	convolution	[1,5] [8,1]	32 64	[1,1] [1,1]	[16,14,64]
4	max-pooling	[2,2]	-	[2,2]	[8,7,64]
5	convolution	[1,3] [6,1]	128 128	[1,1] [1,1]	[3,5,128]
6	convolution	[1,3] [3,1]	256 512	[1,1] [1,1]	[1,3,512]
7	convolution	[1,3]	1024	[1,1]	[1,1,1024]
8	dense & softmax	-	-	-	num. spk

Input data
- A 100×40 spectrogram

Features
- 40-dimentional MFCC

Speaker embedding
- Layer 7 output

Articulatory Feature Extraction

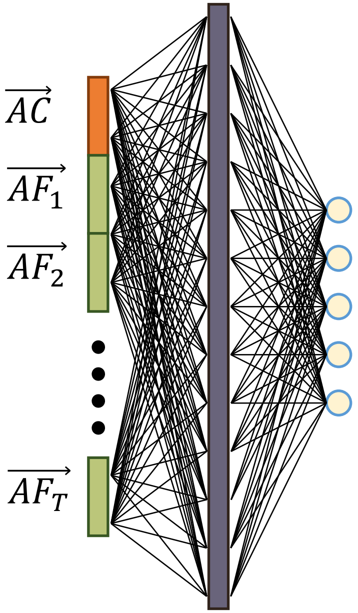
- AFs can be distinguished into different pronunciation places and manners by speaker voices.

Phoneme → Articulatory

- Training step:
  1. The **Kaldi ASR toolkit** is used to align the phone positions of the training speech signals in GMM-HMM-based acoustic model training procedure.
  2. According to the alignment information, every segment of training speech signals can exactly be labeled with the attributes which the phone corresponds to.
  3. A multilayer perceptron (MLP)-based model is trained for AF recognition.
  4. After the model training is completed, the AF embedding is extract from the output of last hidden layer.

Enrolled Speaker Classifier

- The speaker classifier is trained to identify who the speaker is in the recording.



Output targets

$$= \{P(spk_1^{test}|\overrightarrow{AC}, \overrightarrow{AF_1}, \dots, \overrightarrow{AF_T}), \\ P(spk_2^{test}|\overrightarrow{AC}, \overrightarrow{AF_1}, \dots, \overrightarrow{AF_T}), \\ \dots, \\ P(spk_E^{test}|\overrightarrow{AC}, \overrightarrow{AF_1}, \dots, \overrightarrow{AF_T})\}$$

Hidden layer produce the 1,024-dimensional feature for speaker discrimination.

Datasets

- Training data of speaker embedding model
  - **King-ASR-044:** 500 randomly selected speakers; the training data contained 15,396 recordings.
- Training data of AF embedding model
  - **King-ASR corpora:** Approximately 130 hours recordings
    - 2,082 randomly selected speakers from King-ASR-044
    - 1,026 randomly selected speakers from King-ASR-360
- Testing data
  - **LibriSpeech corpus:** 460 hours “clean” speech collected from 1,172 speakers
  - **Speakers in the Wild (SITW) corpus:** the core-core subset (a total of 1,201 recordings were collected from 180 speakers)

Experimental Results

- Comparison on different number of enrolled speakers

King-ASR: 100 randomly selected speakers
- **Enrollment:** 25 recordings for each speaker
- **Evaluation:** 5 recordings for each speaker

LibriSpeech: 1,172 speakers of train-clean subset
- **Enrollment:** 10 recordings for each speaker
- **Evaluation:** 2 recordings for each speaker

Systems	King-ASR 100 speakers	LibriSpeech 1,172 speakers
d-vector (cosine distance)	4.10	13.50
x-vector (cosine distance)	2.92	11.18
d-vector (PLDA)	1.54	10.61
x-vector (PLDA)	1.02	8.25
AC embedding classifier	2.33	7.95
AC & AF embedding classifier	2.41	7.80

Experimental Results

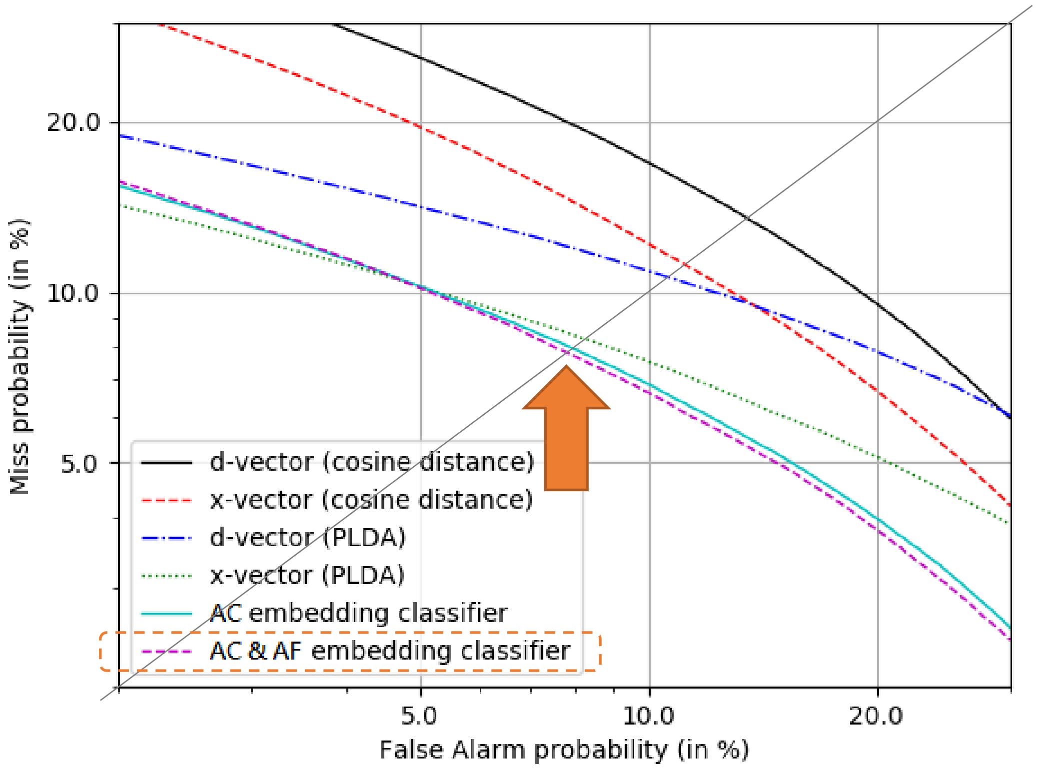
- The effect of signal mismatch

The **SITW corpus** provides samples of same speaker across varying environmental conditions.

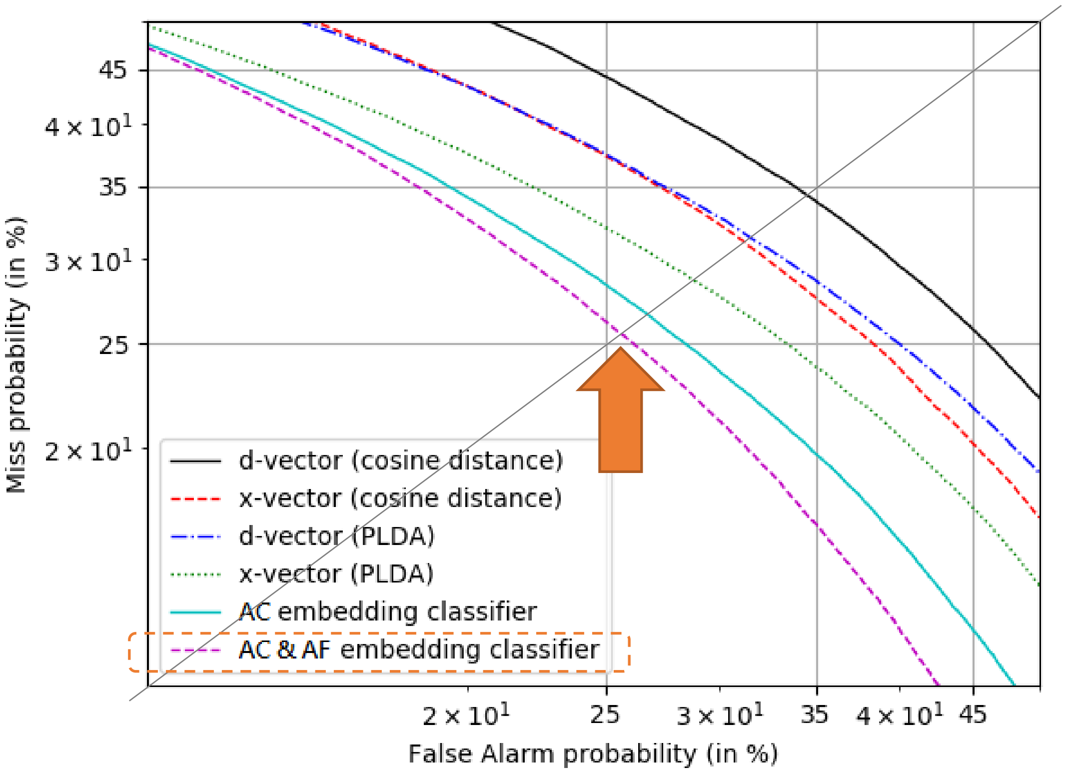
- **Evaluation:** 1 recordings for each speaker
- **Enrollment:** remaining recordings

Systems	SITW core-core 180 speakers
d-vector (cosine distance)	34.42
x-vector (cosine distance)	31.14
d-vector (PLDA)	31.43
x-vector (PLDA)	28.70
AC embedding classifier	26.67
AC & AF embedding classifier	25.19

Experimental Results



DET curves comparison on 1,172 enrolled speakers



DET curves comparison on SITW core-core subset without considering the types of environment

Experimental Results

- In this study, compared to the SOTA x-vector system
  - When the number of enrolled speakers is increased from 100 to 1,172.
    - **X-vector:** the performance is decreased by 87.6% in EER.
      (1.02% → 8.25%)
    - **Our system:** the performance is decreased by 69.1% in EER.
      (2.41% → 7.8%)
  - When recordings are collected from different conditions, it will cause the signal mismatch.
    - **X-vector:** achieved the EER of 28.7%.
    - **Our system:** achieved the EER of 25.19%.

Conclusions

- In this paper, we integrated speaker embedding with AF embedding for speaker identification.
  - We found that training a backend classifier from large number of data for speaker recognition can achieved a better performance than PLDA scoring.
  - Combining the articulatory features to consider the speech attributes, it can help us to build a more robust speaker recognition model.
  - Even though the all systems achieved poor performances in the case of signal mismatch, our proposed system is still superior to the baseline systems.

Future Work

- In the future, we will try to train the speaker recognition with noisy data augmentation
  - To deal with the signal mismatch problem.
- We will investigate the potential of attention mechanisms
  - To further consider the different status in speech
    - Speaking style
    - Emotion

Thank you for your attention