Optimizing Speech Multi-View Feature Fusion through Conditional Computation

Weiqiao Shan, Yuhao Zhang, Yuchen Han, Bei Li, Xiaofeng Zhao, Yuang Li, Min Zhang,Hao Yang, Tong Xiao, Jingbo Zhu

School of Computer Science and Engineering, Northeastern University, Shenyang, China
The Chinese University of Hong Kong, Shenzhen, China
Meituan, Beijing, China
Huawei Translation Services Center, China
NiuTrans Research, Shenyang, China

Introduction

- **The SSL-based features (S-feature) alleviate the variability of signals in a fixed representation space using the vector quantization (VQ) method**
  This type of feature:
  I. easy to learn and operate.
  II. offers multi-view representations for speech processing tasks.
  III. suffers from information loss during the VQ process (difficult to achieve the strongest performance compared to spectral features in complicated tasks).

Can we leverage the strengths of the two views of feature to improve complex speech-to-text tasks?

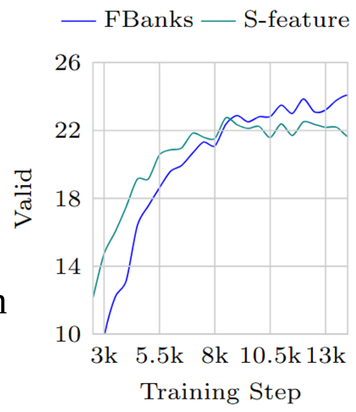


Figure.1: The BLEU scores of the model using only FBanks or S-feature.

Complementary and Interference

- **S-features and Spectral features**
  I. **Complementary** information exists between the two features, Fbank features give better performance but slow convergence, S-features converge faster but have poorer performance.
  II. **Interference** between two features, direct fusion can not take advantage of the advantages of each of the two features.

S-feature from Hubert (960h hours pretrained data)

Model	En-De Test	En-De Epoch	En-Fr Test	En-Fr Epoch	En-Es Test	En-Es Epoch
S2T(FBanks-only)	25.08	40	36.22	35	30.04	37
S2T(S-feature-only)	23.33	27	33.00	25	26.52	23

S-feature from Wavlm (9.4w hours pretrained data)

Model	En-De Valid	En-De Test	En-De Epoch	En-Fr Valid	En-Fr Test	En-Fr Epoch	En-Es Valid	En-Es Test	En-Es Epoch
S2T(FBanks-only)	25.43	25.08	40	30.77	36.22	35	33.27	30.04	37
S2T(S-feature-only)	25.23	25.46	25	30.7	36.06	21	33.22	29.85	25
S2T-Concat	23.53	23.72	29	29.42	34.1	28	33.22	30.3	23

Table.1: BLEU scores of the model using only FBanks or S-feature in multiple speech translation direction.

Better performance **FBanks-only**
Rapid convergence **S-feature-only**
Concatenate
Performance degradation
Relative rapid convergence

Gradient-sensitive Gating Network (GSGN)

- **Our primary analysis reveals a significant divergence in the update directions between FBanks and S-feature.**
  I. We calculated the cosine similarity between the gradients of the two features (Figure 3, left) and measured the probability of the cosine value being less than zero across the entire training set (Figure 3, right).
  II. In the model trained from scratch, the gradient cosine similarity fluctuated significantly during training, indicating that the directions of the two gradients increasingly diverged—reflecting gradient instability.
  III. A cosine similarity less than zero implies that the gradients have opposing components, pointing to a notable gradient conflict issue, particularly in the pre-trained model.

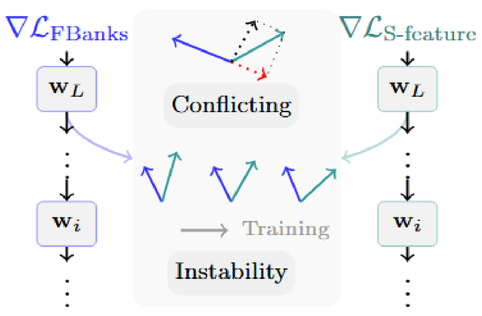


Figure.2: Gradient conflict and gradient instability.

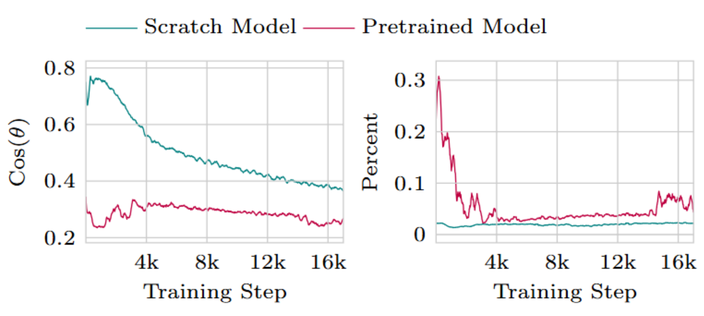


Figure.3: The Cos(θ) between two gradients generated by each feature.

- **we propose a gradient-sensitive gating network (GSGN). The system consists of an acoustic encoder, a textual encoder and a decoder, and an fusion layer.**

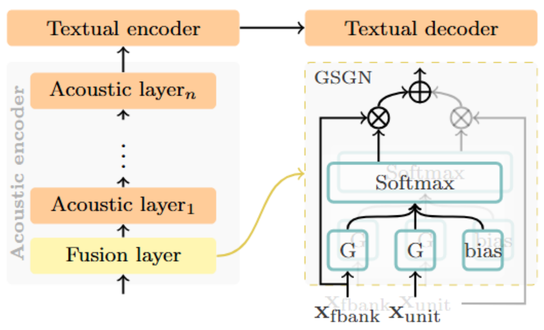


Figure.3: ST model architecture.

$$\mathbf{x}_{\text{fusion}} = \mathbf{g}_{\text{fbank}} \circ \mathbf{x}_{\text{fbank}} + \mathbf{g}_{\text{unit}} \circ \mathbf{x}_{\text{unit}}$$

$$\begin{cases} \mathbf{1} \circ \vec{\boldsymbol{g}}_i + \mathbf{1} \circ \vec{\boldsymbol{g}}_j & \text{if } \cos(\theta) \geq 0 \\ \left(\mathbf{1} - \frac{\|\vec{\boldsymbol{g}}_j\| \cos(\theta)}{\|\vec{\boldsymbol{g}}_i\|}\right) \circ \vec{\boldsymbol{g}}_i + \mathbf{1} \circ \vec{\boldsymbol{g}}_j & \text{else} \end{cases}$$

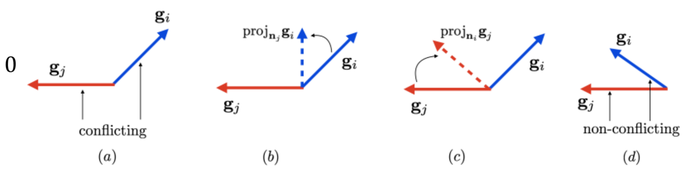
$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \mathbf{g}_{\text{fbank}} \circ \frac{\partial \mathcal{L}_{\text{fbank}}}{\partial \mathbf{w}_i} + \mathbf{g}_{\text{unit}} \circ \frac{\partial \mathcal{L}_{\text{unit}}}{\partial \mathbf{w}_i}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \underbrace{(\mathbf{g}_{\text{fbank}} \circ \mathbf{x}_{\text{fbank}} + \mathbf{g}_{\text{unit}} \circ \mathbf{x}_{\text{unit}})^{\text{T}}}_{\text{Input-dependent}} \times \underbrace{\sum_{j=0, j\neq i}^{n} (\mathbf{w}_j^{\text{T}} + \mathbf{1}) \frac{\partial \mathcal{L}}{\partial \mathbf{h}_{\text{out}}}}_{\text{Input-independent}}$$

we find that the gradient when use the fused features can be expressed as a weighted summation of the gradient use two single feature separately.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \mathbf{g}_{\text{fbank}} \circ \frac{\partial \mathcal{L}_{\text{fbank}}}{\partial \mathbf{w}_i} + \mathbf{g}_{\text{unit}} \circ \frac{\partial \mathcal{L}_{\text{unit}}}{\partial \mathbf{w}_i}$$

This weighted exactly corresponding to the gating weights in the fusion layer.



We prevent the problem of conflicting gradients of different features by constraining the gating weights.

Multi-stage Gropout

- **To enhance the model robustness to each input feature and ensure that all features can be fully utilized, we dynamically sample from three features during the training stage and input them into the next layer**

$$\mathbf{x} = \begin{cases} \mathbf{x}_{\text{fbank}}, & \text{when} \quad p < \delta_{\text{fbank}} \\ \mathbf{x}_{\text{unit}}, & \text{when} \quad \delta_{\text{fbank}} <= p < \delta_{\text{fbank}} + \delta_{\text{unit}} \\ \mathbf{x}_{\text{fusion}}, & \text{when} \quad \delta_{\text{fbank}} + \delta_{\text{unit}} <= p \end{cases}$$

Experimental Setups

- **Datasets:** MuST-C (English-German, English-French and English-Spanish)
- **Metrics:** sacreBLEU

Model Settings	Value
Encoder layers	12
Textual encoder layers	6
Textual decoder layers	6
Hidden size	512
Feedforward size	2048
Attention head	8
Hidden size (Fusion layer)	512

Training Settings	Value
Attention dropout	0.1
Residual connection dropout	0.2
Adam	$\beta_1 = 0.9, \beta_2 = 0.98$
Warmup	4000
Early stop	10

Inference Settings	Value
Ensemble	10
Beam	5
Length penalt	1.0

Main Results

I. Our approach achieves fast convergence while guaranteeing model performance by effectively fusing the two features.
II. Our approach can achieve better performance than spectral features when based on a better S-features which from wavlm.

S-feature from Hubert (960h hours pretrained data)

Model	En-De Test	En-De Epoch	En-Fr Test	En-Fr Epoch	En-Es Test	En-Es Epoch
S2T(FBanks-only)	25.08	40	36.22	35	30.04	37
S2T(S-feature-only)	23.33	27	33.00	25	26.52	23
S2T-GSGN	24.18	28	33.91	25	28.61	24
S2T-GSGN+Drop	25.26 (+0.18)	30 (×1.33)	36.24 (+0.02)	30 (×1.17)	30.24 (+0.20)	30 (×1.23)

S-feature from Wavlm (9.4w hours pretrained data)

Model	En-De Valid	En-De Test	En-De Epoch	En-Fr Valid	En-Fr Test	En-Fr Epoch	En-Es Valid	En-Es Test	En-Es Epoch
S2T(FBanks-only)	25.43	25.08	40	30.77	36.22	35	33.27	30.04	37
S2T(S-feature-only)	25.23	25.46	25	30.7	36.06	21	33.22	29.85	25
S2T-Concat	23.53	23.72	29	29.42	34.1	28	33.22	30.3	23
S2T-GSGN	25.76	25.49 (+0.41)	26 (×1.54)	31.04	36.27 (+0.05)	25 (×1.4)	33.26	30.36 (+0.32)	22 (×1.68)
S2T-GSGN+Drop	26.22	26.21 (+1.13)	33 (×1.21)	31.66	37.29 (+1.07)	32 (×1.09)	34.06	30.97 (+0.93)	33 (×1.12)

Table.3: The final result of our method.

Conclusion

- **Feature fusion based on conditional computation:** Our proposed method effectively resolves the conflict between the two features, which can provide comparable performance and average 1.24 times training speed up.
- **Better S-feature:** Improved S-features may lead to greater gains in speech translation and potentially benefit a wider range of speech processing tasks.
- **Future work:** Enhance the precision of our method, more accurately measure the relationship between two features.