On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction

Shuai Wang[1,2], Johan Rohdin[2], Lukáš Burget[2], Oldřich Plchot[2], Yanmin Qian[1], Kai Yu[1], Jan Černocký[2]

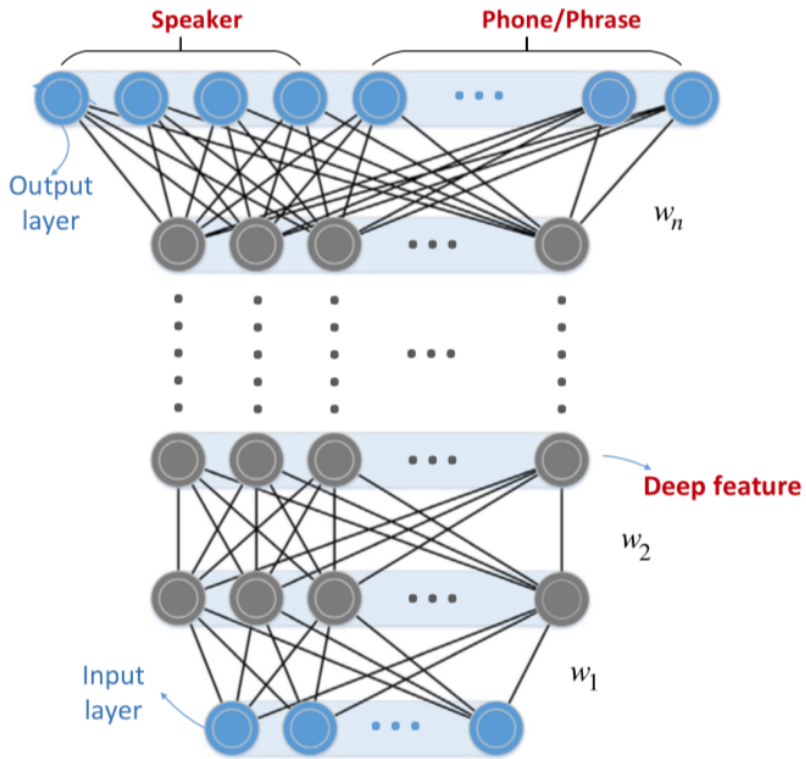
[1]Speech Lab, Shanghai Jiao Tong University, China
[2]Speech@FIT, Brno University of Technology, Czechia

September 2019

- Background
  - The development of ASR techniques has greatly inspired the SID community
  - Researchers investigated to incorporate phonetic information for speaker embedding learning
- Basic assumption for this paper:
  Good text-independent speaker embeddings are not expected to be affected by spoken content, it might be helpful to explicitly **suppress** the phonetic variability **in the final embeddings**

Related work

multi-task learning in the d-vector framework[1]



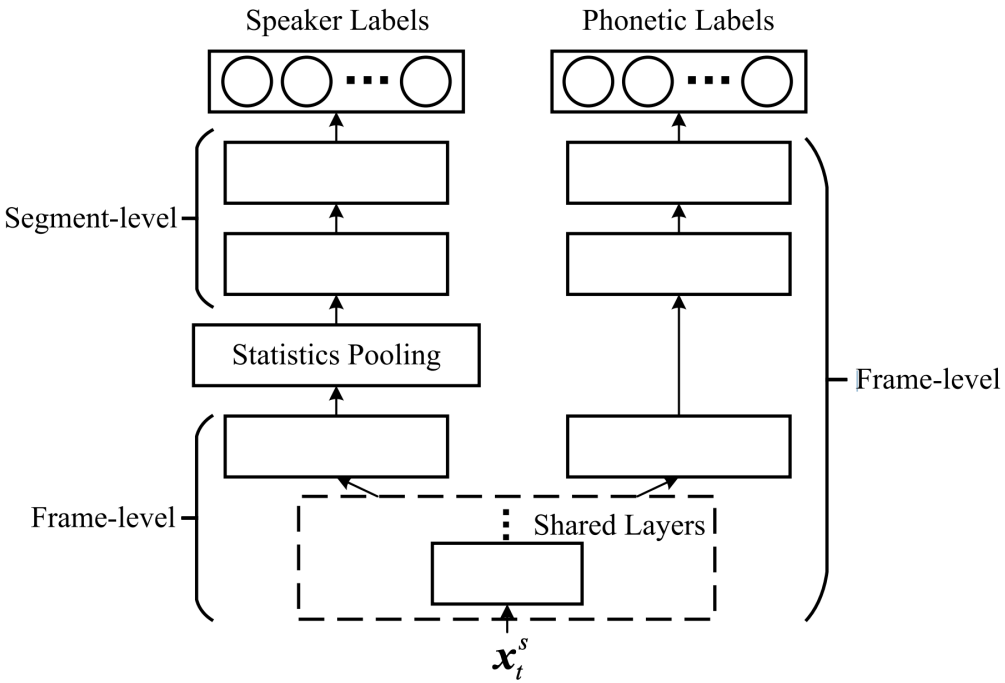
- **Text-dependent** task
- Multi-task at the **frame-level**
- Performance improved

Explicitly modeling phonetic information helps the text-dependent speaker verification task, which is intuitive

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

Related work

multi-task learning in the x-vector framework[2]



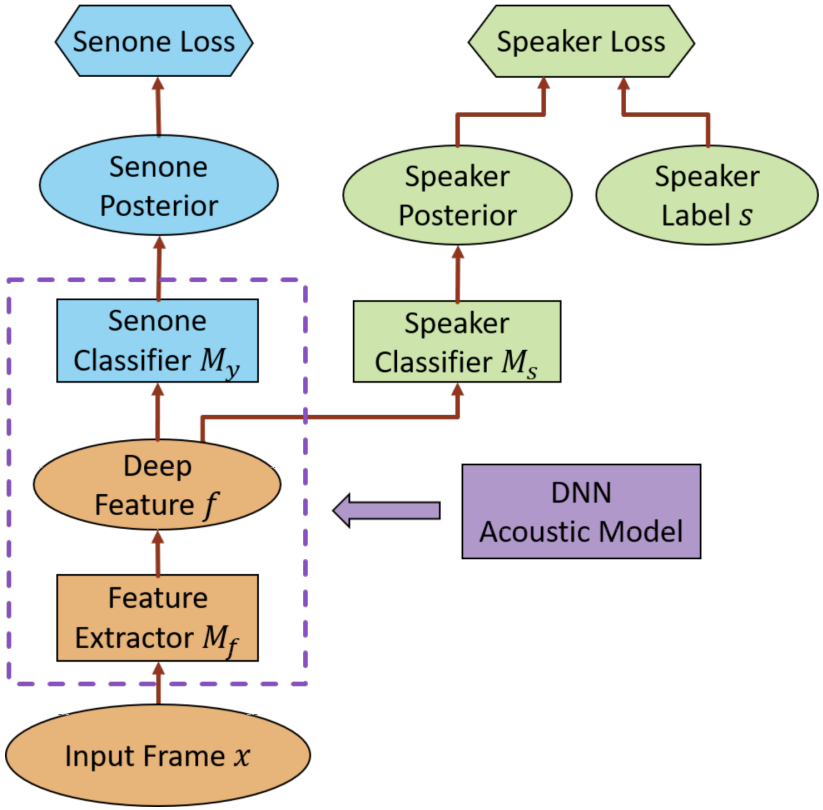
- **Text-independent** task
- Multi-task at the **frame-level**
- Performance improved!

Why explicitly learning phonetic information helps the text-independent speaker verification task? Is it counter-intuitive?

[2] Liu, Yi, et al. "Speaker Embedding Extraction with Phonetic Information." Proc. Interspeech 2018 (2018): 2247-2251.

Related work

Speaker invariant training for ASR [3]



- Acoustic modelling
- Adversarial training suppressing the speaker effect
- Performance improved

[3] Meng, Zhong, et al. "Speaker-invariant training via adversarial learning." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Question & Motivation

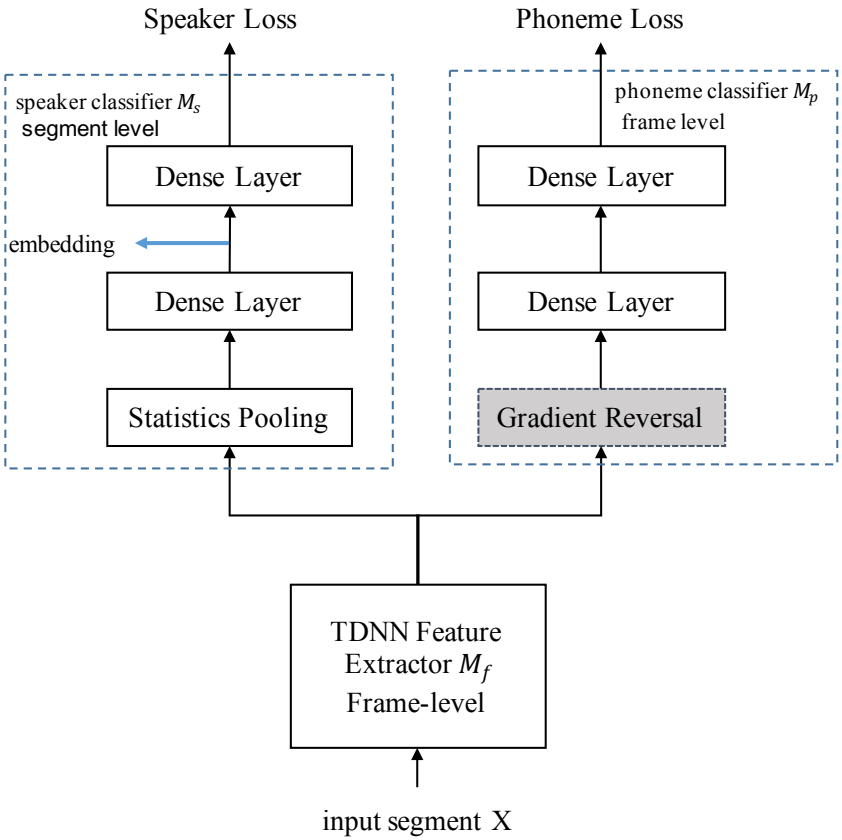
What will happen if we train a speaker classifier, while making it unable to discriminate different phoneme classes?

Adversarial training

Explicitly suppress the phonetic information in speaker embedding

- Consider speaker classification as the primary task and phoneme classification as the secondary task
- Multi-task aims to minimize the classification loss of both tasks
- In adversarial training, the goal is to minimize the speaker classification loss and **mini-maximize** the phoneme classification loss

Frame-level multi-task/adversarial training



$$\mathcal{L}_s = \mathrm{CE}(M_s(M_f(\mathbf{X})), \mathbf{y}^s)$$

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}(M_p(M_f(\mathbf{x}_i)), \mathbf{y}_i^p)$$

$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_p$$

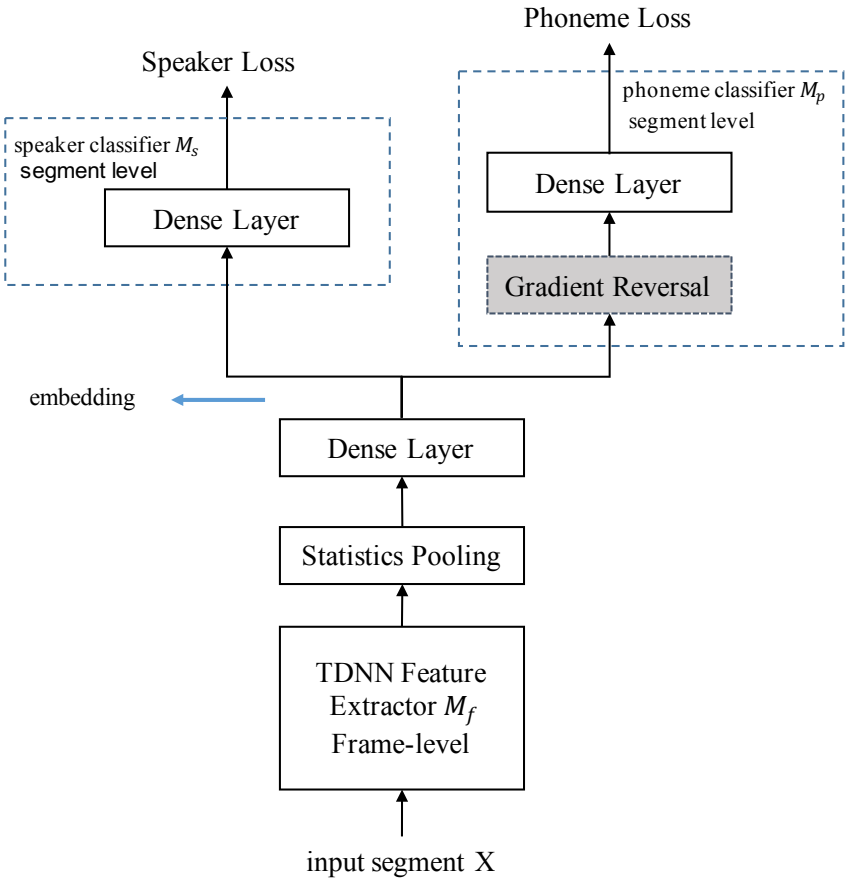
Frame-level multi-task/adversarial training

Problems and possible reasons

However, we observe a big performance degradation for the frame-level adversarial training

- Granularity might be the problem, it's harder to remove fine-grained information from coarse-grained information
- What will happen if we do both tasks at the same granularity?

Segment-level multi-task/adversarial training



$$\mathcal{L}_s = \mathrm{CE}(M_s(M_f(\mathbf{X})), \mathbf{y}^s)$$
$$\mathcal{L}_p = \mathrm{CE}(M_p(M_f(\mathbf{x}_i)), \mathbf{y}^p)$$
$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_p$$

Question: How to represent $\mathbf{y}^p$

Segment-level multi-task/adversarial training

"Normalized histogram": represent phoneme information at the segment level

For a given segment $\mathbf{x}$ with N frames, the corresponding segment-level phoneme label $\mathbf{y}^p$ is represented as

$$\mathbf{y}^p = \{y_1, y_2, \ldots, y_C\}$$
$$y_c = \frac{N_c}{N}$$

where C is the size of the chosen phoneme set. N_c denotes the number of occurrences of the c-th phoneme in $\mathbf{x}$

Experiments

Setups

Dataset

Training data:
Voxceleb1 Dev + Voxceleb2 Dev

Evaluation data:
Voxceleb1 Eval

Speaker Embedding Extractor
All speaker embedding systems are based on the TDNN x-vector

Phoneme recognizer

- The phoneme labels are generated from a phoneme recognizer
- 166 classes:
  position-dependent phonemes
  + SIL and NOISE nodes
- Training follows official Kaldi Tedlium speech recognition recipe

Experiments

Results: Frame-level multi-task/adversarial training

Table: Systems combining frame-level phonetic information, FRM-MT and FRM-ADV denote two systems trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively

System	EER(%)	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.192
FRM-MT	3.38	0.180
FRM-ADV	5.24	0.269

Experiments

Results: Segment-level multi-task/adversarial training

Table: Systems combining segment-level phonetic information, SEG-MT and SEG-ADV denote two systems trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively

System	EER(%)	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.192
SEG-MT	3.71	0.175
SEG-ADV	3.35	0.159

Experiments

Results: Combining frame-level multitask and segment-level adversarial learning

Table: Systems combining frame-level multitask and segment-level adversarial learning

System	EER(%)	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.192
FRM-MT	3.38	0.180
SEG-ADV	3.35	0.159
FRM-MT + SEG-ADV	3.17	0.163

Conclusions

- Main Contribution
  - The architecture of training multi-task/adversarial systems at the segment level
  - Experiments to examine the impact of phonetic information on text-independent speaker embedding learning
- Our experiments show that for text-independent speaker embedding learning, it's beneficial to
  - enhance the fine-grained phonetic information at the frame-level part
  - suppress phonetic information at the segment-level part