Temporal Dynamic Convolutional Neural Network for Text-Independent Speaker Verification and Phonemic Analysis

ICASSP 2022 Oral Presentation
@May 12th, 2022

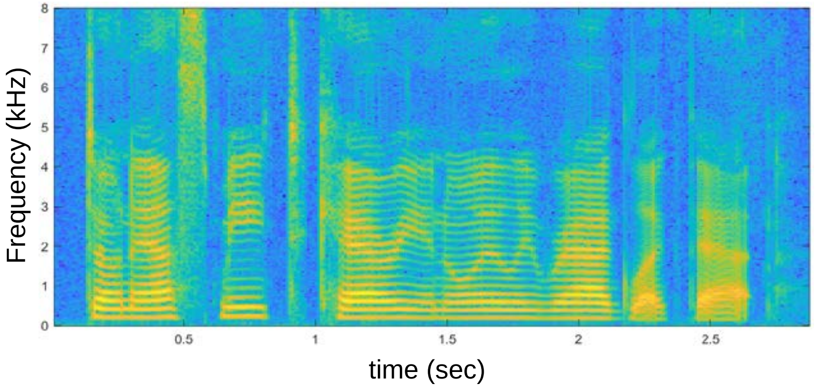
Seong-Hu Kim , Ph.D. Candidate
seonghu.kim@kaist.ac.kr

Human-Machine iNteraction LAB
Dept. Mechanical Eng., KAIST

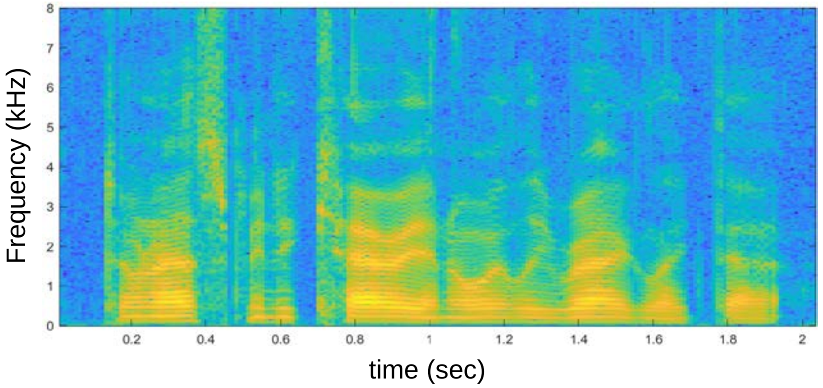
Background

- **Text-dependent speaker verification task**

SA2 utterance of Speaker FSJK1
(Don't ask me to carry an oily rag like that)

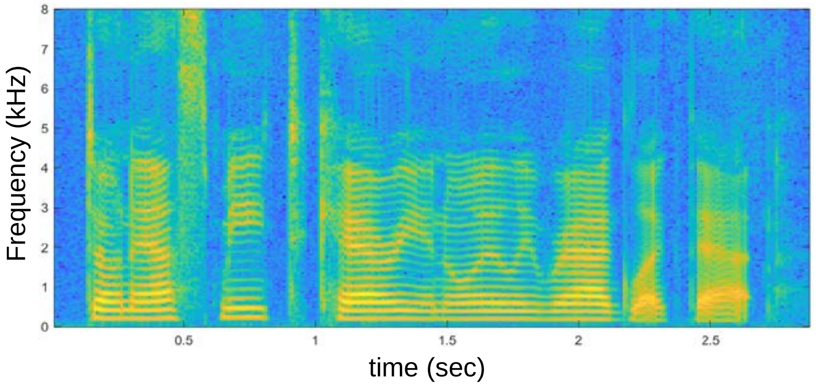


SA2 utterance of Speaker MPGH0
(Don't ask me to carry an oily rag like that)

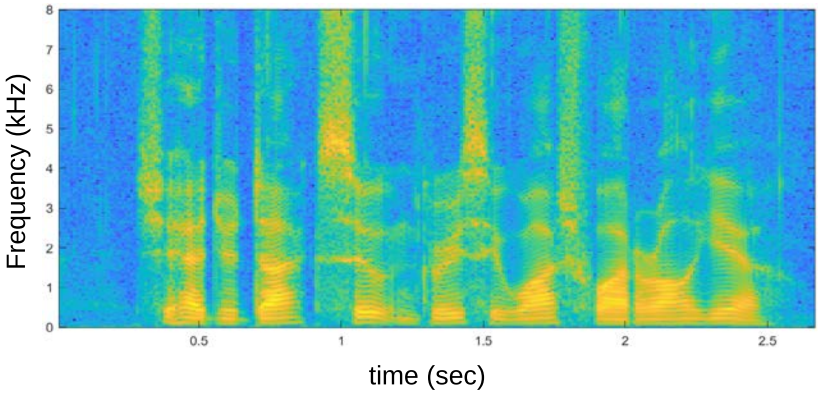


- **Text-independent speaker verification task**

SA2 utterance of Speaker FSJK1
(Don't ask me to carry an oily rag like that)



SA1 utterance of Speaker MPGH0
(She had your dark suit in greasy wash water all year)



Introduction

- In text-independent speaker verification task, speech has **phoneme-varying characteristics along the time axis depending on random text,** but conventional static neural models do not reflect it.

- We propose **temporal dynamic convolutional neural network (TDY-CNN).**

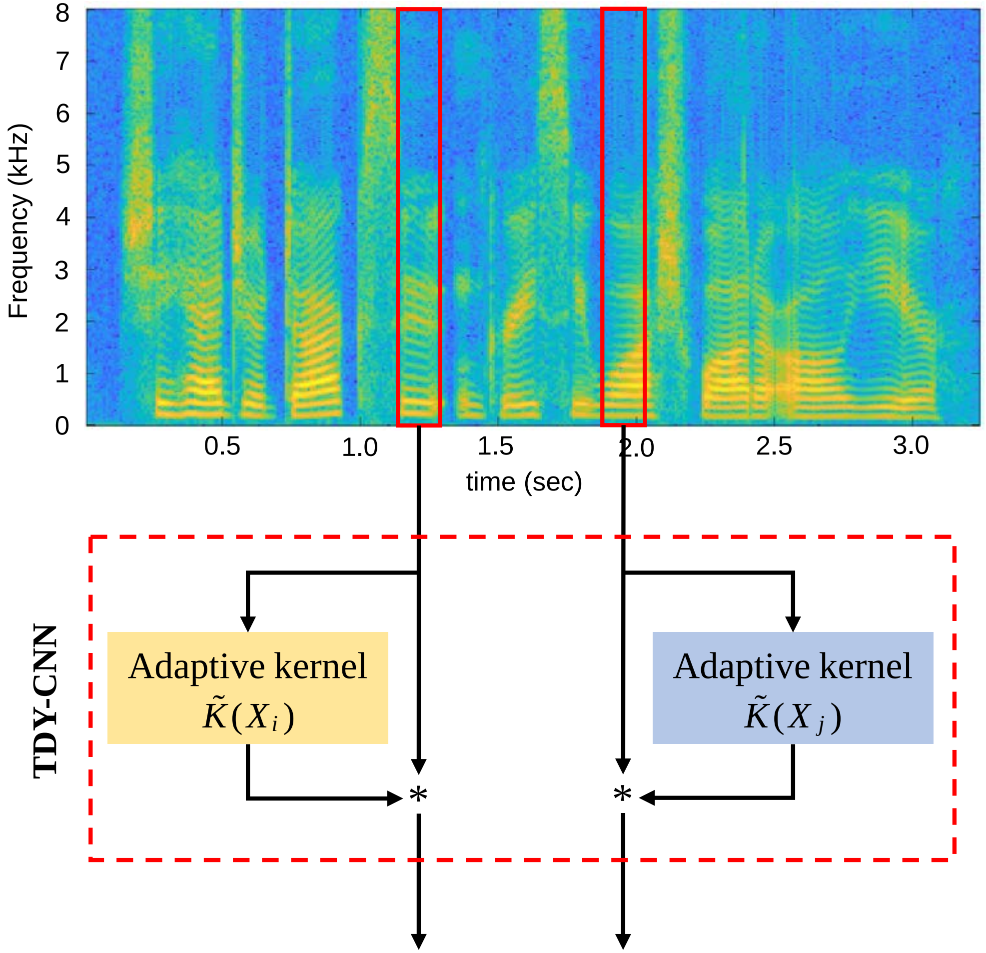
- Contributions of this paper as follows :

1. We propose **CNN kernels adaptive to each time bin** in order to effectively capture the time-varying information in utterances.

2. This is the **first work to perform phonemic analysis on temporal dynamic model** for text-independent speaker verification.

3. We verified that **adaptive kernels change with the acoustic characteristics of phonemes** and **extract speaker information regardless of phonemes** while static kernels do not.

'She had your dark suit in greasy wash water all year' utterance in TIMIT[1]



[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

Temporal Dynamic CNN Module

- TDY-CNN was proposed by **referring to dynamic convolutional neural network (DY-CNN)[2]**, which generate the adaptive kernel with **weighted sum of basis kernels**.

$$y_k(f,t) = W_k * x(f,t) + b_k$$

$$y(f,t) = \sigma\left(\sum_{k=1}^{K} \pi_k(t) \cdot y_k(f,t)\right)$$

- x, y : input and output
- f, t : frequency and time features
- W_k, b_k: k-th basis kernel and bias
- K : total number of basis kerels
- $\pi_k(t)$: temporal attention weights

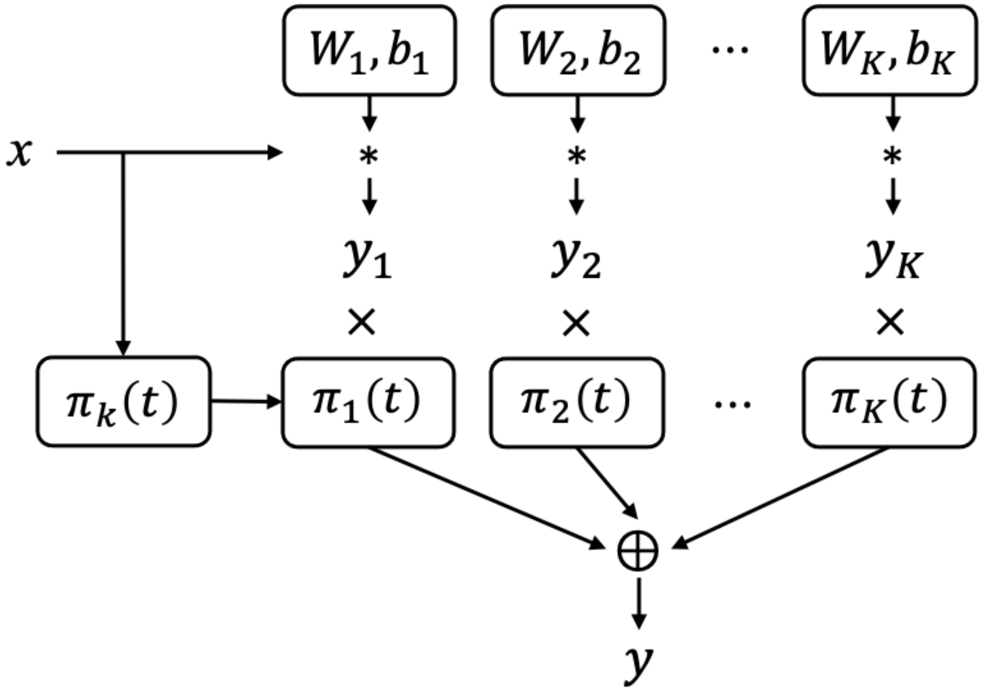


Figure 1. Structure of temporal dynamic convolutional neural network (TDY-CNN)

[2] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030-11039.

Model Structure and Experiment Details

- We used 64-dimensional log Mel spectrogram as the model input.

- TDY-CNN is applied to VGG-M[3] and ResNet-34[4] with a quarter and half channel:
  VGG-M / ResNet-34 (×0.25) / DY-ResNet-34 (×0.5)

- The models were trained on Voxceleb2 development set and tested on Voxceleb1 test set.

- The models are trained using **a loss function combining the Angular Prototypical loss with the vanilla softmax loss**, which shows better performance [5].

  - optimizer : Adam
  - initial learning rate : 10^{-3}
  - learning rate decaying : 0.75 every 15 epochs
  - batch size : 800
  - data augmentation : NO

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, Conference Proceedings, pp. 770–778.
[5] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," arXiv preprint arXiv:2003.11982, 2020

Results and Analysis

[The number of basis convolution kernels]

- TDY-ResNet-34 (×0.25) with ***K* = 6** showed best performance for text-independent speaker verification.
- The error rate was increased when ***K* = 8** because of the **difficulty of optimization** for larger models and **overfitting**.
- We set ***K* = 6** and continued analysis.

Table 1. Text-independent speaker verification performance of temporal dynamic models

TDY-ResNet-34(×0.25)	EER (%)	MinDCF
$K = 2$	1.99	0.140
$K = 4$	1.62	0.128
$K = 6$	**1.58**	**0.116**
$K = 8$	1.69	0.133

KAIST

Results and Analysis

[Comparison of static model and utterance/frame-level dynamic model]

- Each model information :
  - **Static model** : ResNet-34 (×0.25)
  - **Utterance-level dynamic model** : DY-ResNet-34 (×0.25)
  - **Frame-level dynamic model** : TDY-ResNet-34 (×0.25)
- **TDY-CNN**, which considers frame-level speaker information, **showed the better verification performance than DY-CNN**, which considers only utterance-level speaker information.
- Thus, **TDY-CNN is suitable for text-independent speaker verification**.

Table 2. Text-independent speaker verification performances of models using dynamic convolution with frame-level and utterance level.

Network	EER (%)	MinDCF
ResNet-34(×0.25)	2.43	0.184
DY-ResNet-34(×0.25)	2.07	0.157
TDY-ResNet-34(×0.25)	**1.58**	**0.116**

KAIST

Results and Analysis

[Text-independent speaker verification results]

- **TDY-ResNet-34 (×0.5)** showed the best performance with 1.48% of EER.
- **All models to which TDY-CNN was applied show better performance** than the models using static CNN.
- The proposed models show good verification performance but lag slightly behind the state-of-the-art performance.

Table 3. Text-independent speaker verification performances of the networks without data augmentation.

Network	#Parm	EER (%)	MinDCF
VGG-M	4.16M	3.77	0.287
TDY-VGG-M	71.2M	3.04	0.237
ResNet-34(×0.25)	1.86M	2.43	0.184
TDY-ResNet-34(×0.25)	13.3M	1.58	0.116
ResNet-34(×0.5)	6.37M	1.79	0.134
TDY-ResNet-34(×0.5)	51.9M	**1.48**	**0.118**
ResNet-50 [19]	67.0M	3.95	0.429
Thin ResNet-34 [20]	12.4M	2.87	0.310
H/ASP [22]	8.00M	1.29	0.091
ECAPA-TDNN [25]	14.7M	**1.18**	**0.088**

Results and Analysis

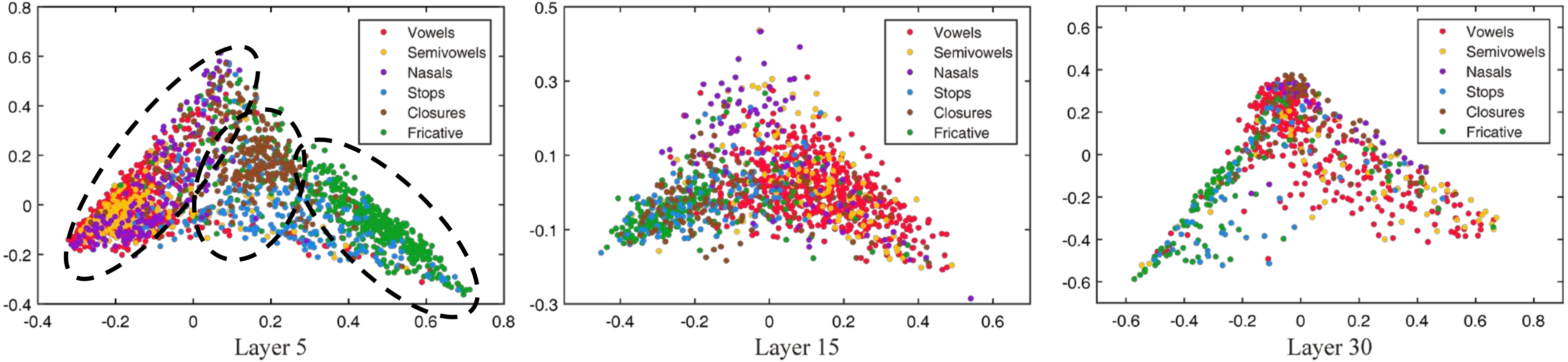
[Analysis of adaptive kernels in relation to phonemes]

- We verified how the kernels adapt to phonemes by **comparing attention weights of basis kernels depending on phonemes**.
- A total of 52 phonemes are classified into 6 categories :
  vowels / semivowels and glides / nasals / fricatives and affricates / stops / closures.
- Correlation between attention weights and phonemes was analyzed in several layers (5, 15, 30 layer) using TIMIT dataset which provides phoneme labels.
- The attention weights were extracted from trained **TDY-ResNet-34(×0.25) with K = 6** and displayed using principal component analysis (PCA).

Results and Analysis

[Analysis of adaptive kernels in relation to phonemes]

➢ **Speaker FDAS1** (the largest variance of attention weights)



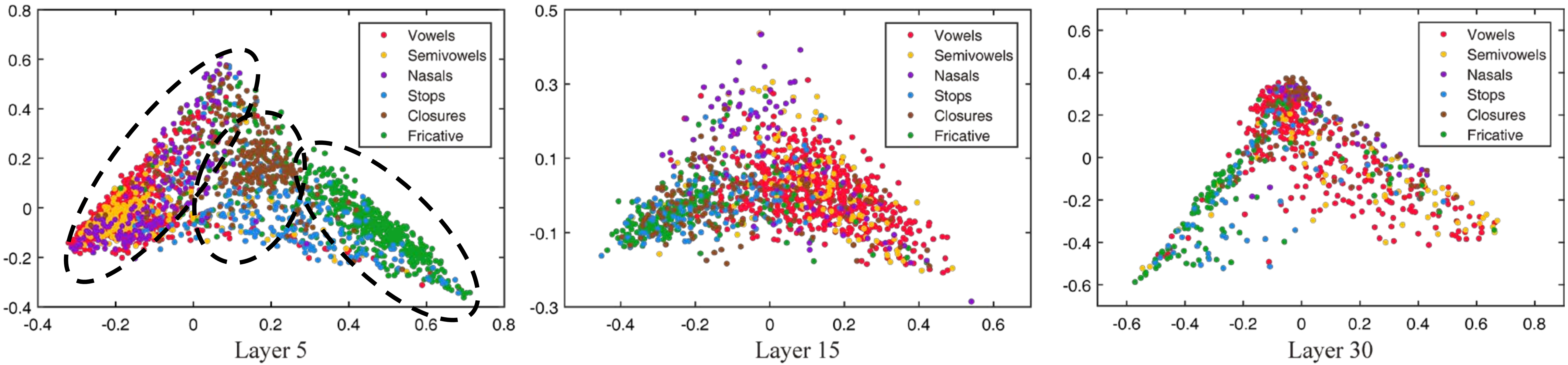
- **At layer 5**, the distribution of attention weights can be divided into three groups:
  - **voiced sounds** : vowels + semivowels + nasals
  - **stops** : stops + closures
  - **fricative-likes** : fricative

grouping with **similar acoustic characteristics and phoneme generation mechanisms!**

Results and Analysis

[Analysis of adaptive kernels in relation to phonemes]

➢ **Speaker FDAS1** (the largest variance of attention weights)



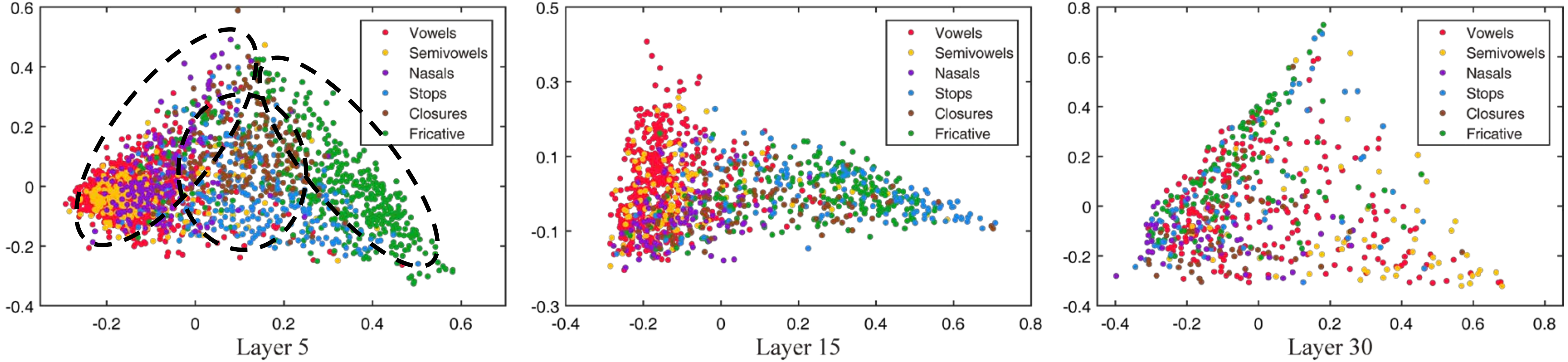
- The **boundary** between the distribution of groups starts to seem **vague at layer 15**, and the distributions completely merge and groups become **indistinguishable at layer 30**.

Attention weights are more phoneme-specific at earlier layers.
Only speaker information has been remained at later layers.

Results and Analysis

[Analysis of adaptive kernels in relation to phonemes]

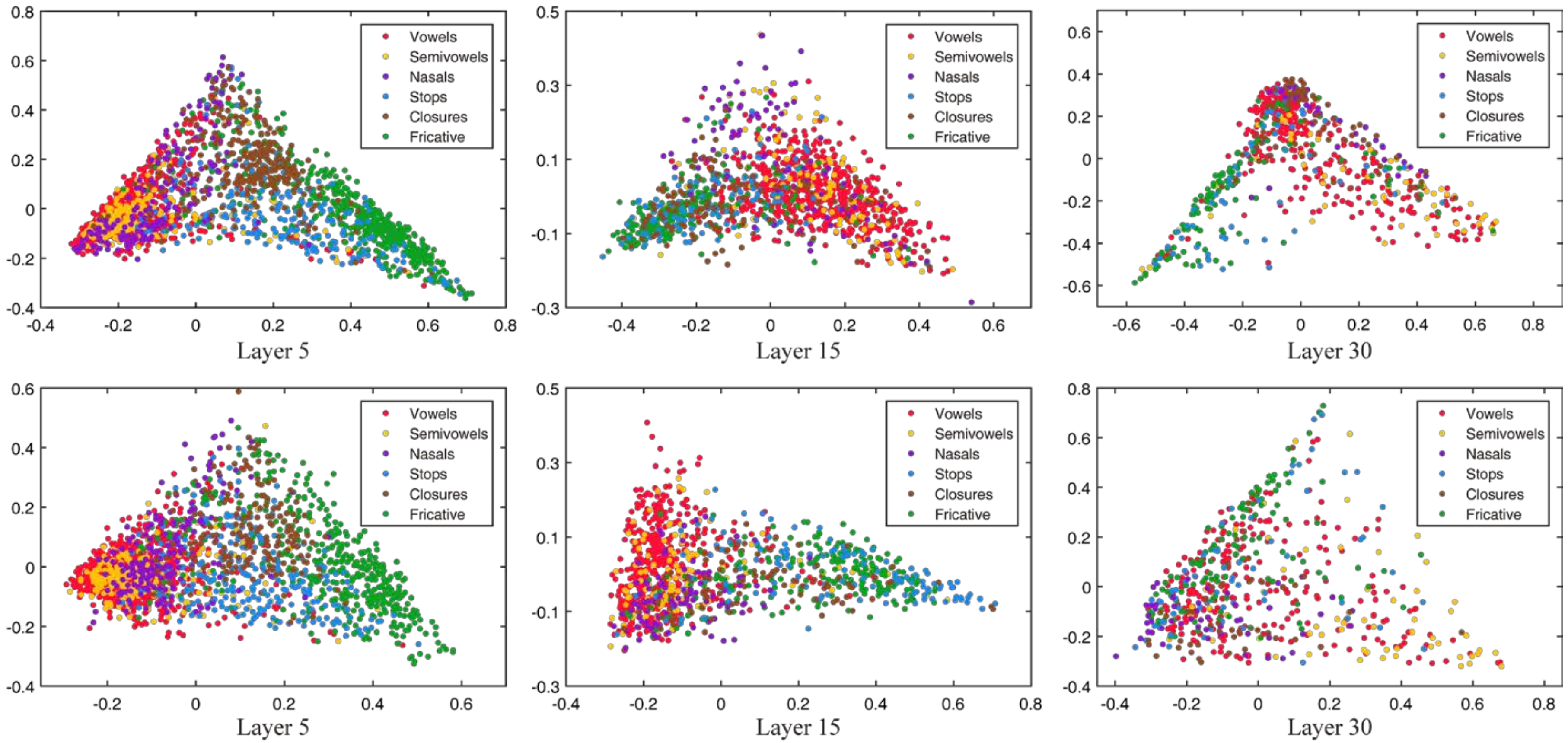
- **Speaker MHRM0** (the smallest variance of attention weights)



- **Similar results** were shown in the other speaker case.
  - The distribution of attention weights can be divided into **three groups**.
  - The distribution of groups **merge at later layers**.

Results and Analysis

[Analysis of adaptive kernels in relation to phonemes]



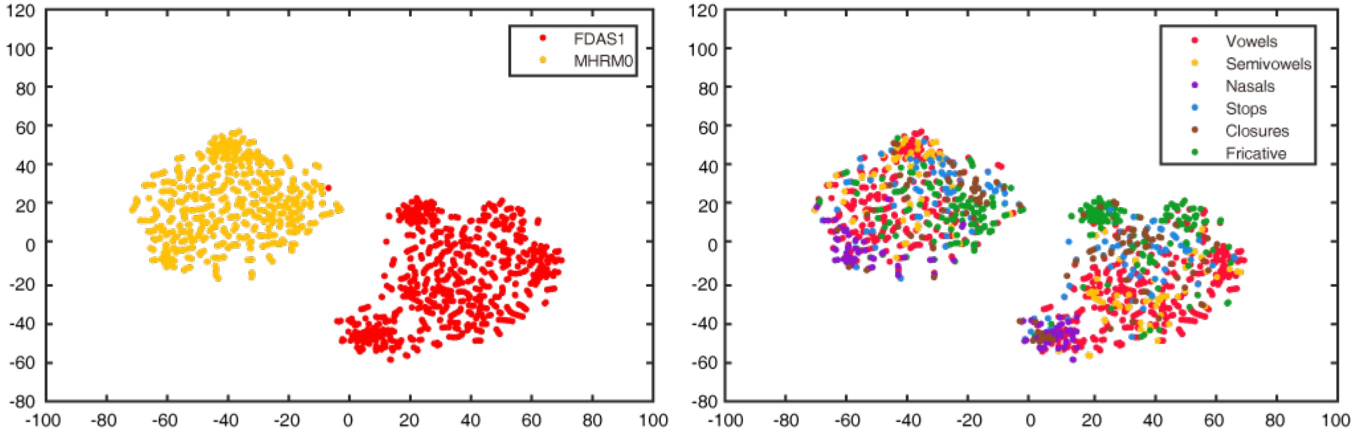
- ➢ The kernels were **adapted to phoneme groups**.
- ➢ **Only the speaker information is extracted** without phonetic information **in the speaker embeddings**.

Results and Analysis

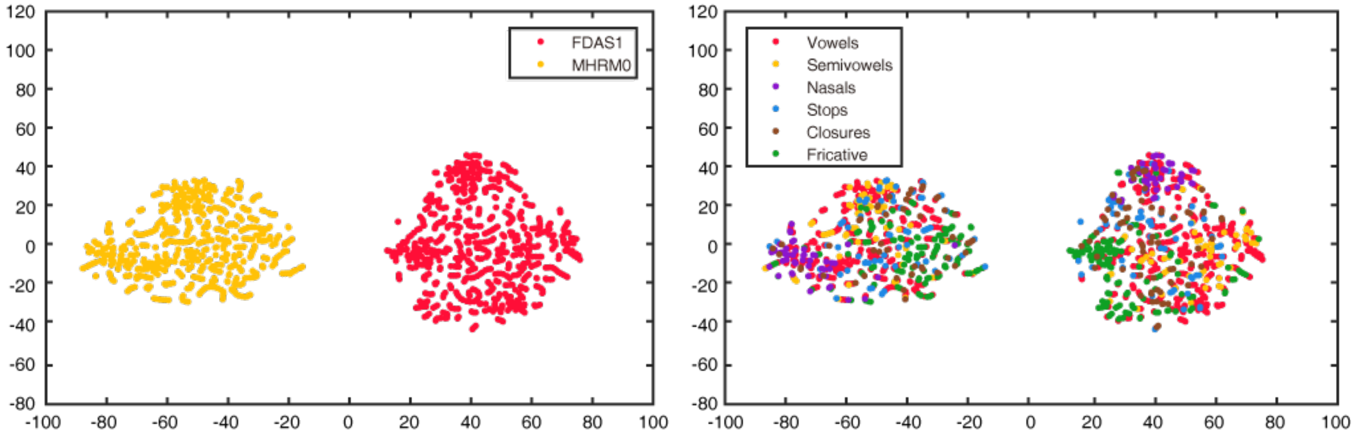
[Analysis of frame-level embeddings in relation to phonemes]

- Speaker embeddings are well gathered.
- Embeddings of nasals and fricatives within MHRM0 are far from the center of group in ResNet-34(×0.25).
- Embeddings by TDY-ResNet-34(×0.25) are closely gathered regardless of phoneme groups .
- Therefore, **TDY-CNN adapts to phonemes** and **extract consistent speaker embeddings regardless of phonemes**.

t-SNE projection of frame-level speaker embeddings



ResNet-34(×0.25)



TDY-ResNet-34(×0.25)

Conclusion

- **TDY-CNN** was proposed to **extract consistent speaker information** on **different time bins** for text-independent speaker verification.

- Models with TDY-CNN <u>**extract consistent speaker embeddings regardless of phonemes using phoneme-adaptive kernels**</u> and **improve speaker verification performance**.

- This work is the **first to analyze how temporal dynamic models work** depending on time bins and phonemes.

- The results indicate that **temporal dynamic models are suitable** and **consideration of phoneme information is crucial** in **text-independent speaker verification**.

icassp 2022 Singapore

KAIST HUMAN-MACHINE INTERACTION LAB

Thank You

seonghu.kim@kaist.ac.kr