清华大学 | 讲义分享

语言识别技术研讨会

2019.3.24
olr.cslt.org

XMUSPEECH Systems for OLR Challenge 2018

Miao Zhao(赵淼), Shijiang Yan(颜世江), Zheng Li(李铮),
Lin Li(李琳), Qingyang Hong(洪青阳)

2019.03

Outline

- About XMU Speech Lab
- XMUSPEECH Systems
    1.Data Preparation
    2.Our Systems
    3.Results and Conclusions
- Future Direction

About XMU Speech Lab

- Research directions:
  - Speaker/Language Recognition
  - Speech Recognition
  - Speech Synthesis
  - Microphone Array

- Research teams:
  - Three professors
  - More than twenty graduate students

- Website: speech.xmu.edu.cn

第十四届全国人机语音通讯学术会议
The 14th National Conference on Man-Machine Speech Communication, NCMMSC2017
闭 幕 式

1.Data Preparation

Three tasks

- Short-utterance identification task.
- Confusing-language identification task.
- Open-set recognition task.

The dataset used in our experiments:

Name	Datasets	Total Utt.	Length
BaseTrain	AP16-OL7, AP17-OL3-train/dev	72234	full
Thchs30-train	THCHS30-train	10000	full
Task_1_dev	AP17-OLR-test-1s	22051	1s
Task_1_eval	Task_1	21456	1s
Task_2_enroll	AP16-OL7, AP17-OL3-train/dev (Cantonese, Korean and Mandarin)	16553	full
Task_2_dev	AP17-OLR-test-all (Cantonese, Korean and Mandarin)	7354	full
Task_2_eval	Task_2	7357	full
Task_3_dev	AP17-OLR-test-all (3k), MinNan (1264)	4264	full
Task_3_eval	Task_3	40416	full

Data augmentation

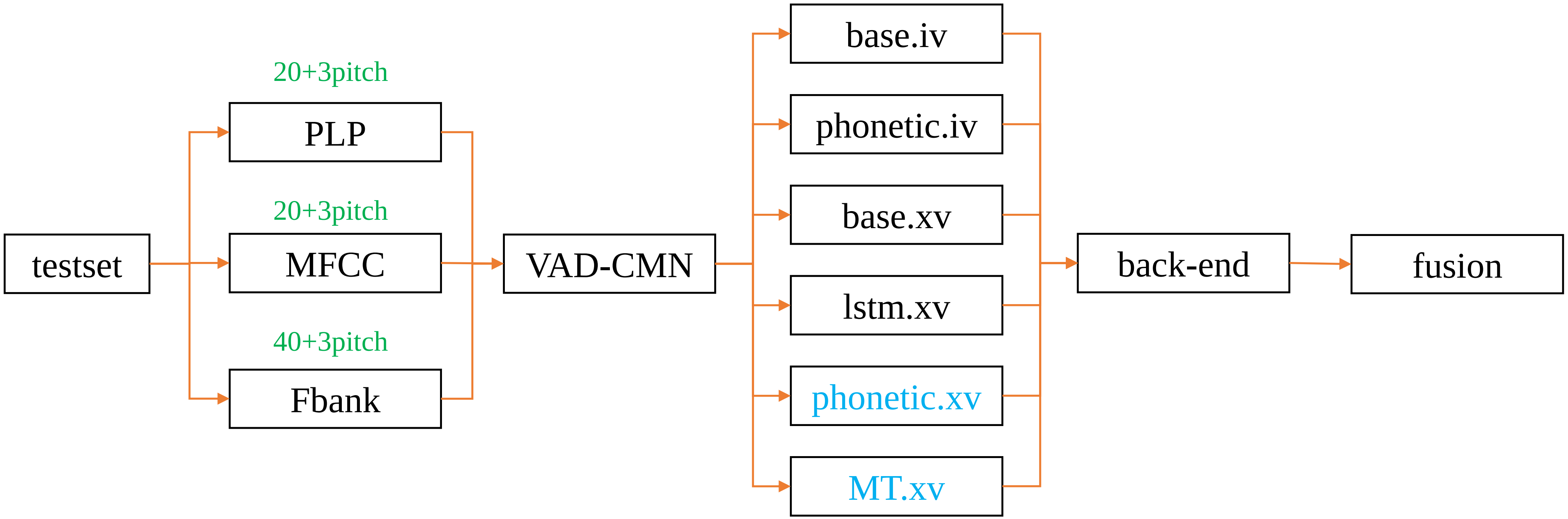
- Speed perturbation with 0.9 and 1.1 factor.
- Volume perturbation with random factor.

Z. Tang, D. Wang, and Q. Chen, “AP18-OLR Challenge: Three Tasks and Their Baselines,” in APSIPA ASC. IEEE, 2018.

2.Our systems

The fusion system

- The three tasks share this fusion system with a little difference in back-end and fusion step.
- At first, there are 18 sub-systems with 3 acoustic features and 6 extractors, but very redundant.
- The “iv” means i-vector model and the “xv” is x-vector framework.



testset

20+3pitch

PLP

20+3pitch

MFCC

40+3pitch

Fbank

VAD-CMN

base.iv

phonetic.iv

base.xv

lstm.xv

phonetic.xv

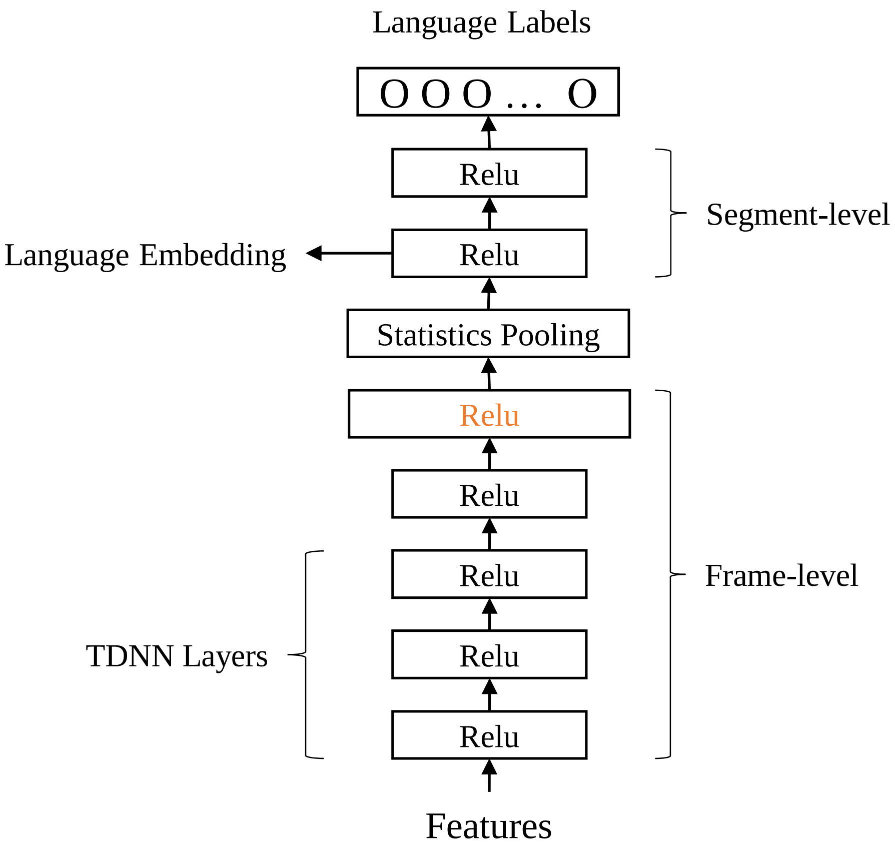
MT.xv

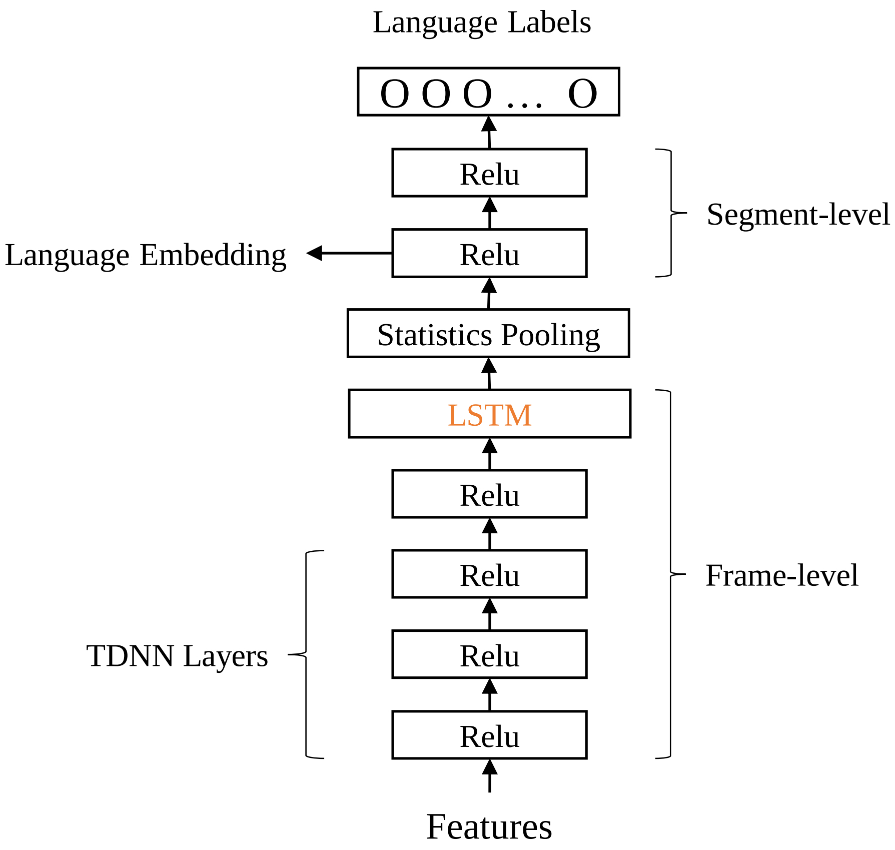
back-end

fusion

2.Our systems

The base.xv is a standard x-vector framework proposed in speaker recognition.



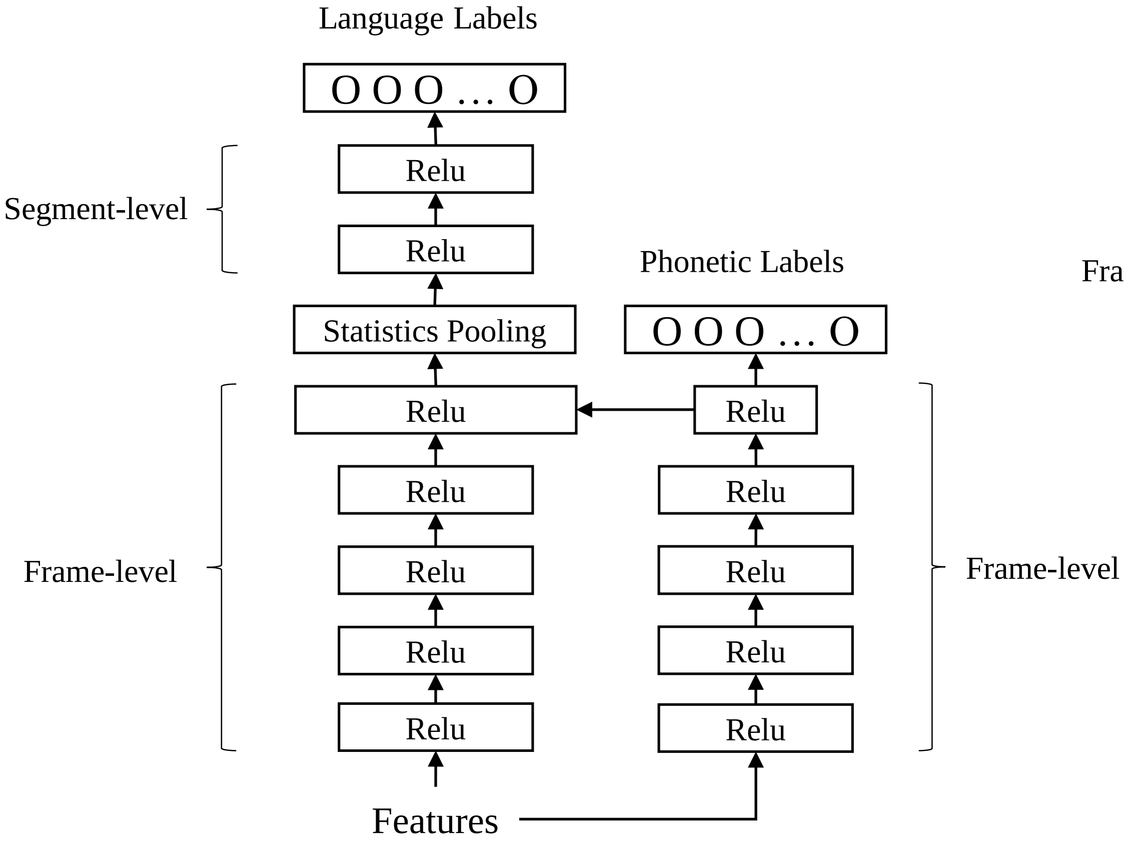


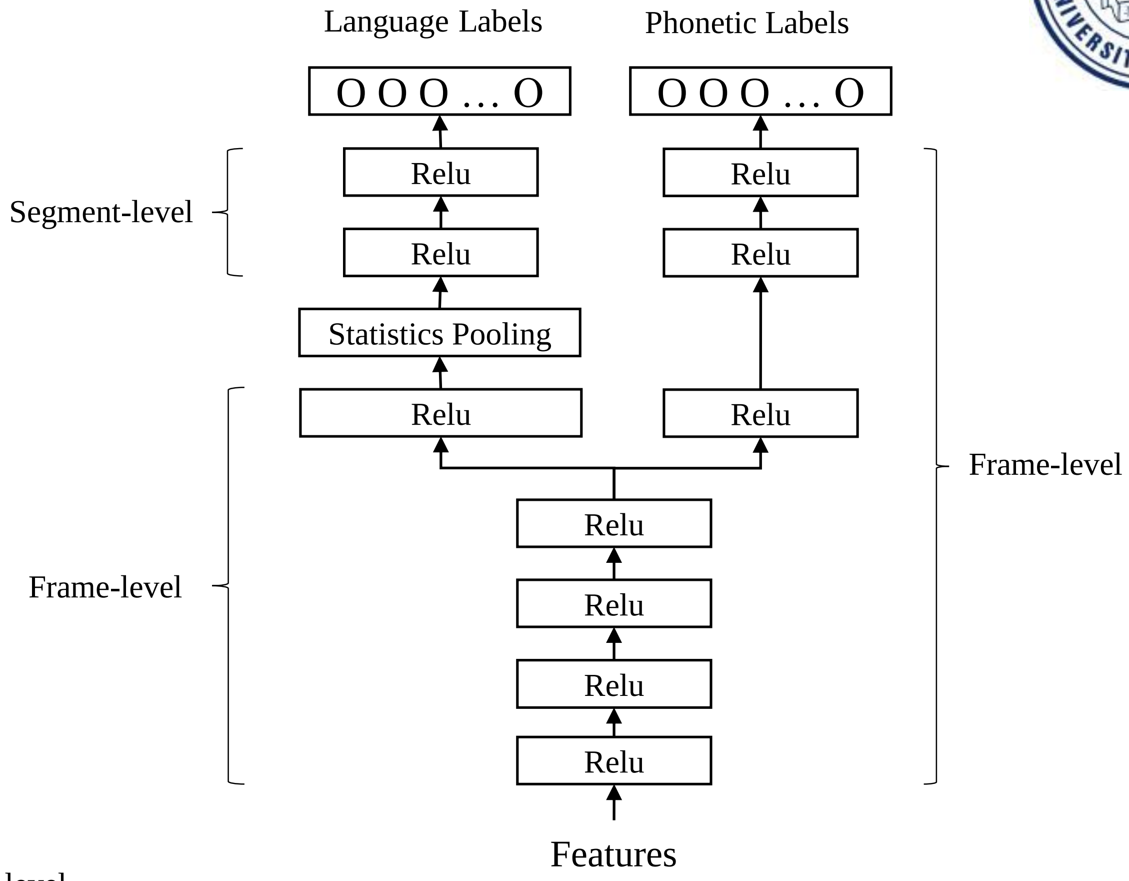
Based on the x-vector framework, we replace the 5'th hidden layer with LSTM, which is our attempt but not work well.

D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," Interspeech, 2017.

2.Our systems

The phonetic.xv is x-vector framework with phonetic information by splicing phonetic embeddings in hidden layers.





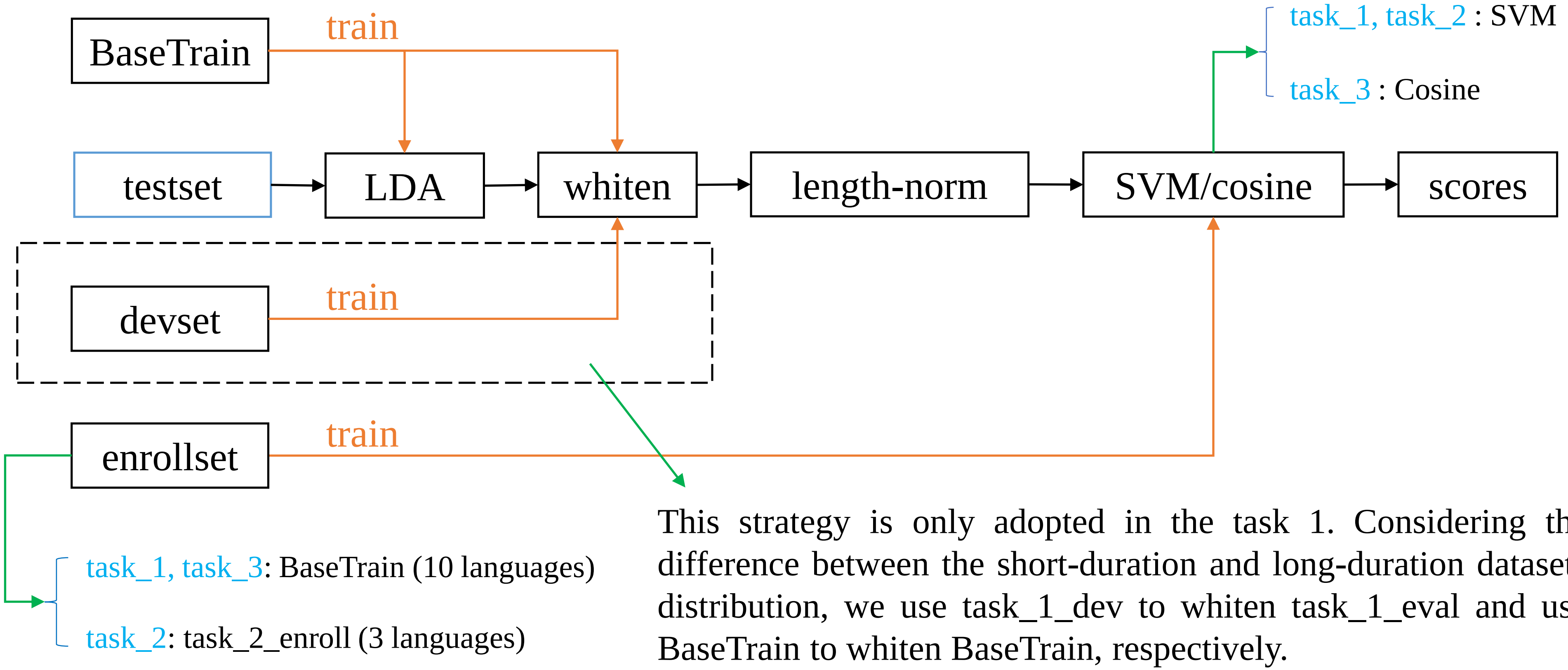
The MT.xv is a multitask x-vector framework and its first four hidden layers are shared layers.

Y. Liu, L. He, J. Liu, M. T. Johnson, “Speaker Embedding Extraction with Phonetic Information,” Interspeech, 2018.

2.Our systems

The back-end

- In the task 1 and 2, we use SVM model with RBF kernel as the back-end classifier.
- In the task 3, we use the Cosine Distance Scoring to compute the final scores, because it is an open-set task.

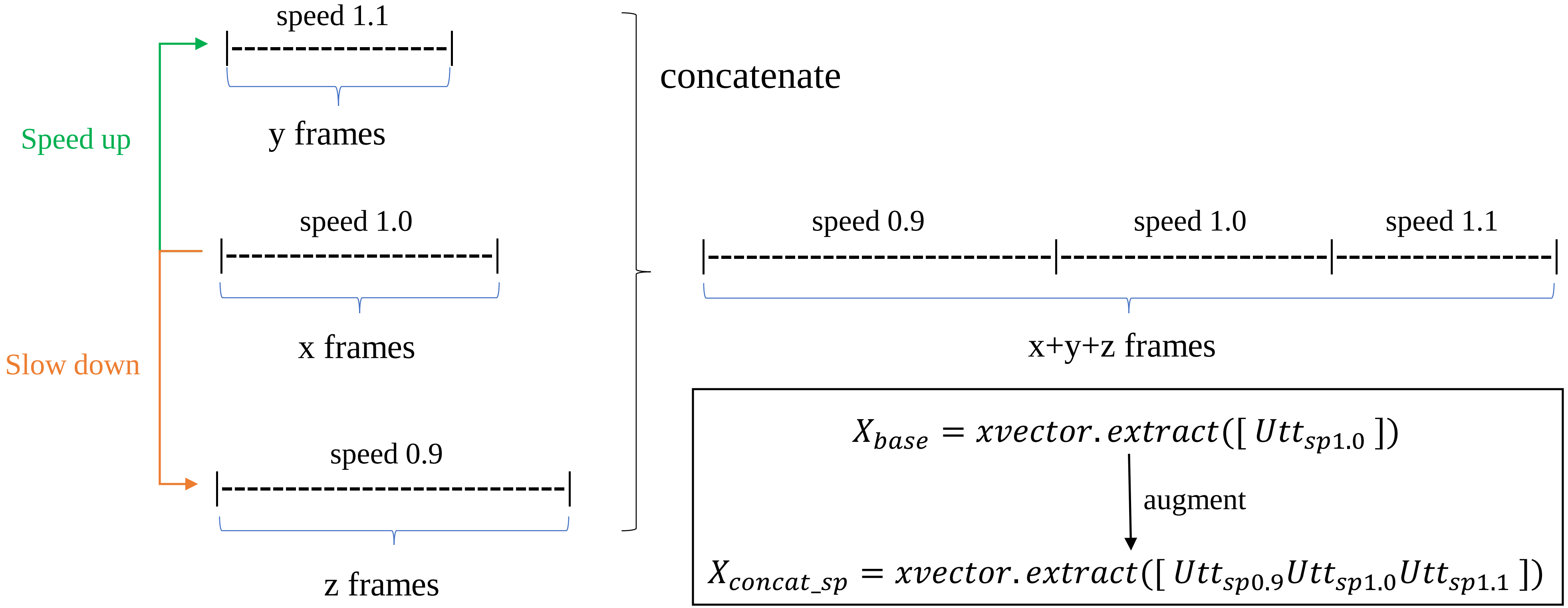


This strategy is only adopted in the task 1. Considering the difference between the short-duration and long-duration datasets distribution, we use task_1_dev to whiten task_1_eval and use BaseTrain to whiten BaseTrain, respectively.

2.Our systems

Augment testset by concatenating different speed utterances (concat-sp) before extracting it’s x-vectors.

- There are information compensation with light speed perturbation, especially for short utterances.
- The x-vector model should be trained by speed perturbation augmentation to match this process.
- It should be used with whitening method to normalize the data distribution.



$$\begin{aligned} X_{base} &= xvector.extract([\,Utt_{sp1.0}\,]) \\ &\Big\downarrow \text{augment} \\ X_{concat\_sp} &= xvector.extract([\,Utt_{sp0.9}Utt_{sp1.0}Utt_{sp1.1}\,]) \end{aligned}$$

X. X. MIAO, J. ZHANG, H. B. SUO, R. H. ZHOU, Y. H. YAN, “Expanding the length of short utterances for short-duration language recognition,” Journal of Tsinghua University (Science and Technology), 2018, 58(3): 254-259.

2.Our systems

Fusion methods

- Fusing with SVM weights for task 1 and fusing with equal weights for task 2 and task 3.
- The SVM fusion method is also used to remove the redundant subsystems whose weights are negative or too small.

SVM fusion method

Scores	system1	system2	system3	……	Label
Utt1	4.23	3.14	1.55	……	target
Utt2	-0.4	1.0	2.0	……	nontarget
⋮	⋮	⋮	⋮		⋮

We expect the target scores are as big as possible and the nontarget scores are the opposite, which helps to reduce the equal error rate.

These scores coming from different subsystems is combined as a vector sample to train a linear SVM model with their target or nontarget labels, to discriminate the target fusion scores and nontarget fusion scores.

Training by $y \cdot \left(\sum_{i=1}^{n} w_i \cdot score_i + b \right) \geq 1, y = 1, -1$ *then* $score_{fusion} = \sum_{i=1}^{n} w_i \cdot score_i$

The weights of linear kernel are the final fusion weights which are trained by devset and used in testset.

3.Results and Conclusions

The equal error rate of a part of subsystems.

Subsystems /EER%	**Feature type**	**task 1**		**task 2**		**task 3**	
		dev	eval	dev	eval	dev	eval
base.xv	PLP	7.934	7.140	3.318	1.509	4.667	4.334
base.xv	MFCC	8.315	7.465	3.046	1.4	4.367	4.073
lstm.xv	MFCC	7.866	7.427	2.584	1.509	4.50	4.712
phonetic.xv	PLP	**6.918**	**6.691**	2.040	1.074	4.133	4.055
phonetic.xv	MFCC	7.040	6.854	2.067	1.169	3.967	3.99
MT.xv	PLP	7.617	7.250	**1.836**	**0.680**	4.267	**3.612**
MT.xv	MFCC	7.848	7.479	3.046	0.938	**3.667**	3.822
Fusion	-	**4.786**	**4.590**	**1.346**	0.690	**3.333**	**3.160**

Conclusions

- PLP is more stable than MFCC in language recognition.
- Data augmentation is very useful.
- Training model with phonetic information could bring better performance.
- The concat-sp method for short duration testset could make results better.
- Fusion could obtain further improvement in general.

Future Direction

Limitation

- We focus on task 1 and have no time to optimize task 2 and task 3 specially.
- The current back-end is not the best process and there are better classifiers, like logical regression.
- There are still redundant when selecting subsystems for fusion.

Future Direction

- Analyze why the concat-sp method needs some special setups to make it work.
- Optimize the multi-task x-vector framework for short duration utterances.
- Make our fusion strategy more robust.

Thank you ! Any questions?