Optimizing Speech Multi-View Feature Fusion through Conditional Computation

ICASSP 2025

Weiqiao Shan[1], Yuhao Zhang[2*], Yuchen Han[1], Bei Li[3], Xiaofeng Zhao[4], Yuang Li[4], Min Zhang[4], Hao Yang[4], Tong Xiao[1,5†], Jingbo Zhu[1,5]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China

[2]The Chinese University of Hong Kong, Shenzhen, China

[3]Meituan, Beijing, China

[4]Huawei Translation Services Center, Beijing, China

[5]NiuTrans Research, Shenyang, China

* Equal contribution

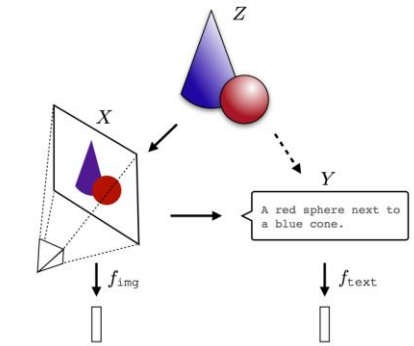
† Corresponding author

Background

- Large-scale datasets often contain a variety of multi-view data
  - ➢ These data have different patterns, sources, and modalities
  - ➢ But they share highly similar high-level semantic information

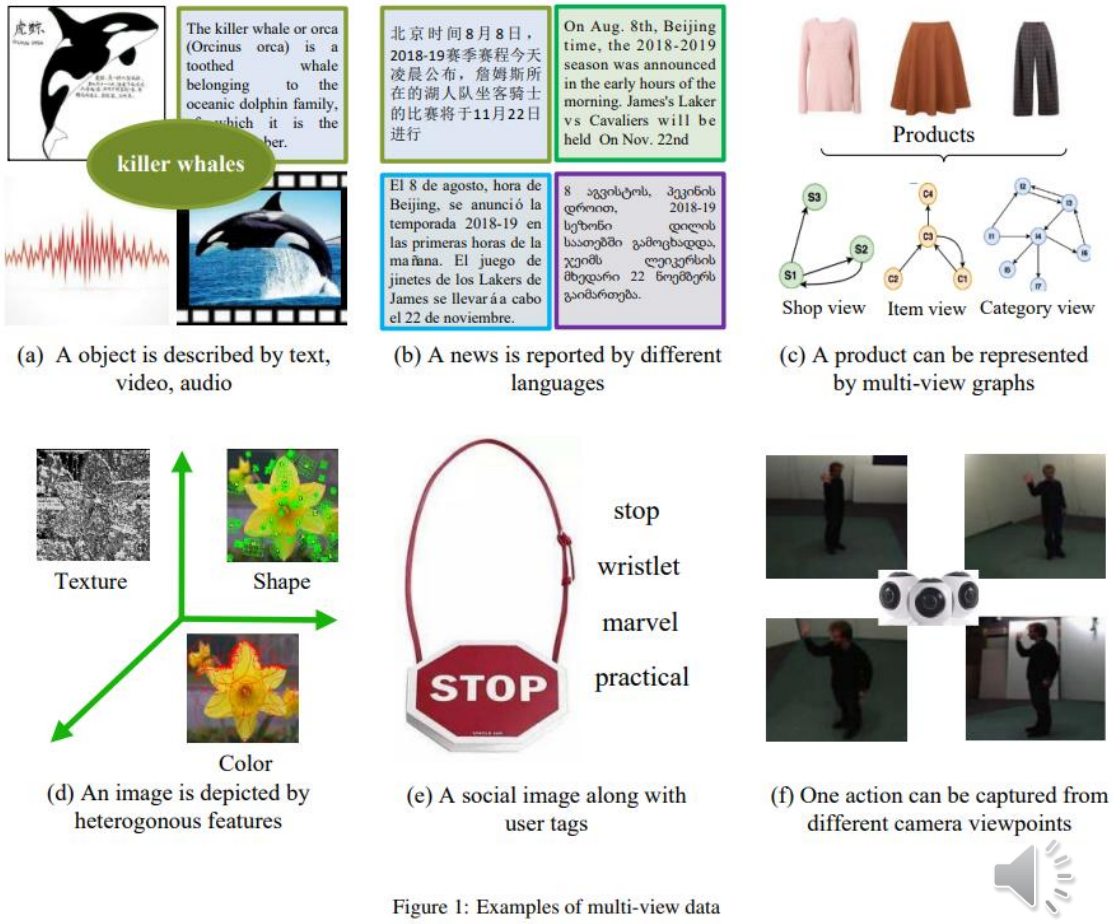
The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



Multi-view learning:

Multi-view learning aims to learn the common feature spaces or shared patterns by combining multiple distinct features or data sources



(a) A object is described by text, video, audio

(b) A news is reported by different languages

(c) A product can be represented by multi-view graphs

(d) An image is depicted by heterogonous features

(e) A social image along with user tags

(f) One action can be captured from different camera viewpoints

Figure 1: Examples of multi-view data

Yan X, Hu S, Mao Y, et al. Deep multi-view learning methods: A review[J]. Neurocomputing, 2021, 448: 106-129.
Huh M, Cheung B, Wang T, et al. The platonic representation hypothesis[J]. arXiv preprint arXiv:2405.07987, 2024.
Oncevay A, Haddow B, Birch A. Bridging linguistic typology and multilingual machine translation with multi-view language representations[J]. arXiv preprint arXiv:2004.14923, 2020.

Background

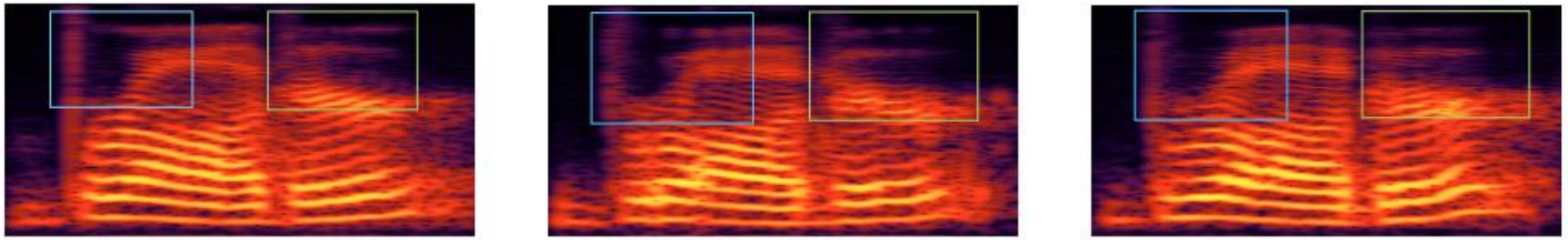
- In speech processing tasks, there is also multi-view data
  - ➢ Multiple Resolutions

High-resolution:

better envelope information in the time domain

low-resolution:

more detailed information in the frequency domain



(a) 20ms (b) 40ms (c) 100ms

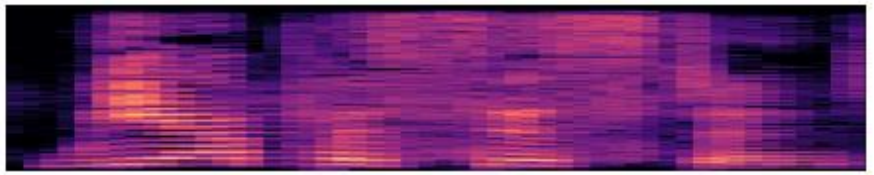
Figure 2: *Speech re-synthesis using features from HuBERT base models at different resolutions. The high-resolution features capture better envelope information in the time domain (shown in the blue box), while the low-resolution features provide more detailed information in the frequency domain (shown in the green box). See Sec. 3.4 for experimental details and discussion.*

Shi J, Tang Y, Inaguma H, et al. Exploration on HuBERT with multiple resolutions[J]. arXiv preprint arXiv:2306.01084, 2023.
Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.

Background

- In speech processing tasks, there is also multi-view data
  - Multiple Resolutions
  - Spectral features (Fbank) and Self-supervised learning features (Unit)

S-features 334 226 666 277 746 991 162 535 271 930
327 905 104 108 778 759 547 241 798 432 ...

spectral 

Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.

Question

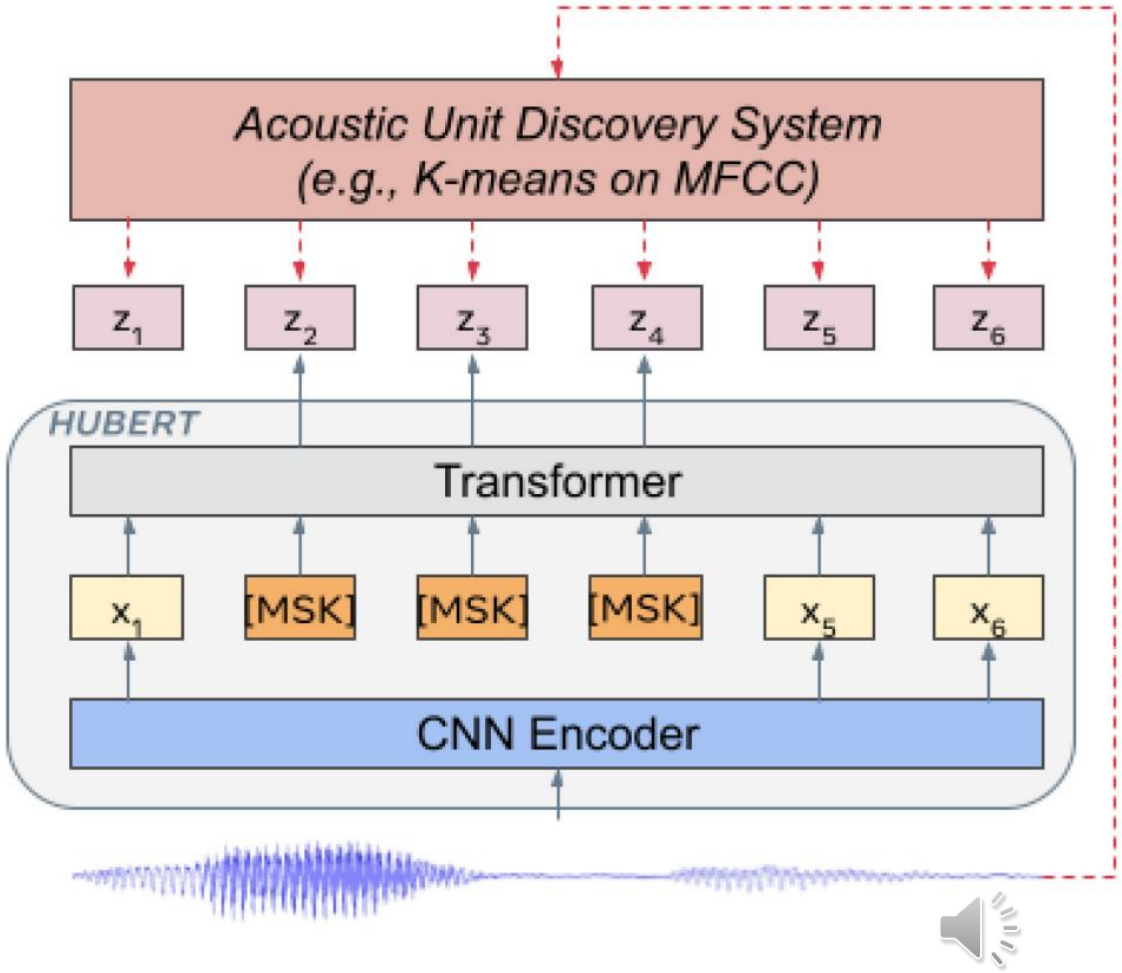
- Self-supervised learning features (S-features):
  - refer to discrete speech units derived from raw audio or spectral features (e.g., MFCC, FBanks) using a BERT-based model with clustering techniques..
  - provide multi-view representations for speech processing tasks and exhibit a strong capability for integrating contextual information.

Question 1:

Do multi-view features provide distinct advantages in speech processing tasks?

Question 2:

If so, are these advantages complementary?



Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.

Phenomena (Speech Translation, ST)

- Phenomenon 1:
  - There exists **complementary** information between the two features: Fbank features yield better performance but exhibit slower convergence, while S-features converge more quickly but provide lower performance
- Phenomenon 2:
  - **Interference** between the two features means that direct fusion does not fully leverage the advantages of each feature

TABLE I
MAIN RESULT OF OUR METHOD.

Hubert

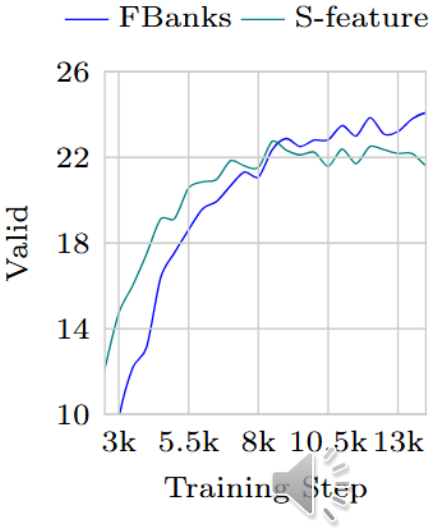
960 hours

Model	En-De		En-Fr		En-Es	
	Test	Epoch	Test	Epoch	Test	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

Wavlm

9.4w hours

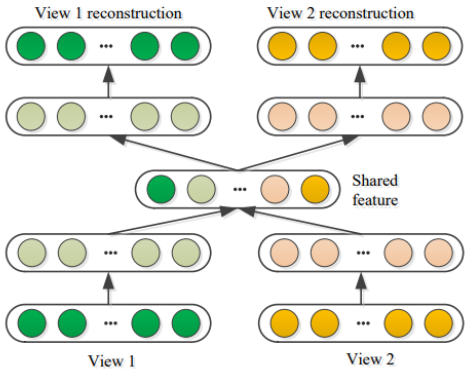
Model	En-De			En-Fr			En-Es		
	Valid	Test	Epoch	Valid	Test	Epoch	Valid	Test	Epoch
S2T(FBanks-only)	25.43	25.08	40	30.77	36.22	35	33.27	30.04	37
S2T(S-feature-only)	25.23	25.46	25	30.7	36.06	21	33.22	29.85	25
S2T-Concat	23.53	23.72	29	29.42	34.1	28	33.22	30.3	23



Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- Traditional multi-view learning requires a large number of additional parameters



Is there a **simple** and **effective** method to leverage both features while minimizing interference between them?

Figure 5: The framework of bimodal auto-encoder (BAE) (adapted from [57]).

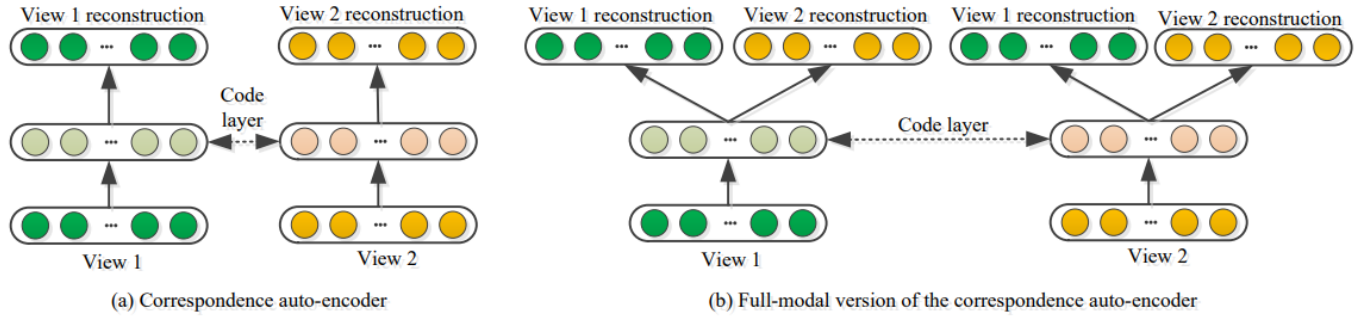


Figure 6: The framework of correspondence auto-encoder (Corr-AE) and its full-modal version (adapted from [35]).

Yan X, Hu S, Mao Y, et al. Deep multi-view learning methods: A review[J]. Neurocomputing, 2021, 448: 106-129.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Architecture:
  - ➢ consists of an **acoustic encoder** with 12 layers, a **textual encoder** and **decoder** with 6 layers each, and a **fusion layer**.

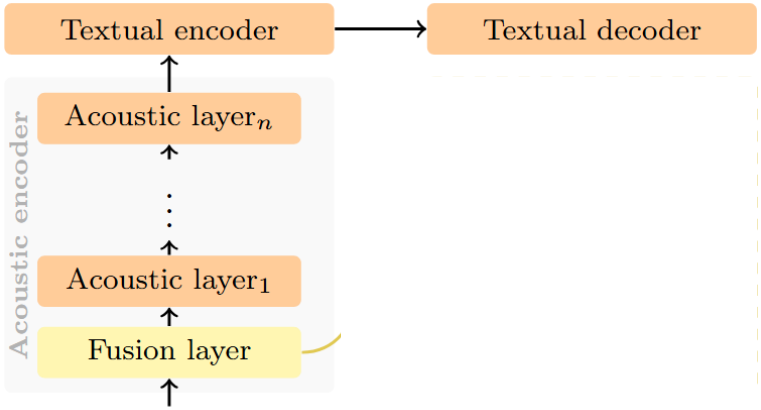
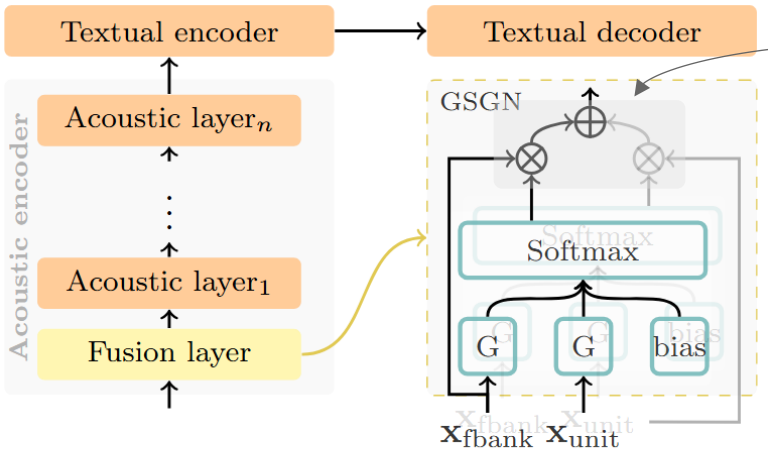


Fig. 2. ST model architecture.

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Architecture:
  - ➢ consists of an **acoustic encoder** with 12 layers, a **textual encoder** and **decoder** with 6 layers each, and a **fusion layer**.



➢(Fusion layer) outputs a weighted sum of the two input features

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

Fig. 2. ST model architecture.

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Architecture:
  - ➢ consists of an **acoustic encoder** with 12 layers, a **textual encoder** and **decoder** with 6 layers each, and a **fusion layer**.

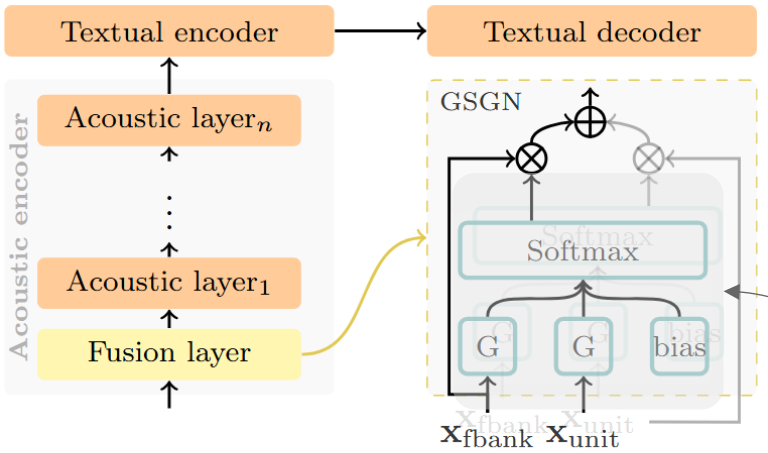


Fig. 2. ST model architecture.

➢(Fusion layer) outputs a weighted sum of the two input features

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

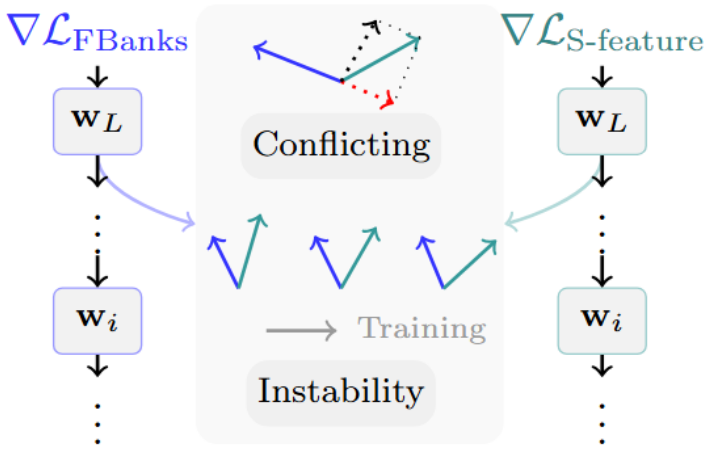
➢(Gating vector $g_{[\cdot]}$) serves as the weight for each input, computed based on a linear mapping applied to each input

$$\mathbf{g}_{[\cdot]} = \text{Sigmoid}(\text{Linear}^1_{[\cdot]}(\mathbf{x}_{\text{fbank}}) + \text{Linear}^2_{[\cdot]}(\mathbf{x}_{\text{unit}}) + \text{bias}_{[\cdot]})$$

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Derive:
  - ➢GSGN for Conflicting Gradient



(b) Angle of gradient

$$\mathbf{h}_{i+1} = \underbrace{F(\mathbf{h}_i)}_{\text{residual}} + \underbrace{\mathbf{h}_i}_{\text{shortcut}} \approx \underbrace{\mathbf{w}_i\mathbf{h}_i + \mathbf{b}_i}_{\text{residual}} + \underbrace{\mathbf{h}_i}_{\text{shortcut}}$$

链式求导

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

(A + B) × C = A × C + B × C (分配律)

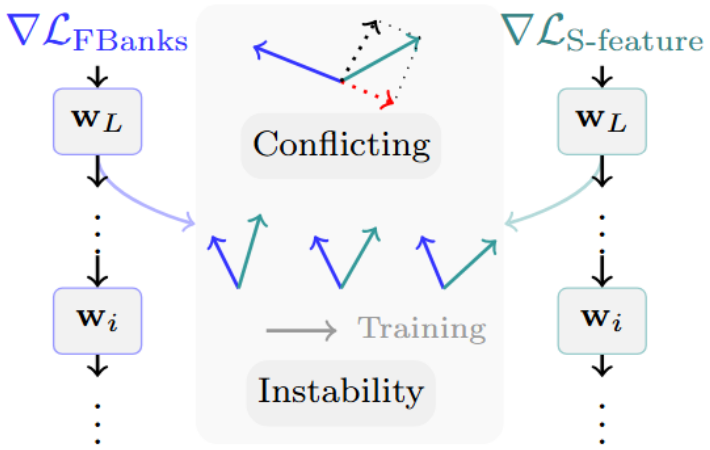
$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}_i} = (\mathbf{g}_{\text{fbank}} \circ \mathbf{x}_{\text{fbank}})^{\text{T}} \times \sum_{j=0,j\neq i}^{n} (\mathbf{w}_j^{\text{T}} + \mathbf{1}) \frac{\partial\mathcal{L}}{\partial\mathbf{h}_{\text{out}}} + (\mathbf{g}_{\text{unit}} \circ \mathbf{x}_{\text{unit}})^{\text{T}}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

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

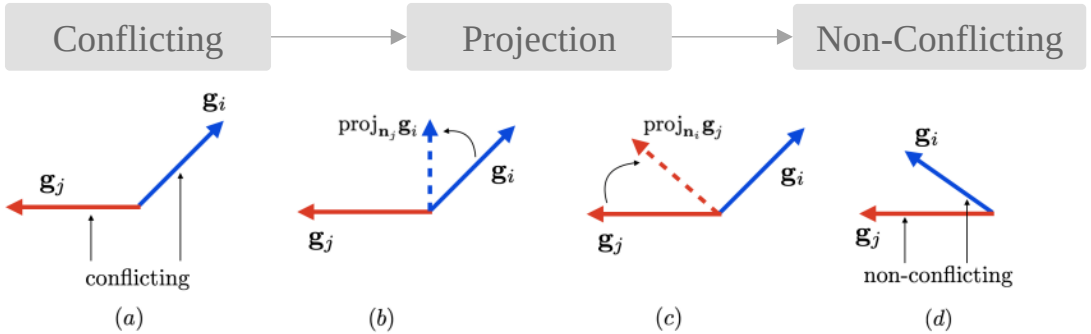
Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Derive:
  - ➢GSGN for Conflicting Gradient



(b) Angle of gradient



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

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

Additional Loss

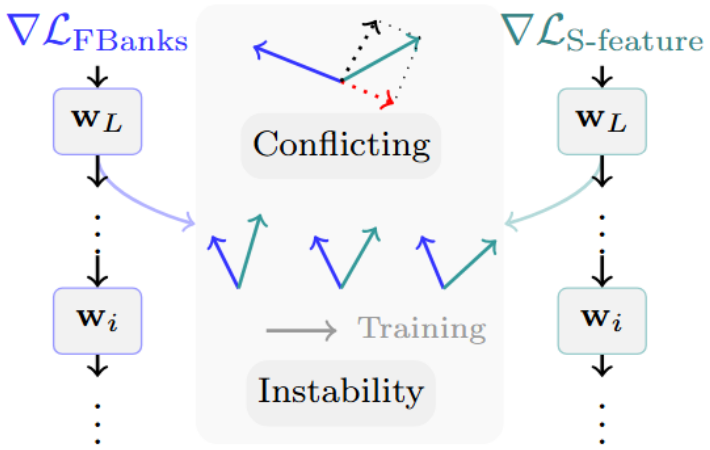
$$\mathcal{L}_{\text{gate}} = \begin{cases} \text{MSE}(\mathbf{g}_{\text{fbank}}, \mathbf{1}) & \text{if } \cos(\theta) >= 0, \\ \text{MSE}(\mathbf{g}_{\text{fbank}}, (\mathbf{1} - \frac{\| \frac{\partial \mathcal{L}_{\text{unit}}}{\partial \mathbf{w}_i} \| \cos(\theta)}{\| \frac{\partial \mathcal{L}_{\text{fbank}}}{\partial \mathbf{w}_i} \|})) & \text{else} \end{cases}$$

Yu T, Kumar S, Gupta A, et al. Gradient surgery for multi-task learning[J]. Advances in neural information processing systems, 2020, 33: 5824-5836.

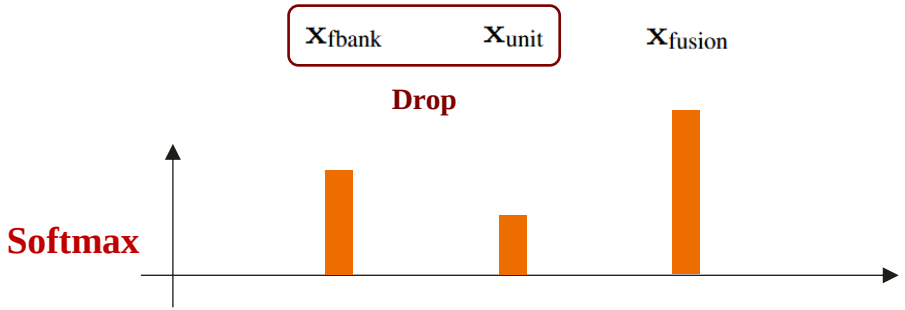
Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Methodology

- The Gradient-sensitive Gating Network (GSGN) :
  - ➢ implements conditional computation for the gradient of parameters
  - ➢ dynamically computes the correlation between heterogeneous features and adjusts the weights of different feature branches to achieve optimal fusion.
- Derive:
  - ➢GSGN for Instability Gradient



(b) Angle of gradient



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

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Result

- Our method :
  - ➢leverages the complementary information between the two features by dynamically adjusting the activation parameters, while preventing interference between them

$$\mathbf{g}_{[\cdot]} = \text{Sigmoid}(\text{Linear}^1_{[\cdot]}(\mathbf{x}_{\text{fbank}}) + \text{Linear}^2_{[\cdot]}(\mathbf{x}_{\text{unit}}) + \text{bias}_{[\cdot]})$$

The dimension of the linear projection in $g_{[\cdot]}$ is $\mathbb{R}^{d_{model} * d_{model}}$, with the parameters approximating those in an attention layer

Hubert

Wavlm

TABLE I
MAIN RESULT OF OUR METHOD.

Model	En-De		En-Fr		En-Es	
	Test	Epoch	Test	Epoch	Test	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)

TABLE II
MAIN RESULT OF OUR METHOD.

Model	En-De			En-Fr			En-Es		
	Valid	Test	Epoch	Valid	Test	Epoch	Valid	Test	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)

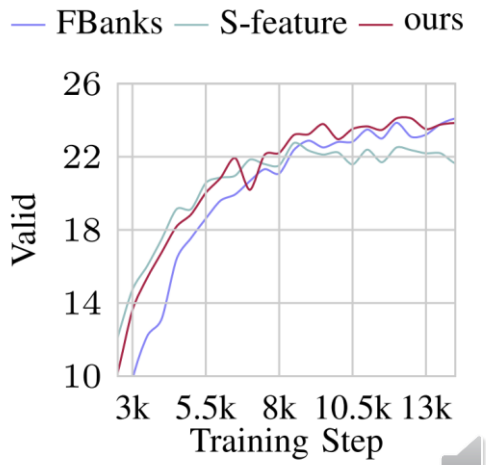


Fig. 5. The final result of our method.

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

ANALYSIS

- Conflicting and Instability :
  - ➢ Pre-trained model compared to the model trained from scratch

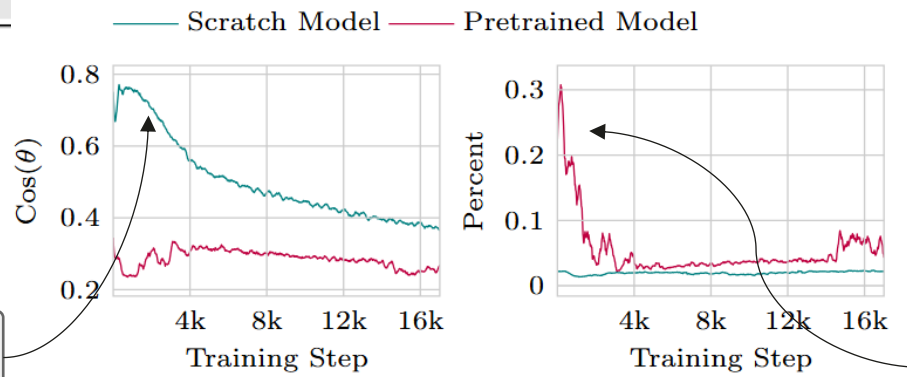
Model	En-De Test	En-De Epoch
S2T(FBanks-only)	25.08	40
S2T(S-feature-only)	23.33	27
S2T-GSGN	24.18	28
S2T-GSGN+Drop	25.26 (+0.18)	30 (×1.33)

In the model trained from scratch

GSGN not good enough, but **GSGN+Drop** can give **better** results

Instability
$\cos\theta$ change significantly

Conflicting
A large percentage of $\cos\theta < 0$



En-De	Test	Epoch
S2T(FBanks-only)	27.98	33
S2T(S-feature-only)	25.60	25
S2T-GSGN	27.61	25
S2T-GSGN+Drop	28.19 (+0.21)	27 (×1.22)

In the Pre-trained model

GSGN performs well, GSGN+Drop does **not** show **substantial improvement** compared to a model trained from scratch

Scratch Model: gradient instability problem becomes more apparent

Pretrained Model: Gradient Conflict Problem Becomes More Obvious

Fig. 3. The $\cos(\theta)$ between two gradients generated by each feature. We find **Instability**(left): The $\cos(\theta)$ becomes smaller and smaller with training for $\cos(\theta) > 0$, and **Conflicting**(right): Percent of conflicting gradients($\cos(\theta) < 0$) among all gradients in an input batch.

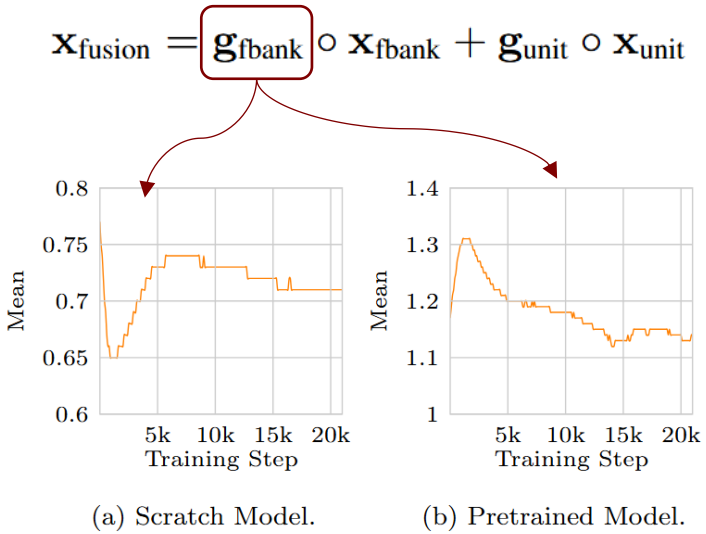
Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

ANALYSIS

- $g_{[\cdot]}$ in pre-trained model and the model training from scratch

Scratch Model:

Since the conflict is not severe in the model trained from scratch, GSGN relies more on S-features during the initial stage of training (g_{fbank} is relatively small), achieving faster convergence.



(a) Scratch Model. (b) Pretrained Model.

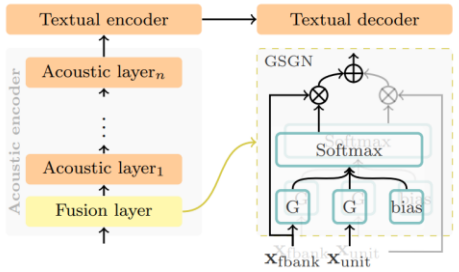
Fig. 4. The mean value of $\mathbf{g}_{\text{fbank}}$ in the model training from scratch (left) and the pre-trained model (right).

Pretrained Model:

The conflict problem is more pronounced in the pre-trained model. Therefore, during the early stage of training, GSGN utilizes more Fbank features to ensure optimal model performance(g_{fbank} is relatively large)

Shan W, Zhang Y, Han Y, et al. Optimizing Speech Multi-View Feature Fusion through Conditional Computation[J]. arXiv preprint arXiv:2501.08057, 2025.

Thank you.



GitHub

Table 1: Main result of our method based on better S-feature from **WavLM** model.

Model	En-De		En-Fr		En-Es	
	Test	Epoch	Test	Epoch	Test	Epoch
S2T(FBanks-only)	25.08	40	36.22	35	30.04	37
S2T(S-feature-only)	25.46	25	36.06	21	29.85	25
S2T-Concat	23.72	29	34.1	28	30.3	23
S2T-GSGN	25.49 (+0.41)	26 (×1.54)	36.27 (+0.05)	25 (×1.4)	30.36 (+0.32)	22 (×1.68)
S2T-GSGN+Drop	26.21 (+1.13)	33 (×1.21)	37.29 (+1.07)	32 (×1.09)	30.97 (+0.93)	33 (×1.12)

TABLE I
MAIN RESULT OF OUR METHOD.

Model	En-De		En-Fr		En-Es	
	Test	Epoch	Test	Epoch	Test	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)

HomePage Scratch Model: gradient instability problem becomes more apparent

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

NEU NLP SINCE 1973

HUAWEI