CHINESE ACADEMY OF SCIENCES
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

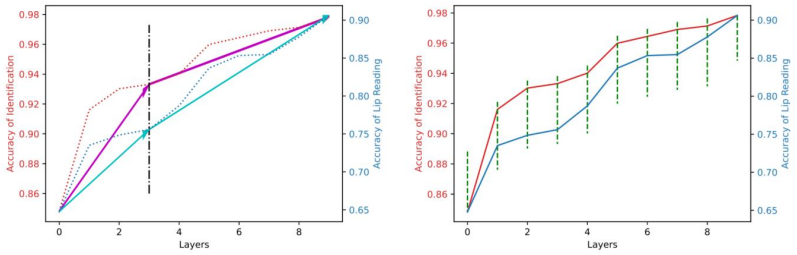
BMVC 2023

code is available

Learning Separable Hidden Unit Contributions for Speaker-Adaptive Lip-Reading

Songtao Luo, Shuang Yang, Shiguang Shan, Xilin Chen

Motivation



(a) Overall Growth Trend (b) Layer-by-layer Growth Trend

Figure 1: Accuracy of Lip Reading and Identification Using the Output at Different Layers

Features extracted from lip reading network's intermediate layers of **varying depths** for both tasks (lip reading and speaker verification):

- **Speaker-dependent** features are well-represented in the **shallow** layers. As the **depth** increases, the level of abstraction improves only slightly.
- **Content-dependent** features have relatively poorer representation in the **shallow** layers. As the **depth** increases, the level of abstraction improves uniformly.

Method

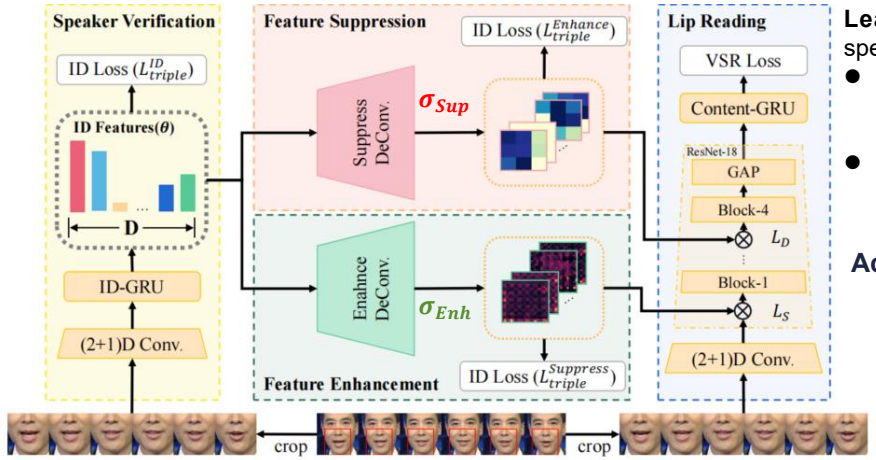
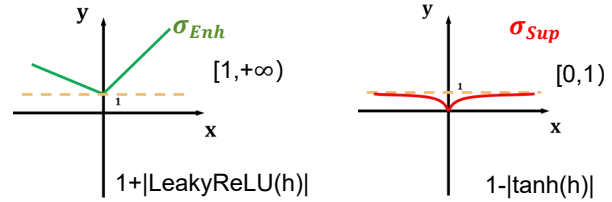


Figure 2: The Overall Architecture of Our Proposed Method.

Learning Separable Hidden Unit Contributions: Differentiate speaker and content contributions at different layers.

- **Shallow Layer Strategy:** Enhance content-dependent features. Use the speaker's features to lead the model to prioritize content-dependent features.
- **Deep Layer Strategy:** Suppress content-independent features. Introduce the speaker's features to further suppress noisy features irrelevant to the content.

Activation Functions Constraint



For Enhancement **For Suppression**

Experiments

Dataset

We establish **CAS-VSR-S68:**

- **68**-hour **11** hosts **3,800** Chinese characters

LRW-ID

- **word-level English 500words**
- **Speaker Adaptation Split of LRW**

GRID

- **sentence-level English fixed corpus**

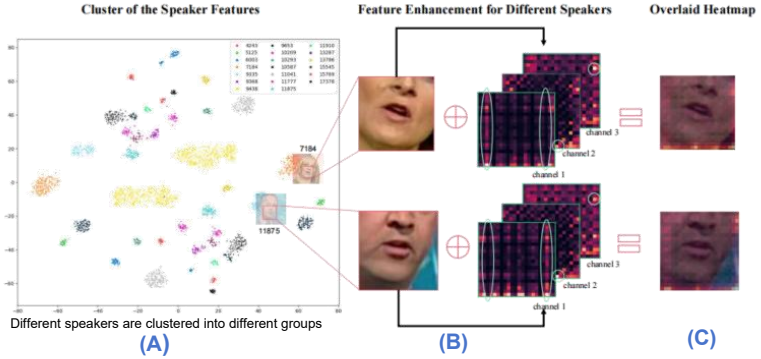


Figure 3: Visualization of the Generated Enhancement Weights

- **(A): Discriminative** Speaker's Features
- **(B): Variability** in enhancement weights across speakers. (in green)
- **(C):** Enhanced regions **beyond lips**.

Ablation Study

Loss Ablation

Method	Acc(%)
Baseline	87.25
Enhance Only	87.83
Suppress Only	87.81
Enhance & Suppress	**87.91**

Module Ablation

Method	L^{ID}_{triple}	L^{Enh}_{triple}&L^{Sup}_{triple}	L^{VSR}_{CE}	Acc(%)
Baseline	-	-	√	87.25
Ours	x	x	√	87.73
	√	x	√	87.74
	x	√	√	87.75
	√	√	√	**87.91**

Comparison with others

LRW-ID (limited adaptation data)

Adapt min.	User-padding[1]	Prompt Tuning[2]	Base -line	Proposed Method
0	85.85	87.54	87.25	**87.91**
1	87.06	88.53	88.52	**89.21**
3	87.61	89.45	89.48	**89.88**
5	87.91	89.99	89.96	**90.45**

CAS-VSR-S68 (limited adaptation data)

Adapt min.	Base -line	Proposed Method
0	19.61	19.37
1	*21.53*	*20.69*
3	18.65	18.55
5	17.55	**16.72**

GRID (no adaptation data)

Method	WER (%)
WAS	14.6
LipNet	11.4
TM-seq2seq	11.7
User-padding[1]	11.12
User-padding*	7.2
Prompt Tuning[2]	12.04
TVSR-Net	9.1
DVSR-Net	7.8
Visual i-vector	7.3
Baseline	10.62
Ours	9.59
Ours*	**6.99**

* apply unsupervised method