INTERSPEECH 2022
Sep 18 – 22 • Incheon Korea

SelfRemaster: Self-Supervised Speech Restoration with Analysis-by-Synthesis Approach Using Channel Modeling

Takaaki Saeki, Shinnosuke Takamichi, Tomohiko Nakamura, Naoko Tanji, and Hiroshi Saruwatari

The University of Tokyo, Japan

Outline

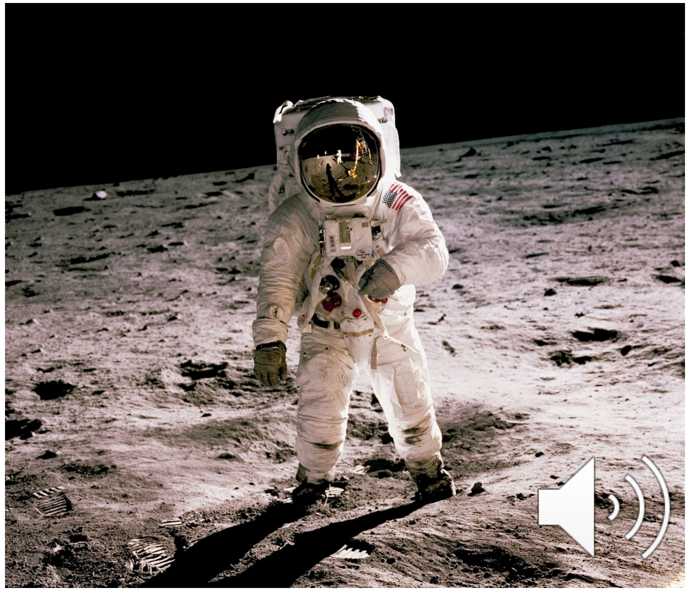
- ❑ Background
- ❑ Related Work
- ❑ Methods
- ❑ Experimental Evaluation
- ❑ Summary

Background: Speech Restoration

Need to use and analyze existing degraded speech data.

E.g.) Historical audio materials, telephone recordings, etc.

Containing low-resource languages, endangered cultures, etc.



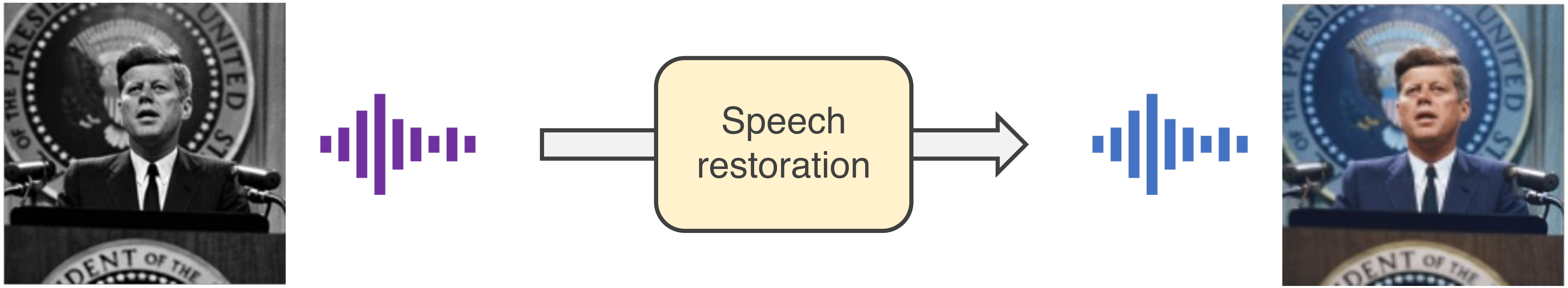
https://parade.com/.image/c_limit%2Ccs_srgb%2Cq_auto:good%2Cw_700/MTkwNTc5NTIyNDMyNTQyNTg4/1-19-martin-luther-king-ftr.webp
https://i.insider.com/5d2faa4d7e76cc3f20437ff6?width=700
https://www.fluentu.com/blog/japanese/wp-content/uploads/sites/6/2021/09/classic-japanese-movies-5.jpg

Background: Speech Restoration

Speech restoration: generating clean speech from degraded speech.

Low-quality old recordings

High-quality restored audio



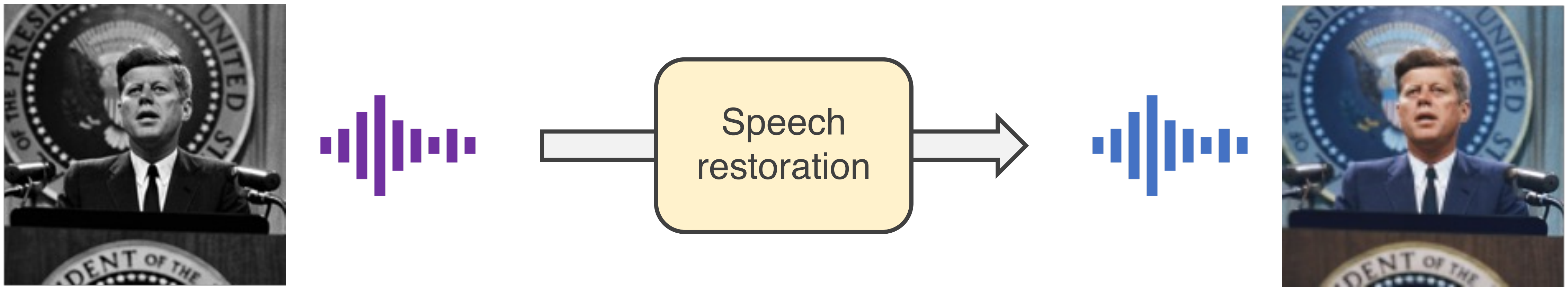
Background: Speech Restoration

Speech restoration: generating clean speech from degraded speech.

Speech restoration of real data is highly challenging.

- Paired training data are not available.
- Cannot use information on acoustics distortions (e.g., audio devices).

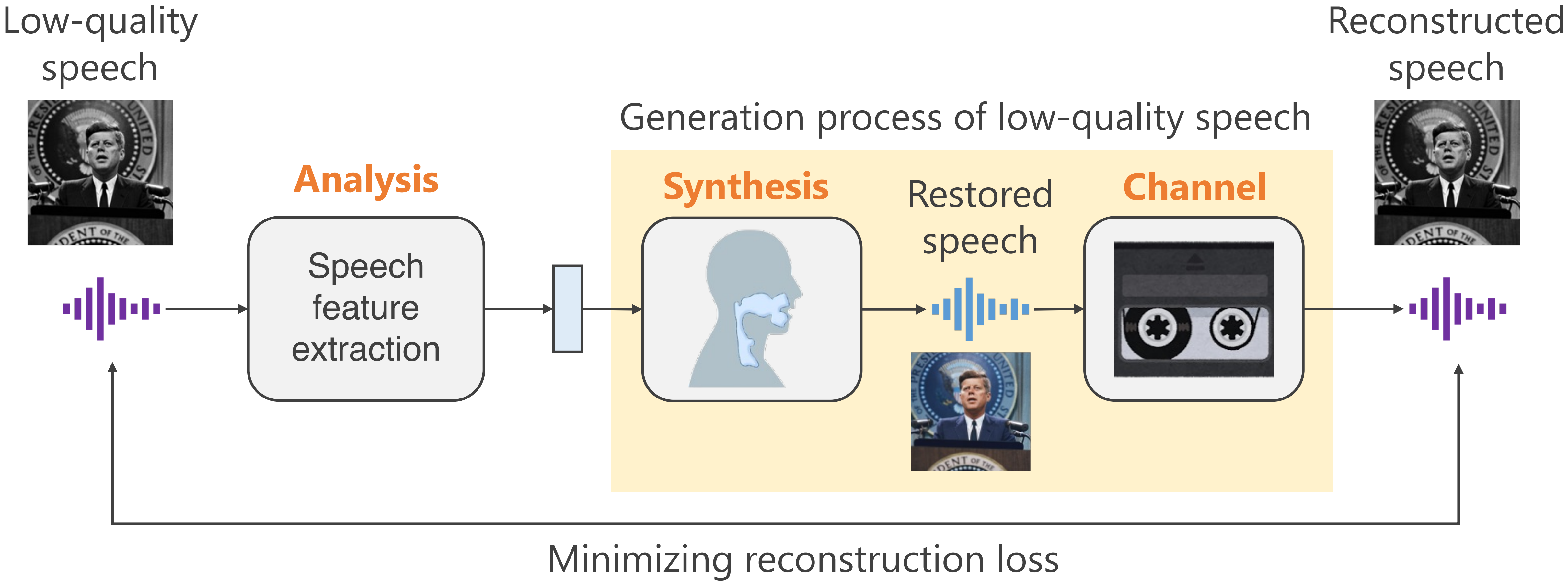
Low-quality old recordings　　　　High-quality restored audio



Our Approach: Self-Supervised Speech Restoration

Learning speech restoration model **without paired data**.

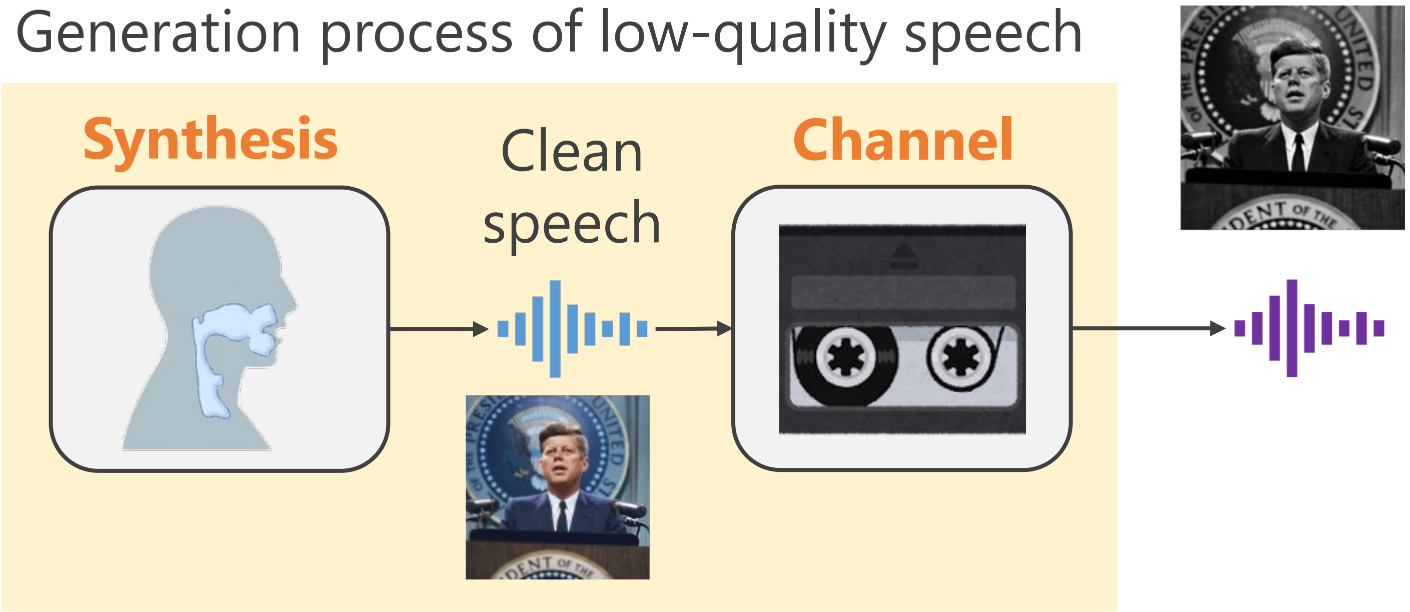
Simulating the generation process of recorded audio.



Our Approach: Self-Supervised Speech Restoration

Learning speech restoration model **without paired data**.

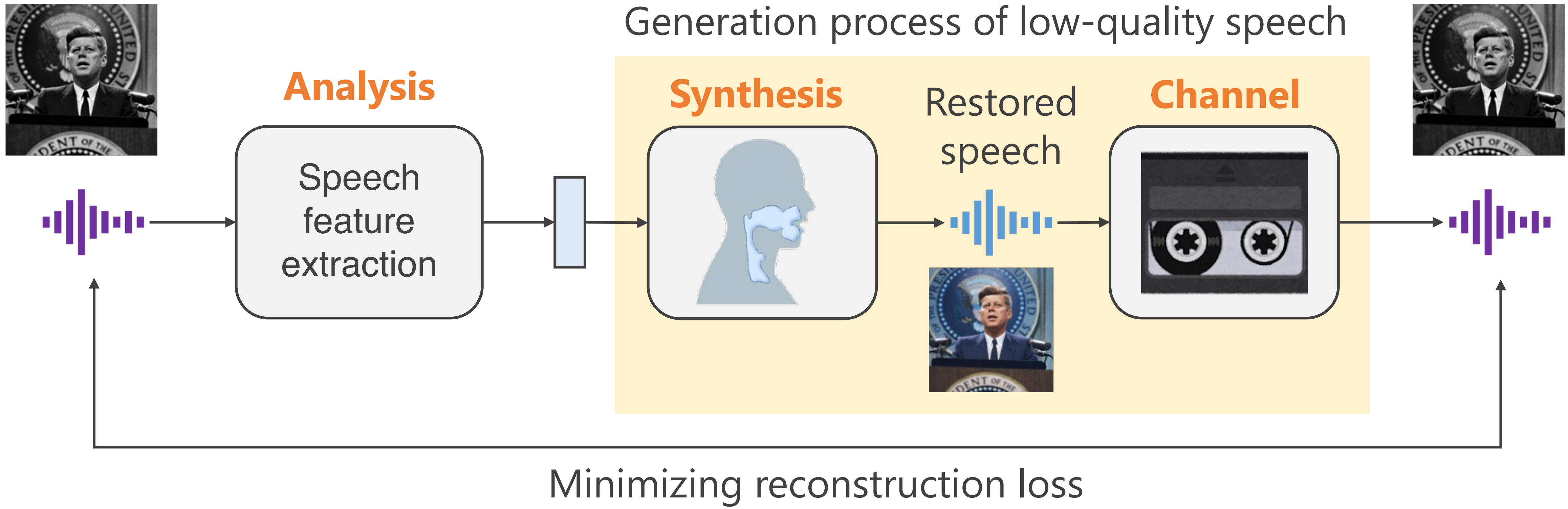
Simulating the generation process of recorded audio.



Our Approach: Self-Supervised Speech Restoration

Learning speech restoration model **without paired data**.

Consisting of **analysis**, **synthesis**, and **channel** modules.

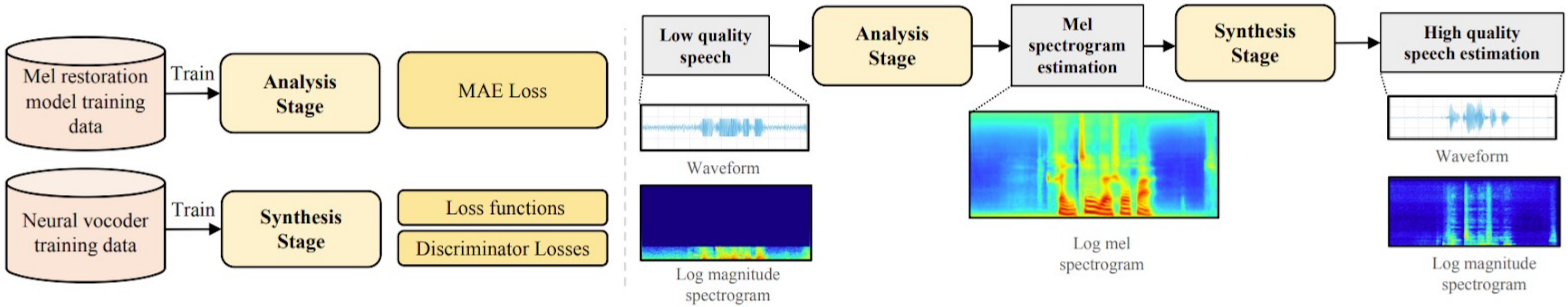


Outline

Supervised learning for speech restoration [Liu+, 2021]

Training **analysis** and **synthesis** modules separately.

Creating artificial **paired** training data.



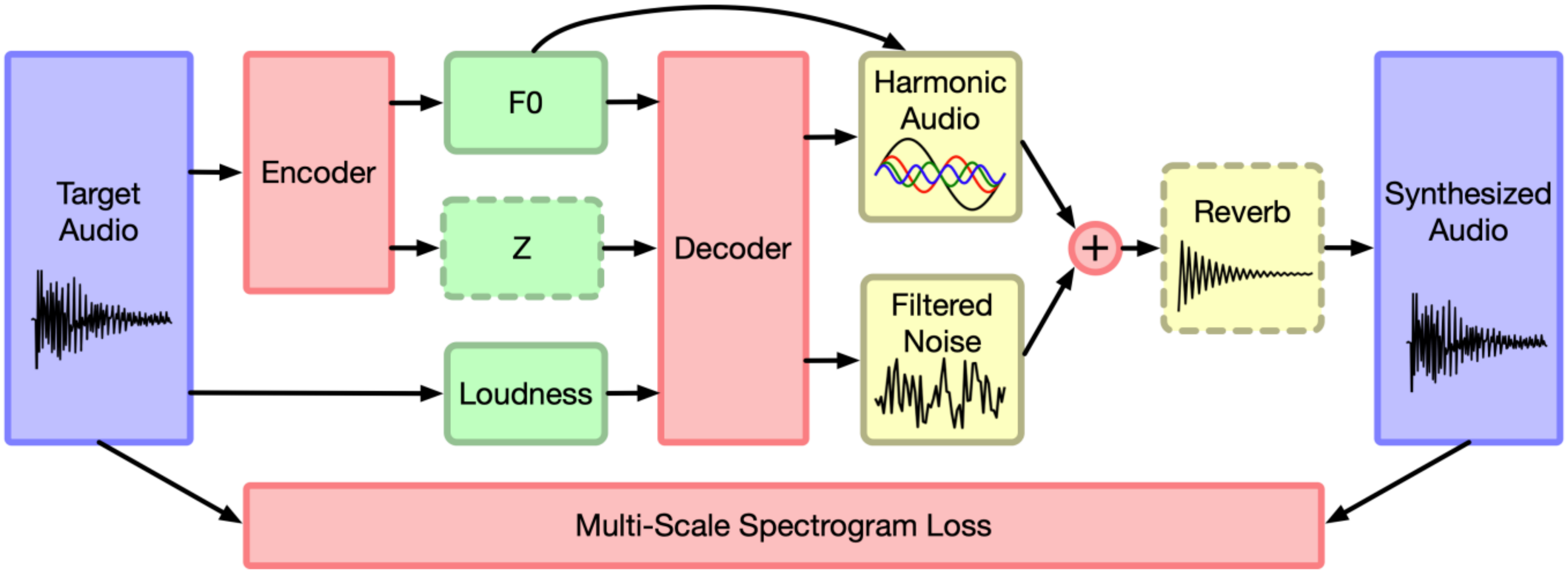
Our approach uses **real data** based on self-supervised learning.

H. Liu et al., "Voicefixer: Toward general speech restoration with neural vocoder," arXiv, vol. abs/2109.13731, 2021

Related Work

Differential digital signal processing (DDSP) autoencoder [Engel+, 2021]

Learning disentangled audio features in a **self-supervised** manner



Our work focuses on **restoration of degraded speech**.

J. Engel et al., "DDSP: Differentiable Digital Signal Processing," Proc. ICLR, 2020.

Outline

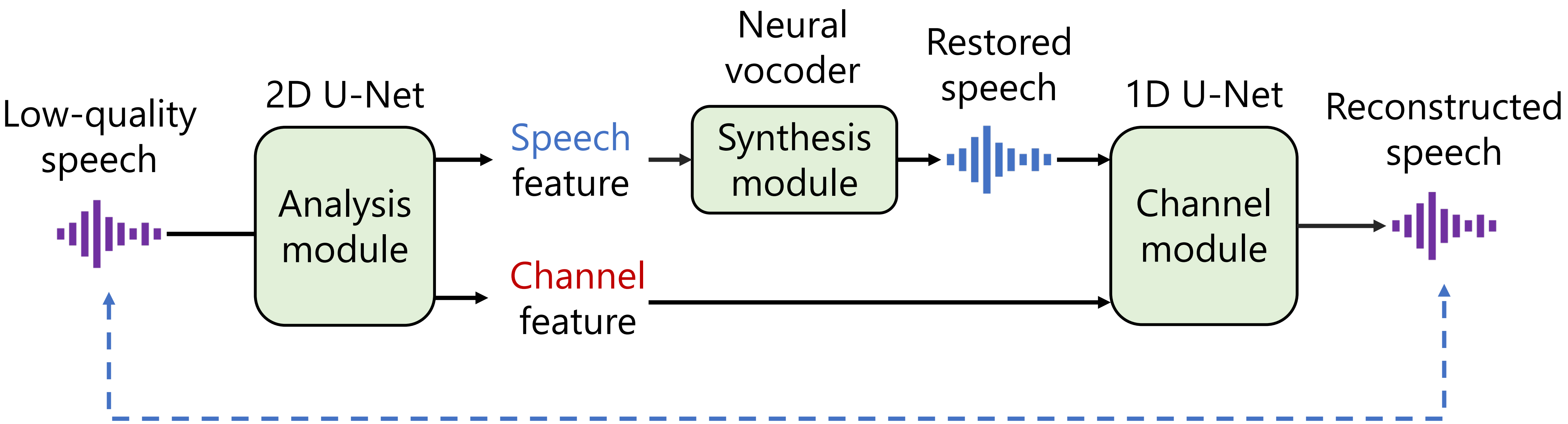
- ❑ Background
- ❑ Related Work
- ❑ Methods
- ❑ Experimental Evaluation
- ❑ Summary

Basic Framework of Proposed Method

Analysis, Synthesis, Channel modules are all composed of neural networks.

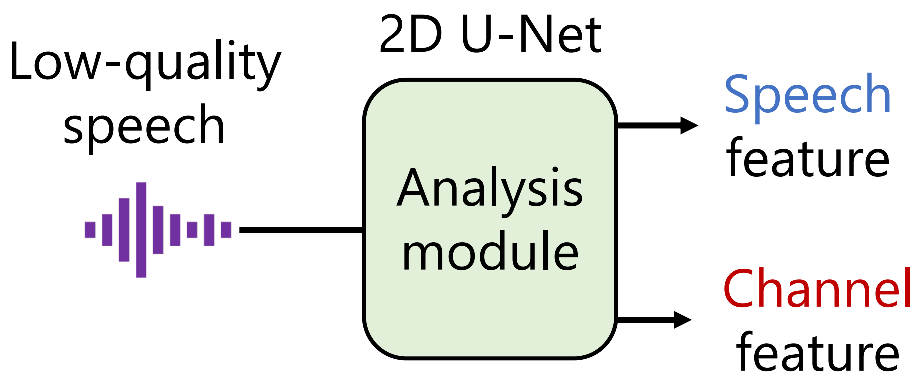
Analysis and **Channel** modules: 2D and 1D U-Net models

Synthesis module: HiFi-GAN [Kong+, 2020]



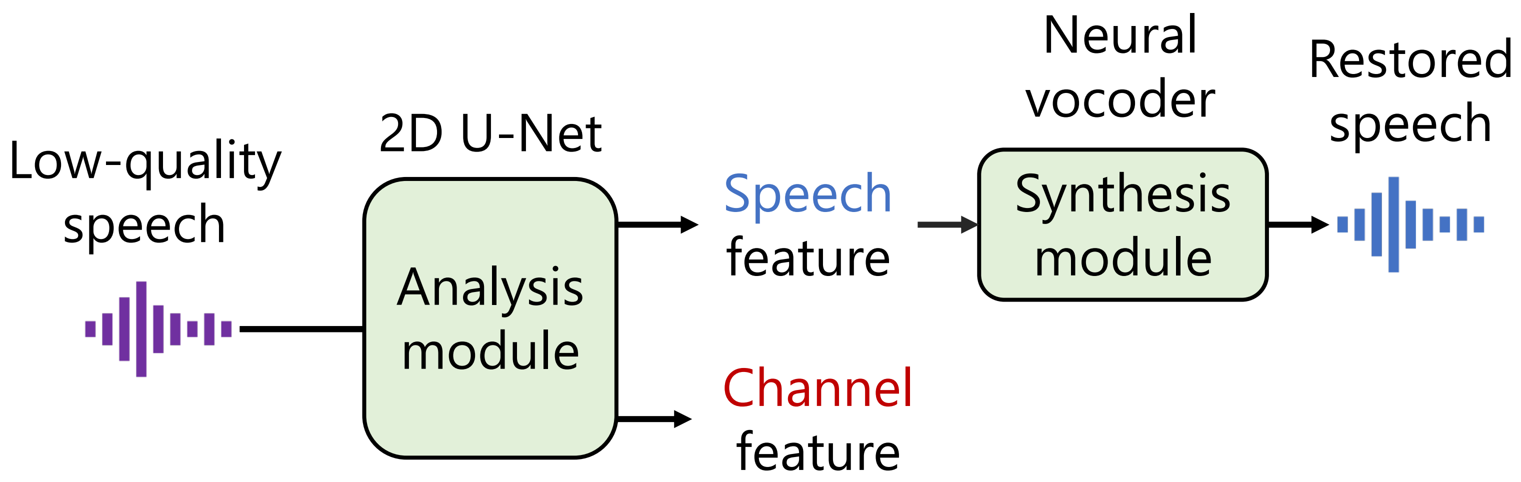
Multi-scale spectral loss $\mathcal{L}_{\text{recons}} = \sum_i \{||\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i||_1 + \alpha|| \log \boldsymbol{s}_i - \log \hat{\boldsymbol{s}}_i||_1\}$

Analysis module estimates speech features and channel features.



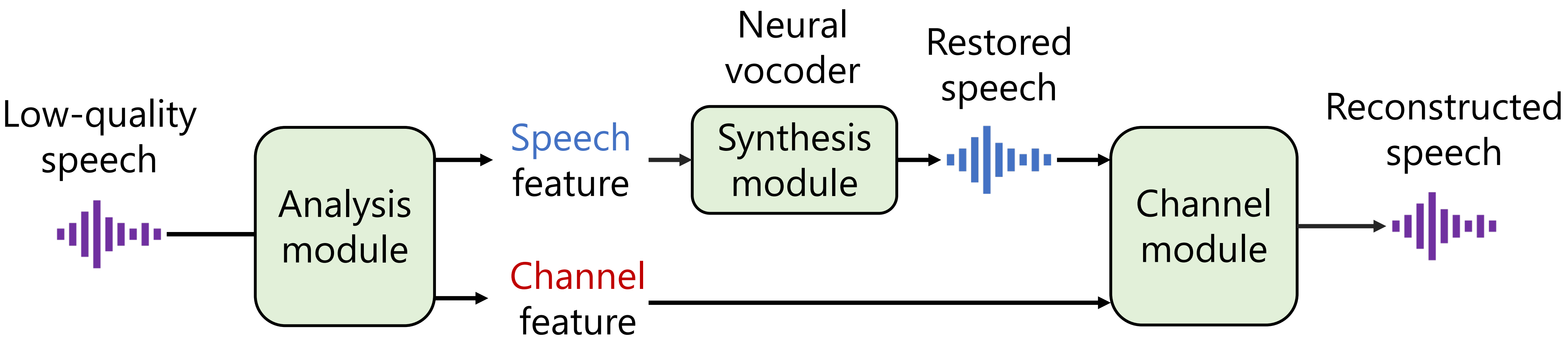
Basic Framework of Proposed Method

Synthesis module synthesizes restored speech from speech features.



Basic Framework of Proposed Method

Channel module adds channel features to restored speech.

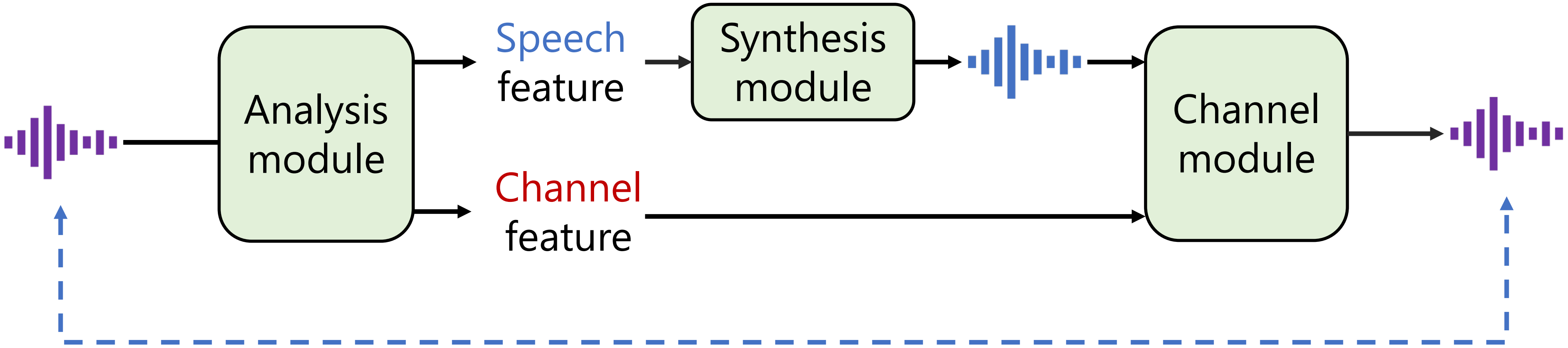


Basic Framework of Proposed Method

Analysis module estimates speech features and channel features.

Synthesis module synthesizes restored speech from speech features.

Channel module adds channel features to restored speech.



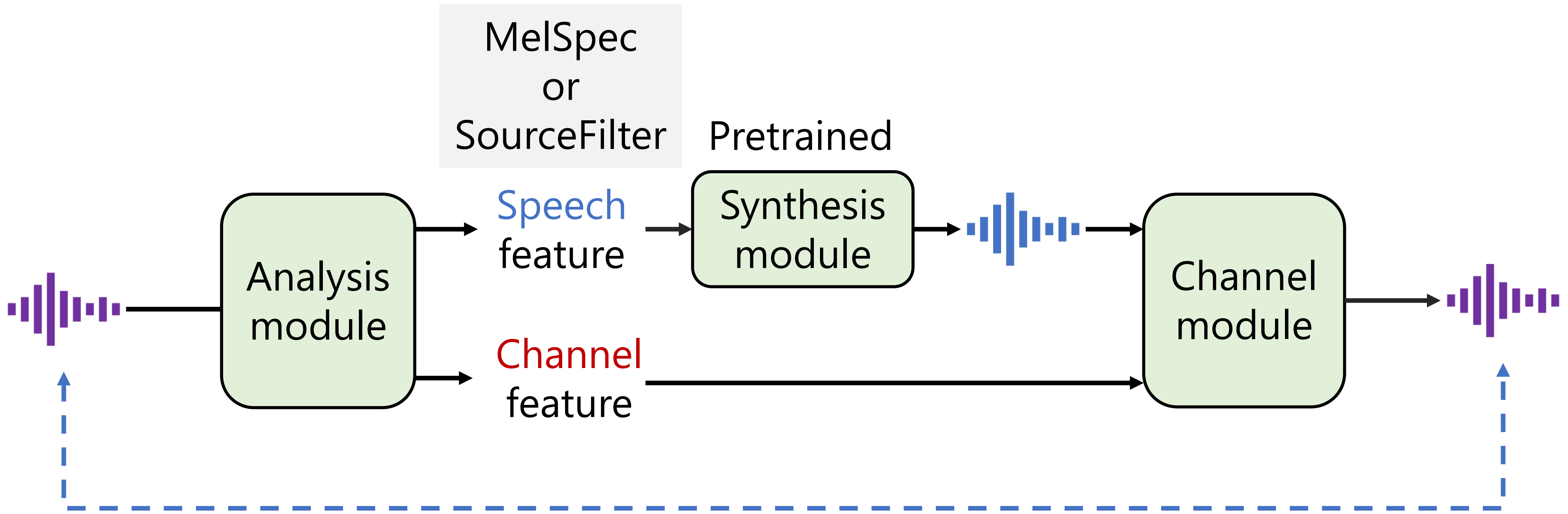
Multi-scale spectral loss $\mathcal{L}_{\text{recons}} = \sum_i \{||\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i||_1 + \alpha|| \log \boldsymbol{s}_i - \log \hat{\boldsymbol{s}}_i||_1\}$

Basic Framework of Proposed Method

MelSpec: Using mel spectrogram to train synthesis module

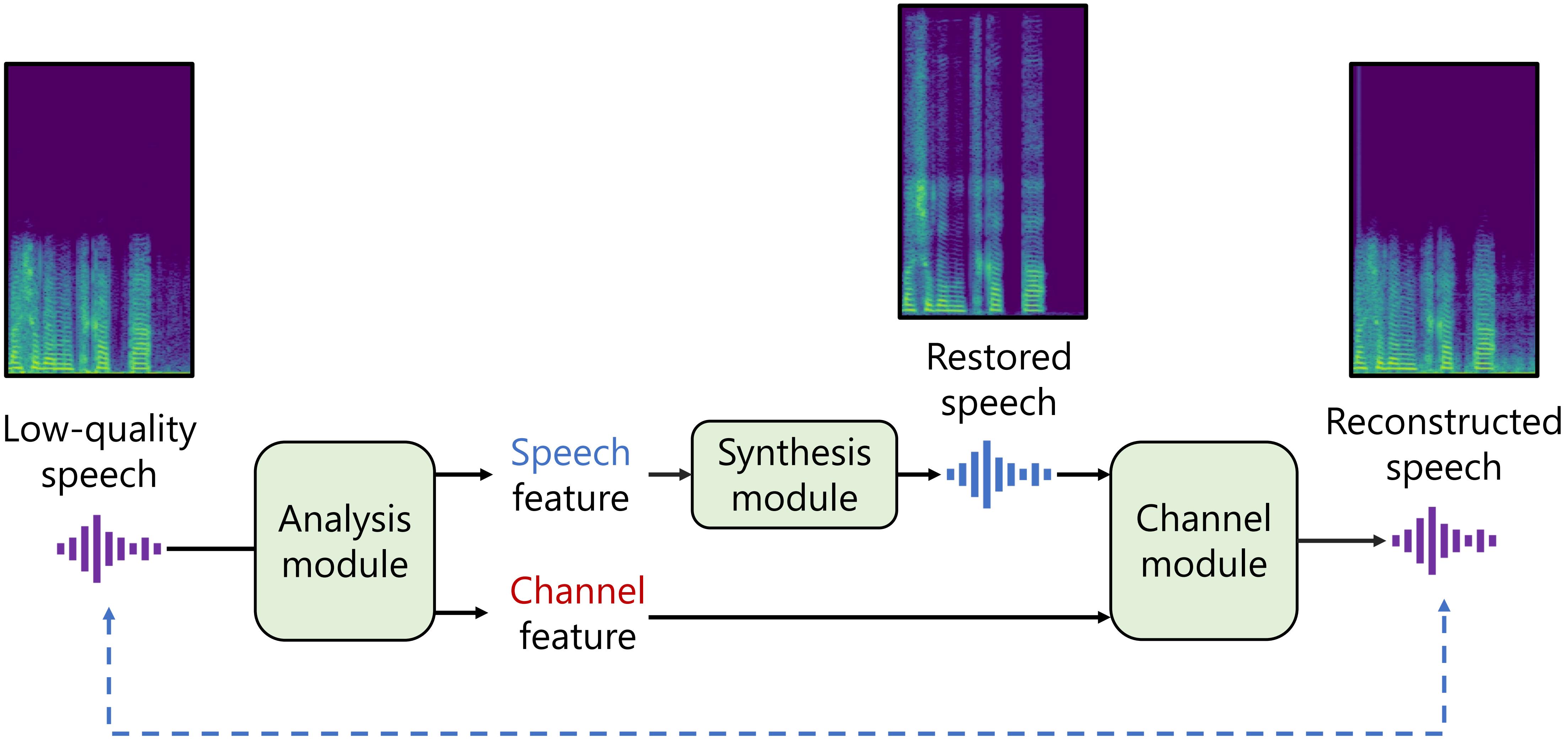
SourceFilter: Using mel cepstrum + F0 to train synthesis module

Only pretraining synthesis module and freezing it.



Multi-scale spectral loss $\mathcal{L}_{\text{recons}} = \sum_i \{||\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i||_1 + \alpha||\log \boldsymbol{s}_i - \log \hat{\boldsymbol{s}}_i||_1\}$

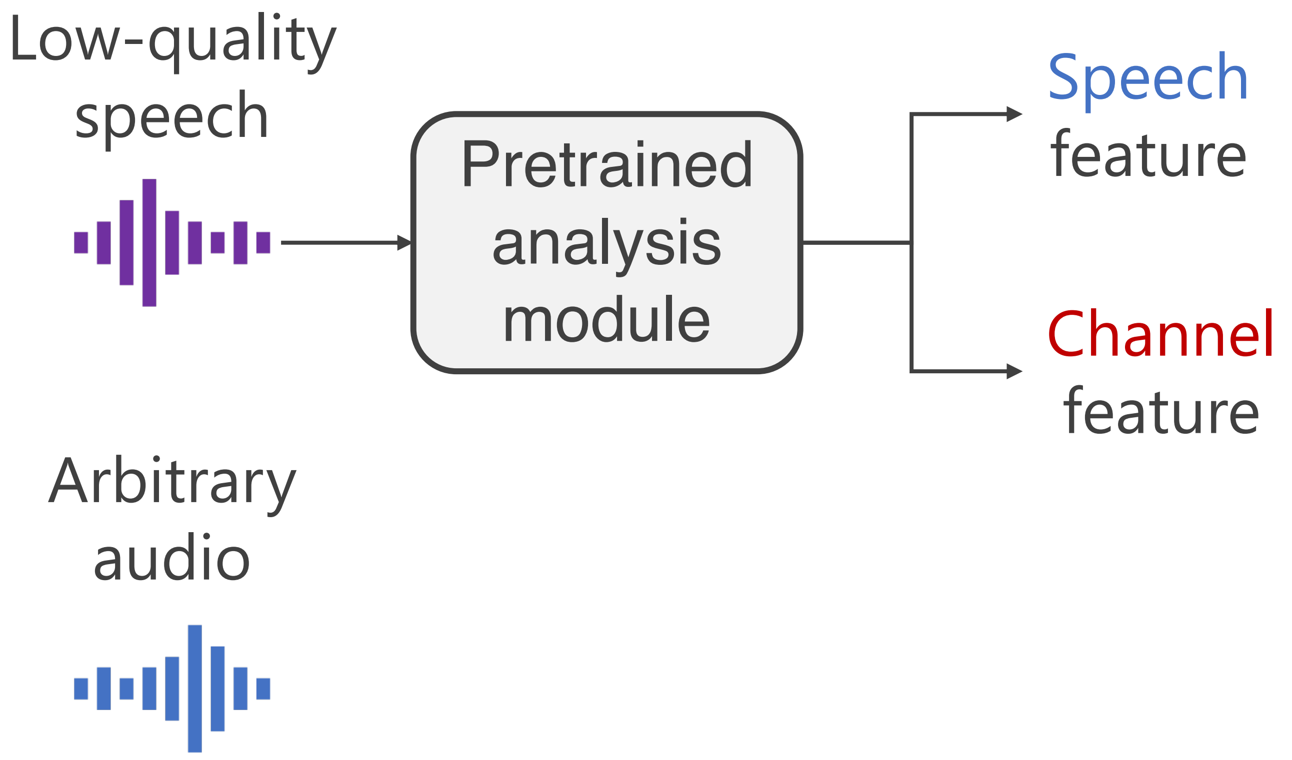
Basic Framework of Proposed Method



Multi-scale spectral loss $\mathcal{L}_{\text{recons}} = \sum_i \{||\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i||_1 + \alpha||\log \boldsymbol{s}_i - \log \hat{\boldsymbol{s}}_i||_1\}$

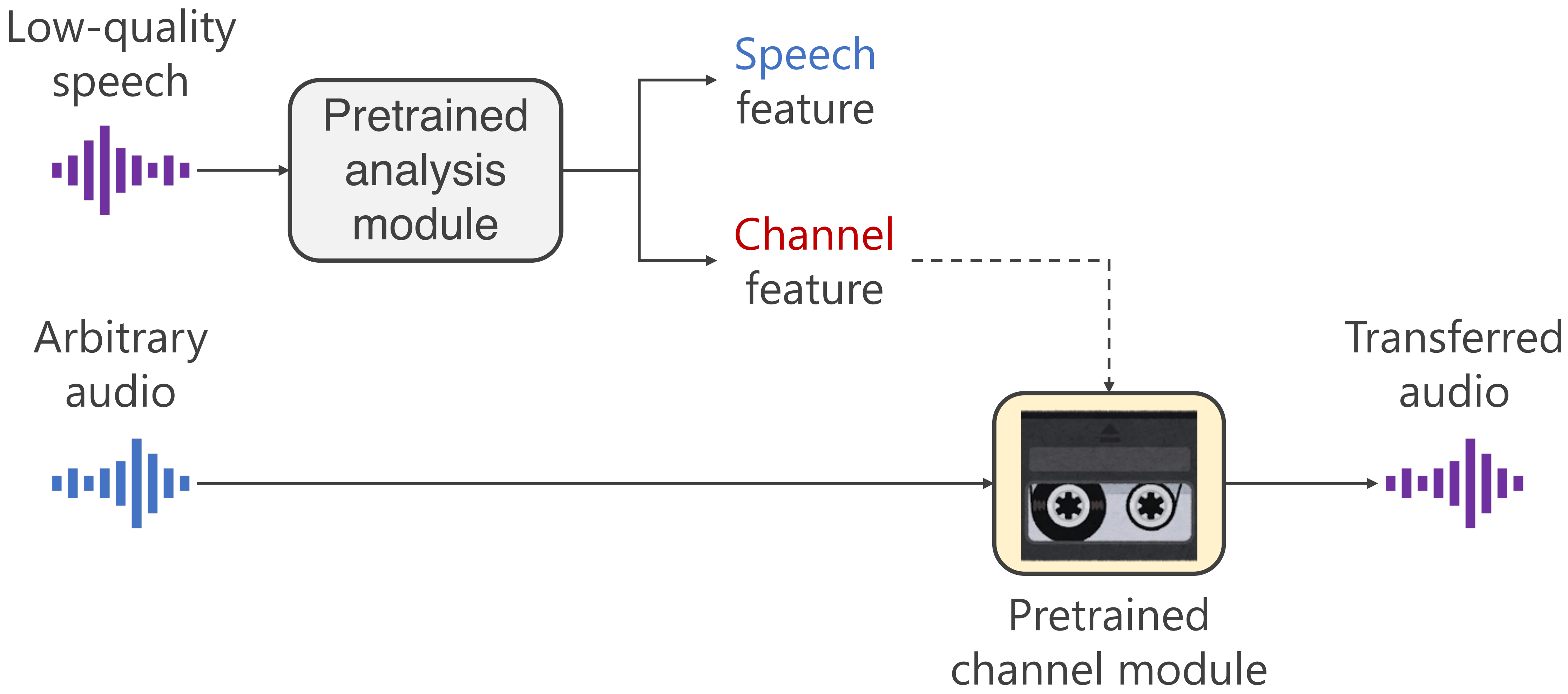
Audio Effect Transfer with Proposed Method

Proposed method **works as Audio effector** to extract and add channel features.

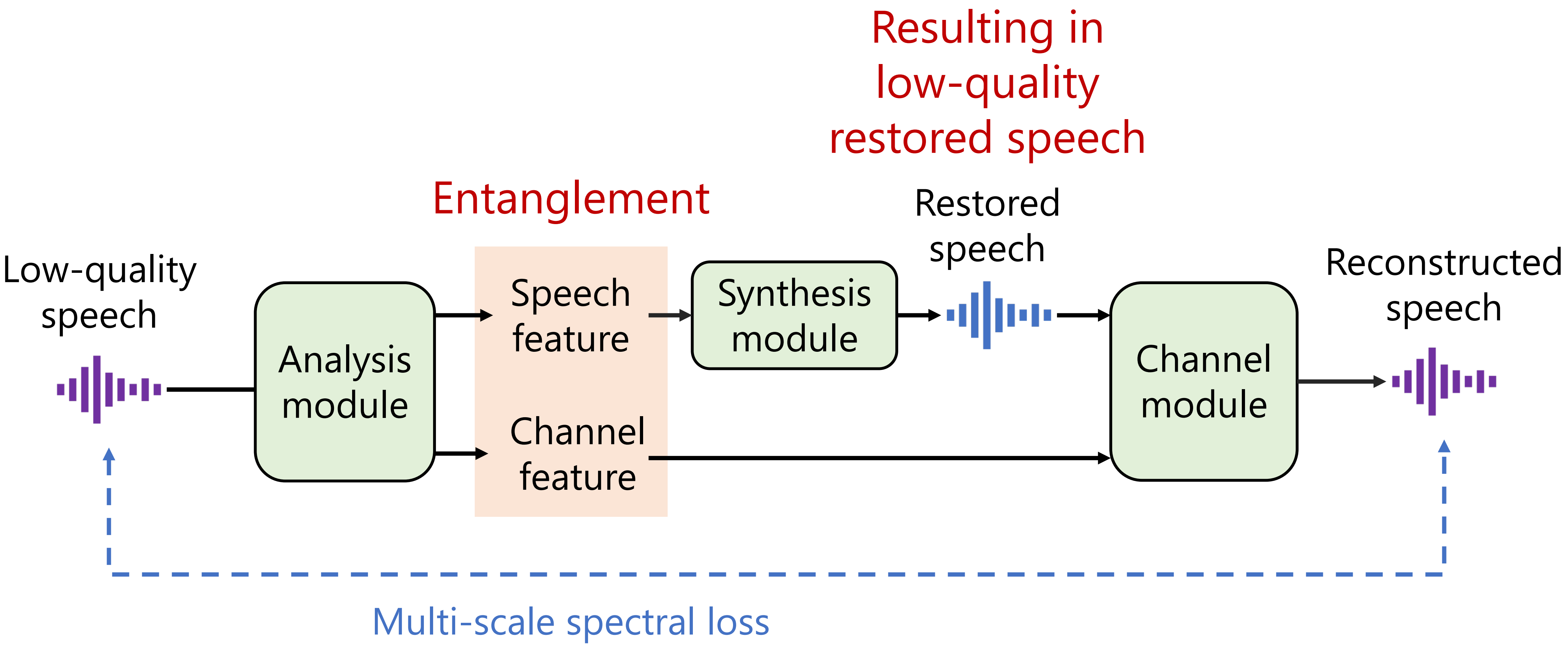


Audio Effect Transfer with Proposed Method

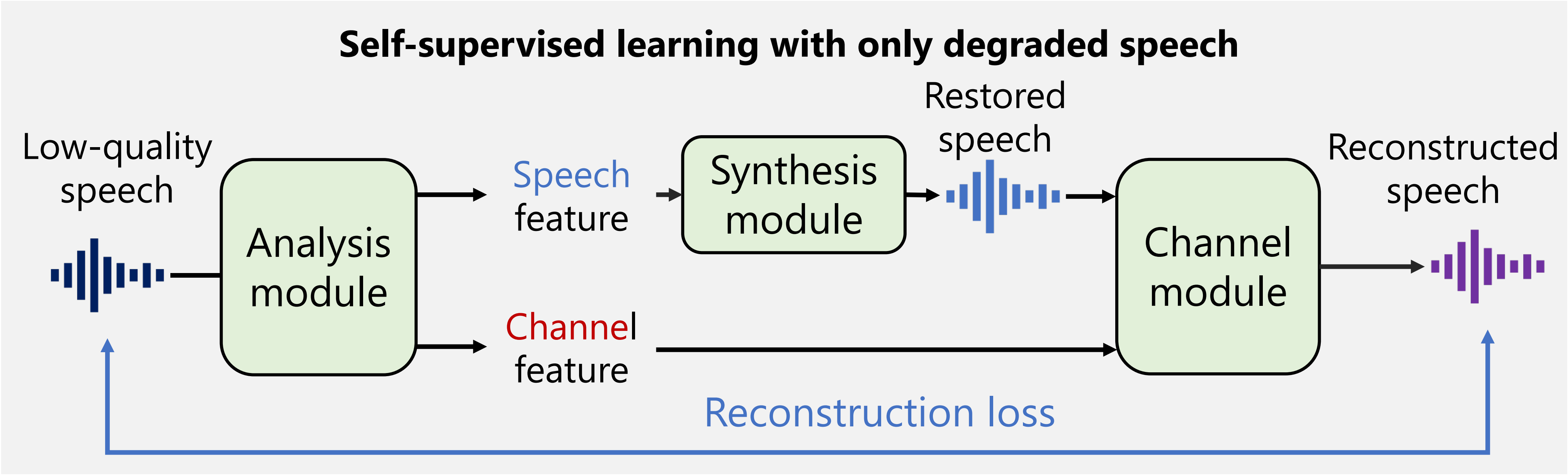
Proposed method **works as Audio effector** to extract and add channel features.



Problem on Basic Framework

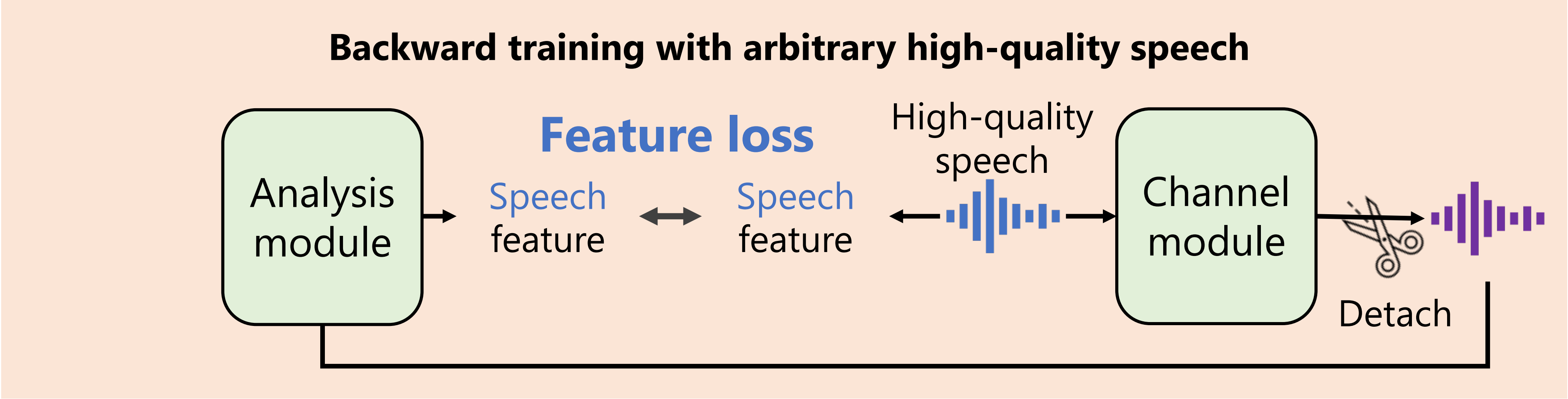


Dual Learning for Stable Training



τ_{forward}

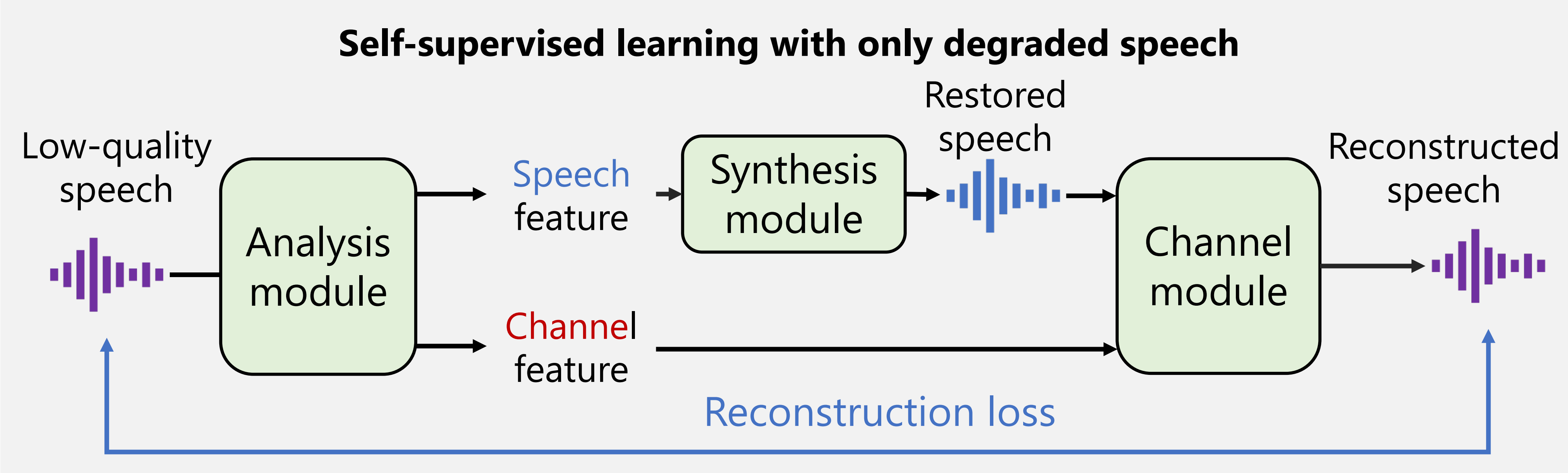
Learning channel module to adapt to input speech



τ_{backward}

Learning analysis module to output clean speech features

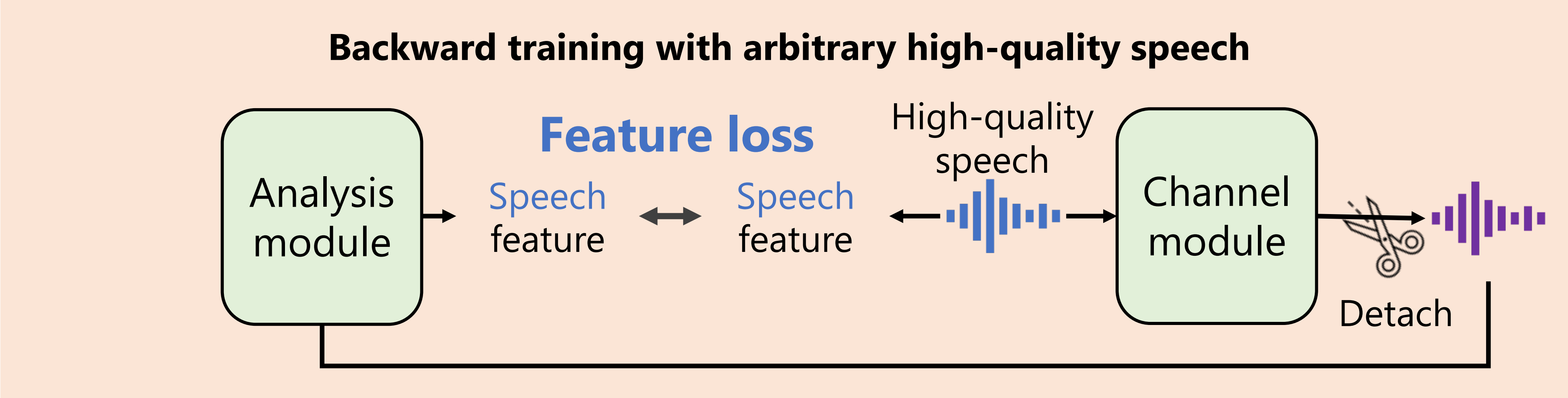
Dual Learning for Stable Training



τ_{forward}

Learning channel module to adapt to input speech

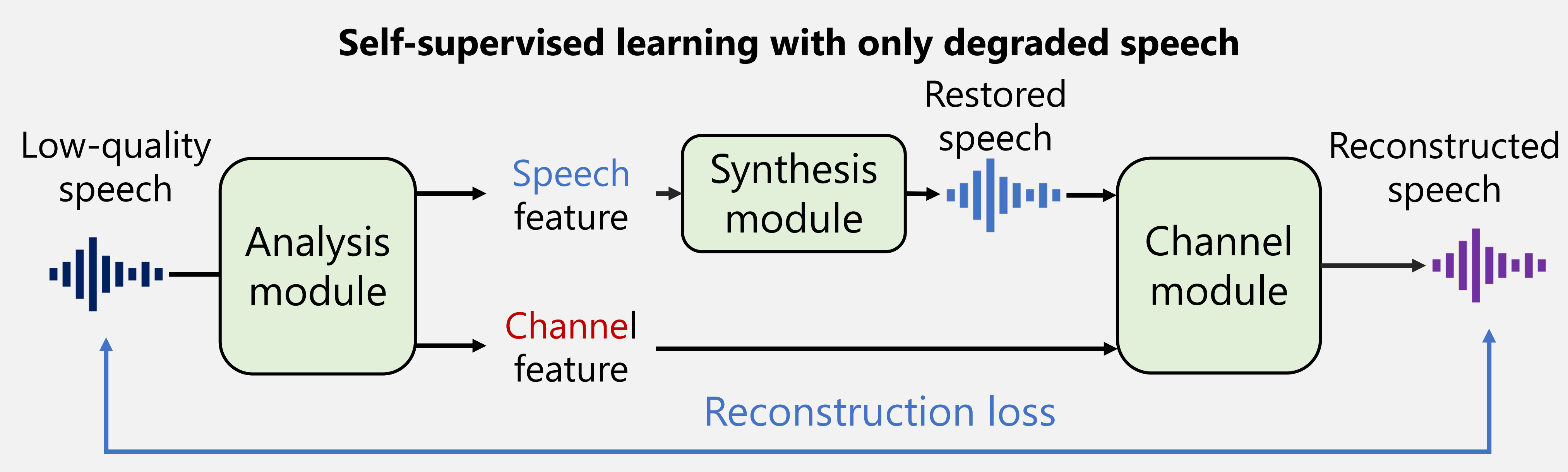
Dual Learning for Stable Training



$\mathcal{T}_{\text{backward}}$

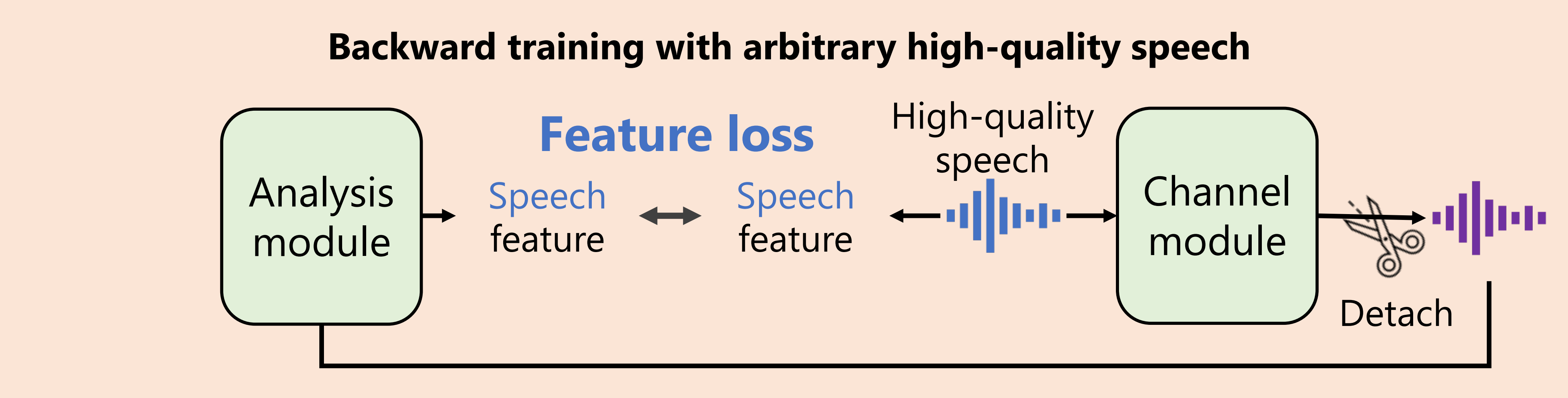
Learning analysis module to output clean speech features

Dual Learning for Stable Training



τ_{forward}

Learning channel module to adapt to input speech



τ_{backward}

Learning analysis module to output clean speech features

Applying Proposed Method to Real Data

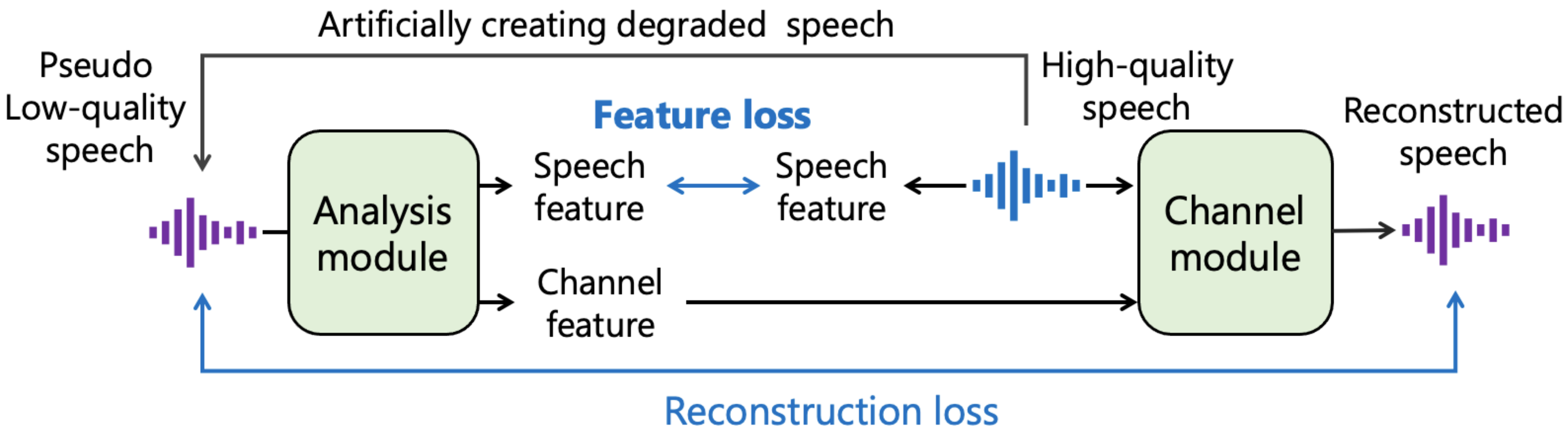
We cannot get so much data from a single historical audio material.

Hard to learn modules with low-resource data (< 1 hour).

- ➢ Introducing supervised pretraining to tackle data scarcity.

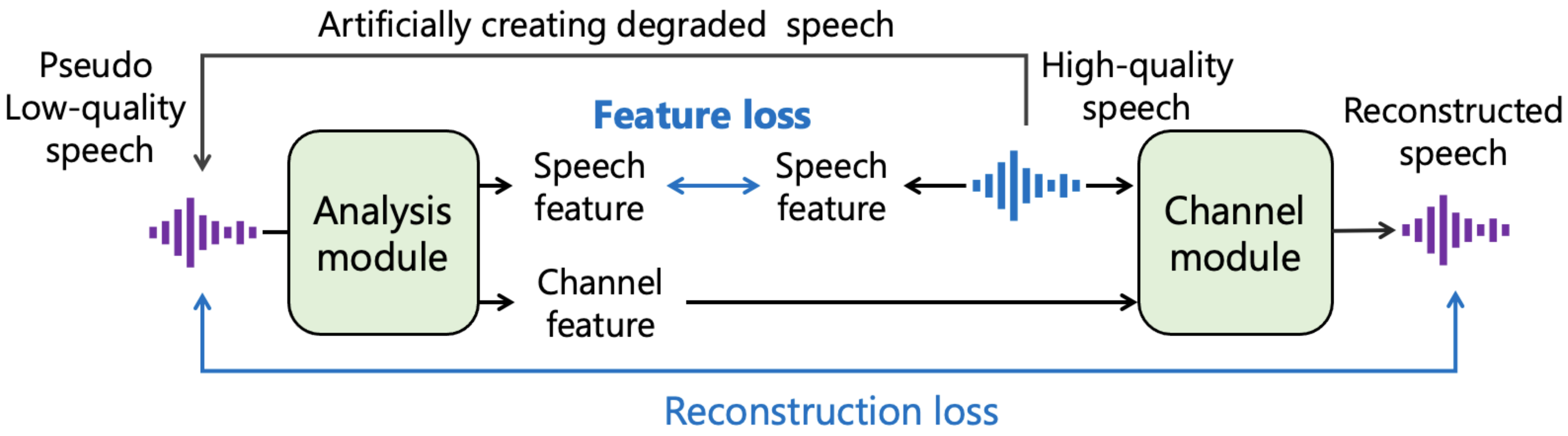
Supervised pretraining for Low-Resource Setting

Supervised pretraining with pseudo low-quality speech data

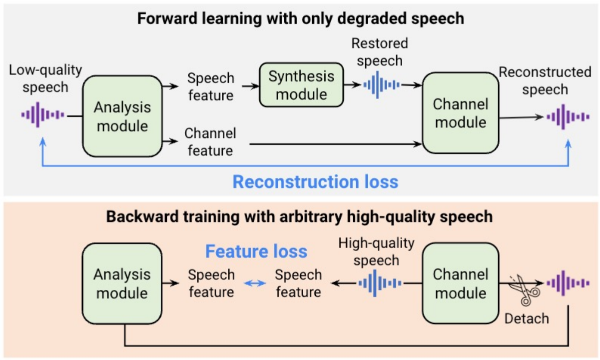


Supervised pretraining for Low-Resource Setting

Supervised pretraining with pseudo low-quality speech data



Self-supervised learning with real data



Outline

Experimental Evaluation

Compared our method and previous supervised method [Liu+, 2021].

1) **Simulated datasets** based on high-quality speech corpus [Takamichi+, 2021]
   Applied four types of distortions to **6-hour** single-speaker data
   a) Band-limited
   b) Clipped
   c) Quantized & Resampled
   d) Overdrive

Experimental Evaluation

Compared our method and previous supervised method [Liu+, 2021].

1) **Simulated datasets** based on high-quality speech corpus [Takamichi+, 2021]
Applied four types of distortions to **6-hour** single-speaker data
   a) Band-limited
   b) Clipped
   c) Quantized & Resampled
   d) Overdrive

2) **Real historical audio material** recorded on an analog tape recorder
Around **20 minutes**' multi-speaker data recorded in 1960s – 1970s

Evaluation Results with Simulated Data

Mean opinion score (MOS) test of speech quality with 40 listeners in each case

	(a) Band-limited		(b) Clipped		(c) Quantized & Resampled		(d) Overdrive	
	MOS	Sample	MOS	Sample	MOS	Sample	MOS	Sample
Ground-truth	4.51		4.58		4.67		4.65	
Input	2.38		2.45		1.73		1.54	
Supervised [Liu+, 2021]	3.74		3.01		2.80		2.00	
Proposed (MelSpec)	**4.20**		**3.49**		**3.27**		**2.68**	
Proposed (SourceFilter)	3.46		2.49		2.66		2.58	

Proposed method achieved significantly higher MOS than previous supervised method.

Evaluation Results with Real Data

Evaluated proposed method with **real historical audio** (around 20 min).

	MOS	Sample
Input	2.98	
Supervised [Liu+, 2021]	2.80	
Proposed (MelSpec)	2.96	
Proposed + pretraining (MelSpec)	**3.06**	

Evaluation Results with Real Data

Evaluated proposed method with **real historical audio** (around 20 min).

Statistical significance
(p-value < 0.05)
in side-by-side test

	MOS	Sample
Input	2.98	
Supervised [Liu+, 2021]	2.80	
Proposed (MelSpec)	2.96	
Proposed + pretraining (MelSpec)	**3.06**	

Effectiveness
for real data

Still needs
pretraining
for real data

Evaluation Results of Audio Effect Transfer

Similarity MOS (SMOS) test for **similarity of audio characteristics**.

- **Source**: Original high-quality audio samples
- **Mean spec. diff**: Applying differential spectrum to original audio samples
- **Proposed**: Performing audio effect transfer with our method

Evaluation Results of Audio Effect Transfer

Similarity MOS (SMOS) test for **similarity of audio characteristics**.

- **Source**: Original high-quality audio samples
- **Mean spec. diff**: Applying differential spectrum to original audio samples
- **Proposed**: Performing audio effect transfer with our method

	Simulated (Quantized & Resampled)	Real
Target	3.98	2.99
Source	1.16	1.30
Mean spec. diff	1.68	-
Proposed	**3.44**	**2.12**

Thank you for your attention!!!

Self-supervised speech restoration without paired data

Confirmed effectiveness with **real data** but need more data.

Code

More audio samples