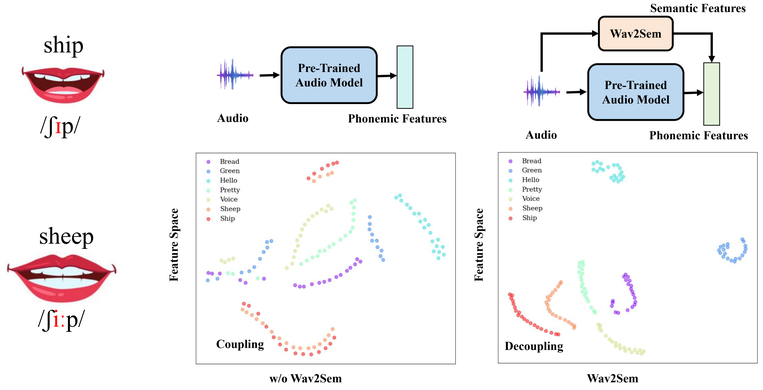
Wav2Sem: Plug-and-Play Audio Semantic Decoupling for 3D Speech-Driven Facial Animation

Hao Li[1,2], Ju Dai[2,*], Xin Zhao[1,2], Feng Zhou[3], Junjun Pan[1,2,*] , Lei Li[4,5]
1 Beihang University, 2 Peng Cheng Laboratory, 3 North China University of Technology,
4 University of Washington, 5 University of Copenhagen

CVPR Nashville JUNE 11-15, 2025

Coupling of homophonic features leads to lip sync error

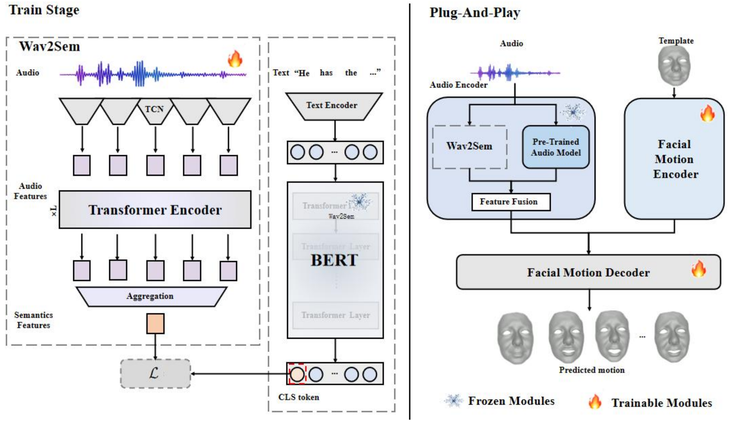


Our Contributions

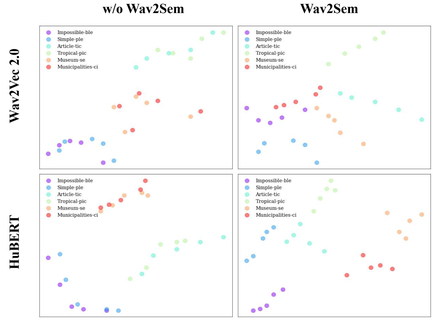
- **Problem:** Current self-supervised audio models overly focus on phoneme-level features, causing entanglement of near-homophonic syllables.
- **Solution:** We propose Wav2Sem, which aligns audio to the text semantic space and decouples the features of near-homophonic through semantic features in the audio.
- **Results:**Wav2Sem significantly enhances facial animation expressiveness and improves phoneme discrimination.

Plug-and-play audio semantic decoupling module

- Wav2Sem utilizes semantic features from input audio to achieve feature decoupling of near sound words.



Audio Feature Space Decoupling



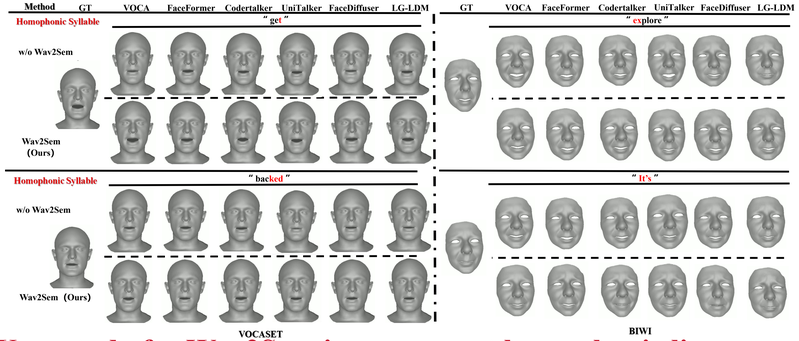
Model	w/o Wav2Sem	Wav2Sem
Wav2Vec 2.0	0.0397	0.0701
HuBERT	0.2689	0.2909

- Wav2Sem effectively improves the coupling phenomenon of near-homophonic syllable features in both quantitative and qualitative statistical results.

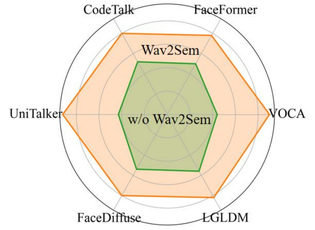
Quantitative evaluation for Wav2Sem by different framework

Methed	Audio Encoder	VOCASET			BIWI		
		MVE ($\times 10^{-5}$)	LVE ($\times 10^{-5}$)	FDD ($\times 10^{-7}$)	MVE ($\times 10^{-3}$)	LVE ($\times 10^{-4}$)	FDD ($\times 10^{-5}$)
VOCA	DeepSpeech	6.1571	4.9245	4.8447	8.3606	6.7158	7.5320
VOCA+Wav2Sem		**6.0015**	**4.8915**	**4.7814**	**8.3175**	**6.6821**	**7.3571**
Faceformer	Wav2Vec 2.0	5.1558	4.1090	4.6675	7.1658	4.9847	5.0972
Faceformer+Wav2Sem		**5.0962**	**3.9891**	**4.5317**	**7.0956**	**4.9571**	**4.9157**
Codetalker	Wav2Vec 2.0	4.6132	3.9445	4.4522	7.3980	4.7914	4.1170
Codetalker+Wav2Sem		**4.5762**	**3.8838**	**4.3654**	**7.2519**	**4.7751**	**3.9181**
UniTalker	Wav2Vec 2.0	4.1271	3.5416	4.7125	6.6104	4.0213	4.1296
UniTalker+Wav2Sem		**4.0973**	**3.1476**	**4.5174**	**6.5101**	**3.9112**	**4.0214**
FaceDiffuse	HuBERT	4.3651	3.7924	4.2647	6.8088	4.2985	3.9101
FaceDiffuse+Wav2Sem		**4.3493**	**3.7739**	**4.2523**	**6.5725**	**4.2521**	**3.8635**
LG-LDM	HuBERT	3.7162	3.7925	3.9154	7.7298	4.9869	4.3637
LG-LDM+Wav2Sem		**3.7042**	**3.7863**	**3.9014**	**7.6952**	**4.9258**	**4.3121**

Qualitative evaluation for Wav2Sem by different framework



User study for Wav2Sem improvement homophonic lip sync error



- Wav2sem can serve as a supplement to existing animation frameworks to improve lip sync error caused by feature coupling.