ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1] , Tal Remez[1] , Bowen Shi[1,3] ,
Jacob Donley[2], Yossi Adi[1,4]

1 FAIR, Meta AI Research
2 Meta Reality Labs Research
3 Toyota Technological Institute at Chicago
4 The Hebrew University of Jerusalem

To learn more:

Meta AI

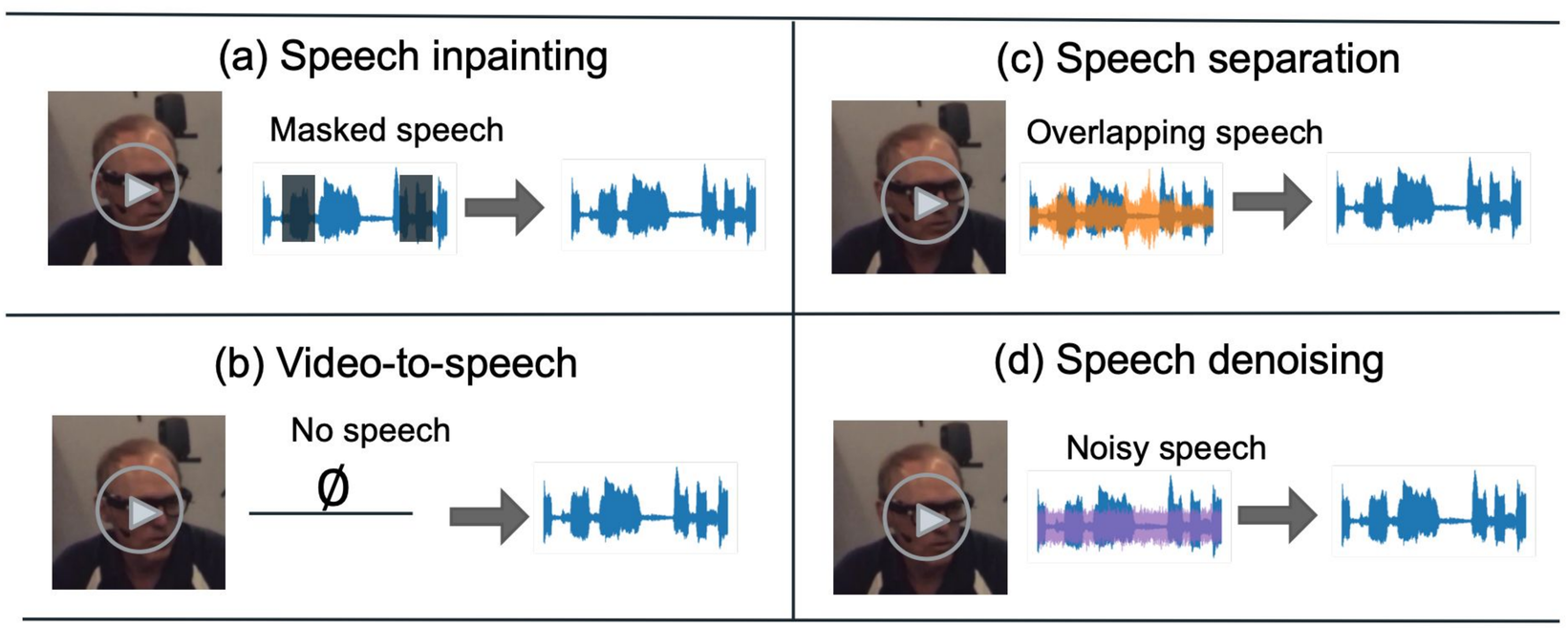
ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

- Improve speech signals given a visual inputs.
- Prior work studies each type of auditory distortion separately.
- Prior work aims at reconstructing the reference rather than preserve the content.
- A lack of “Generalized Speech Enhancement” methods.



To learn more:

Meta AI

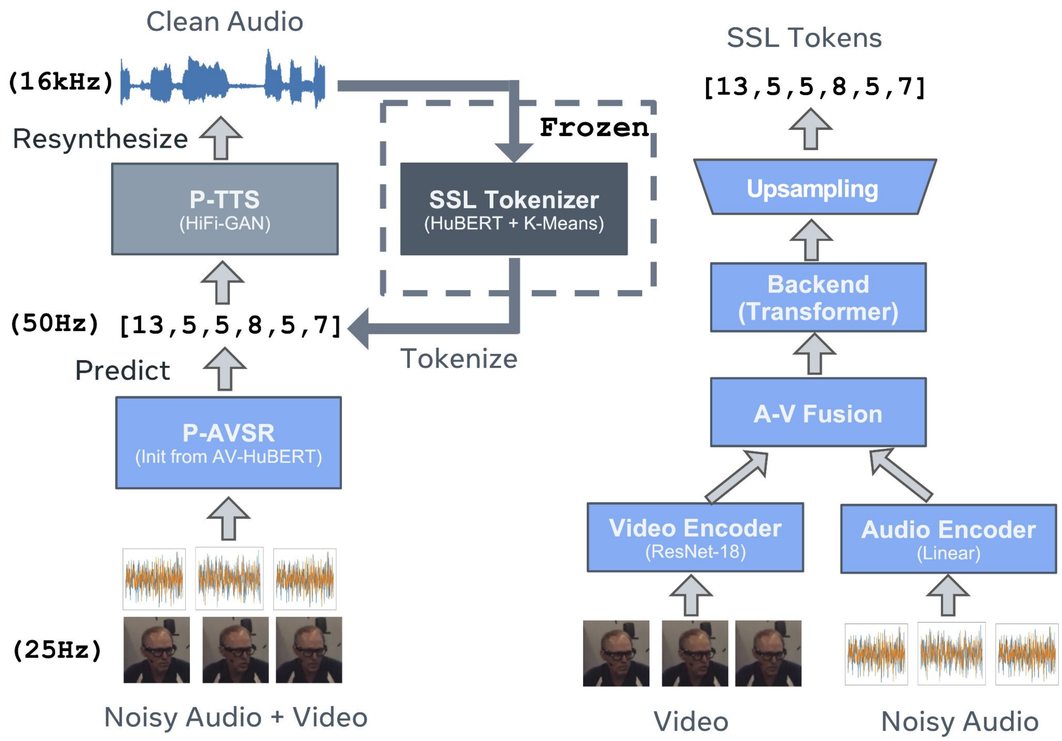
ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Method

- Focuses on intelligibility, quality, and video synchronization.
- Casts the problem as audio-visual speech resynthesis, which is composed of two steps:
  - Pseudo audio-visual speech recognition (P-AVSR)
  - Pseudo text-to-speech synthesis (P-TTS).
- P-AVSR and P-TTS are connected by discrete units.
- Utilizes a self-supervised audio-visual speech model to initialize P-AVSR.



To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

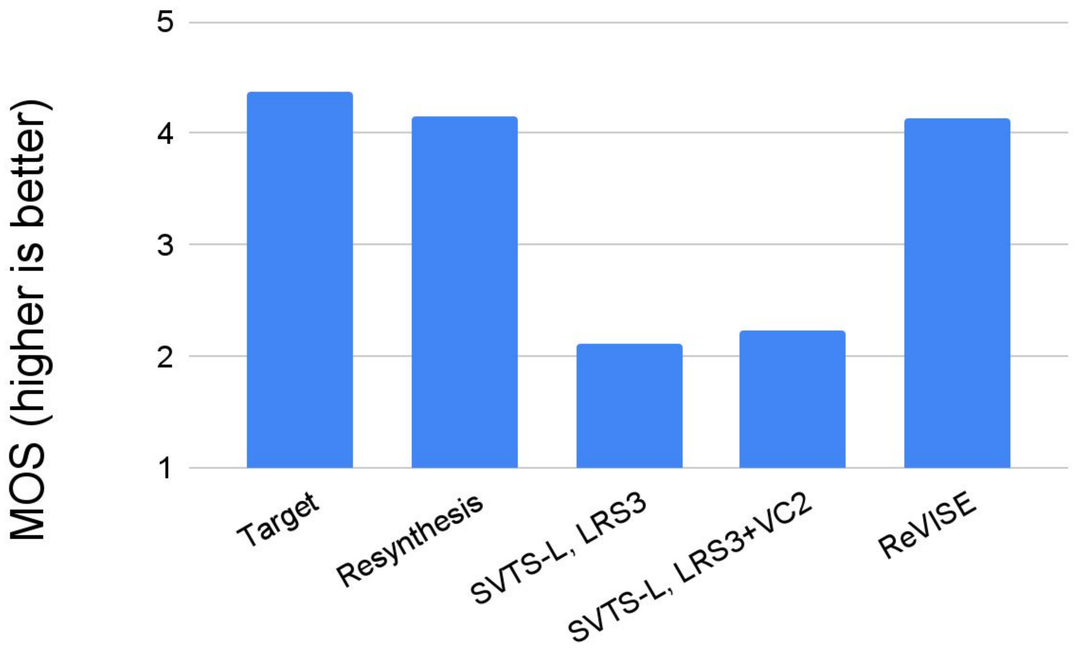
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

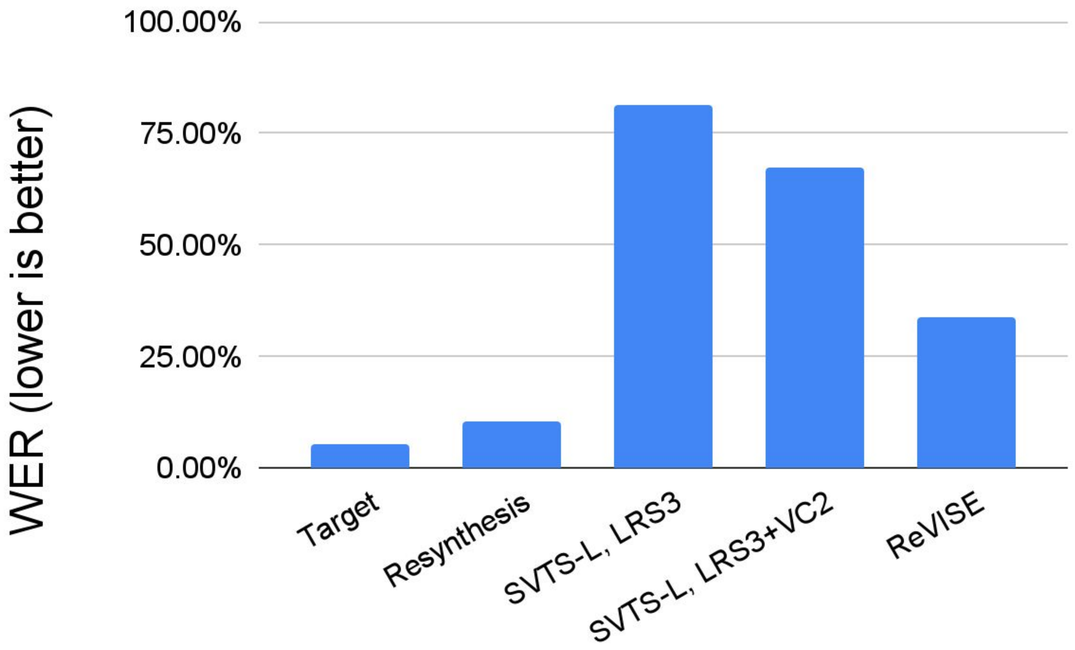
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

LRS3 (Lip-reading Sentences)

A clean dataset based on TED talk videos. It contains **433 hours** of audio-visual speech data and their corresponding text transcripts.

Video-to-speech





To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

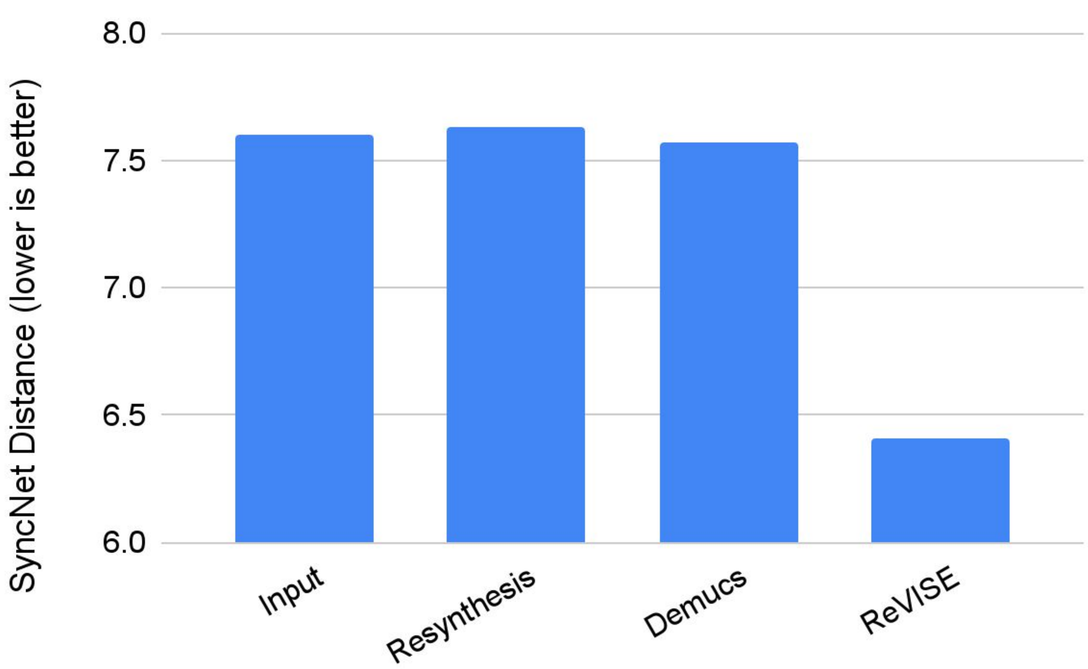
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

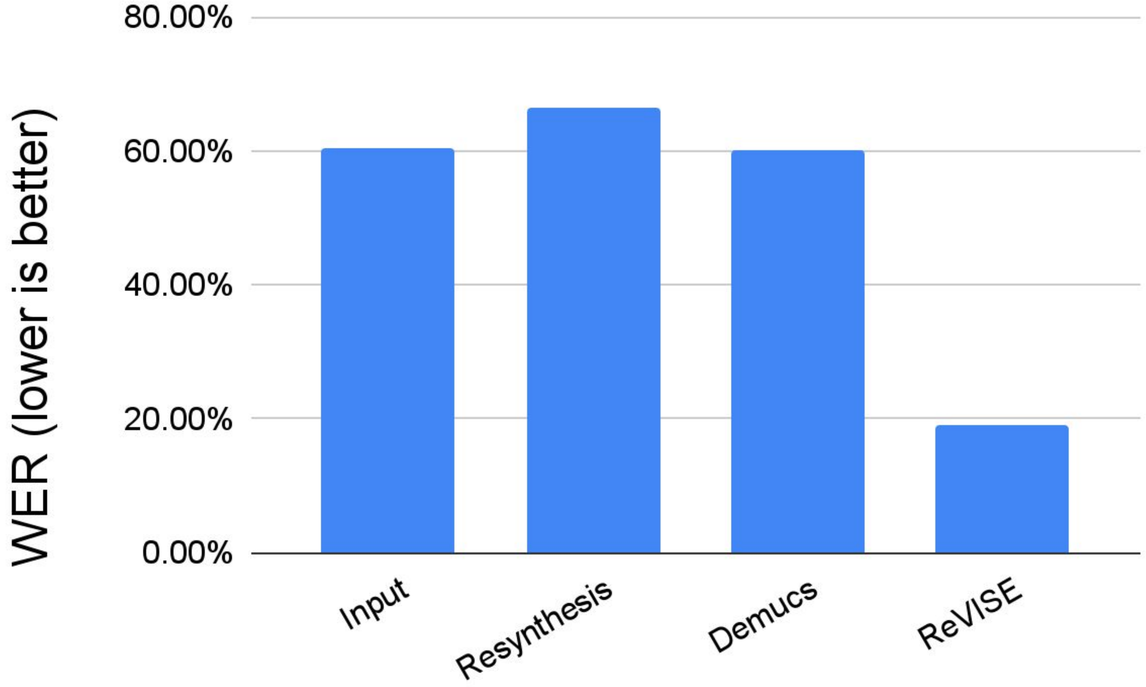
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

LRS3 (Lip-reading Sentences)

A clean dataset based on TED talk videos. It contains **433 hours** of audio-visual speech data and their corresponding text transcripts.

Inpainting





To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

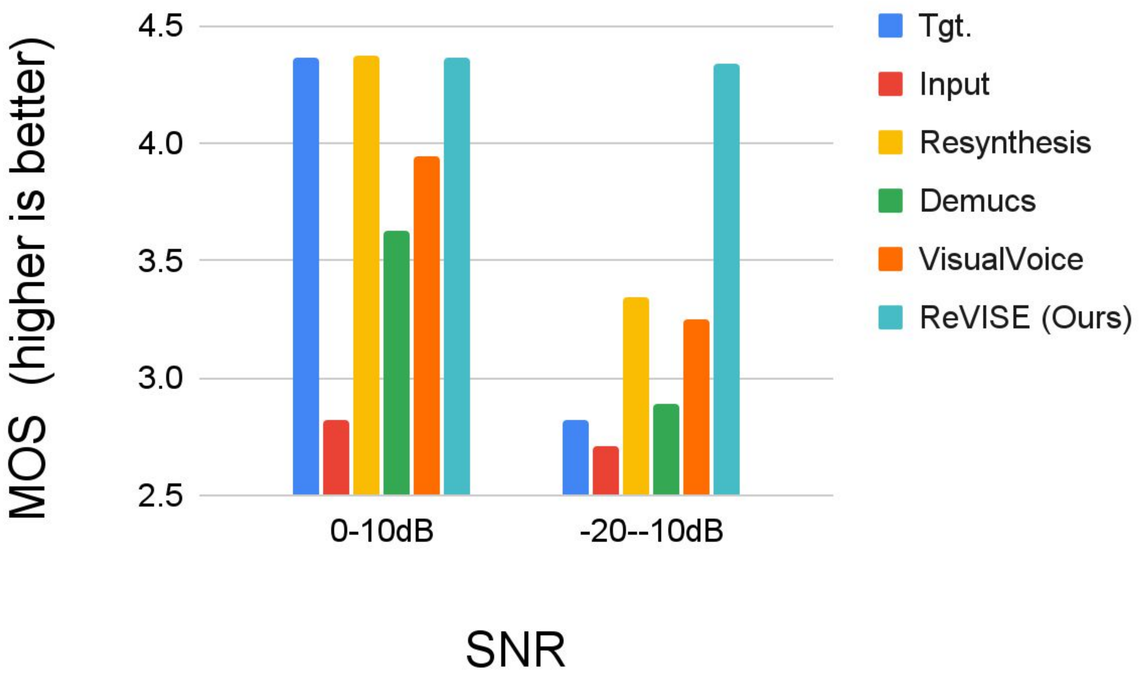
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

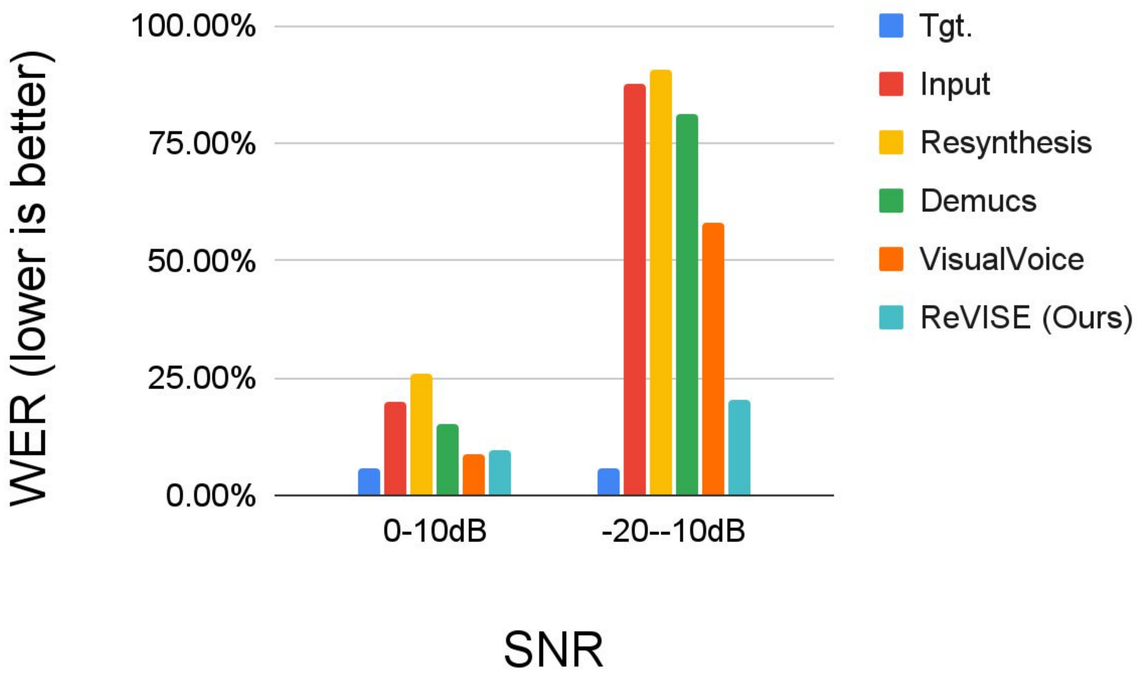
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

LRS3 (Lip-reading Sentences)

A clean dataset based on TED talk videos. It contains **433 hours** of audio-visual speech data and their corresponding text transcripts.

Inpainting





To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

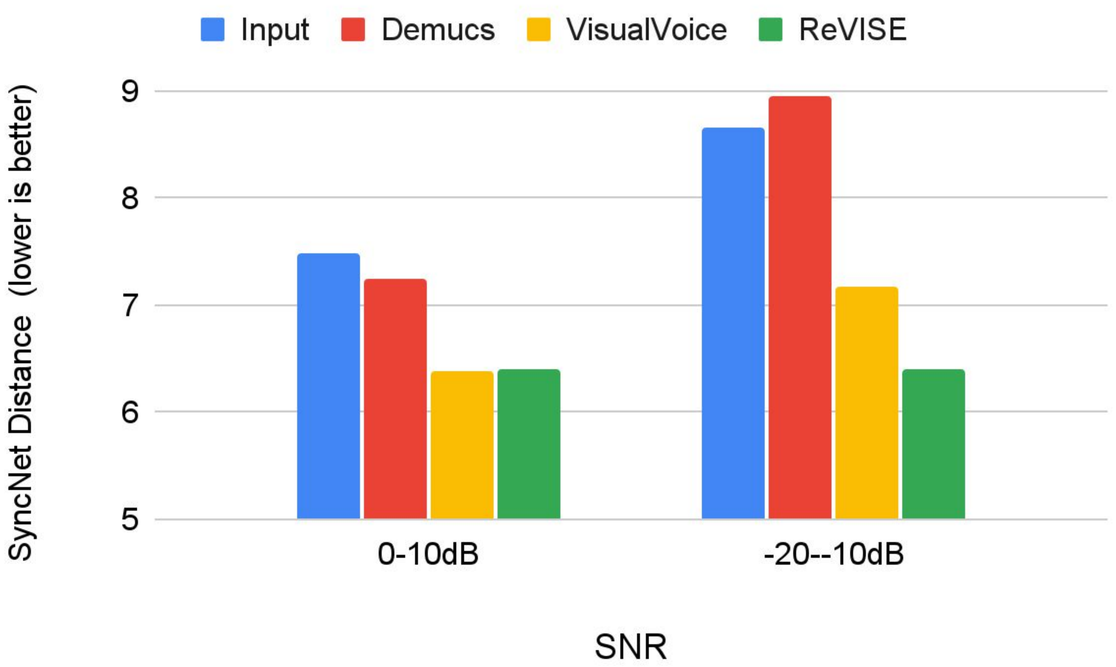
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

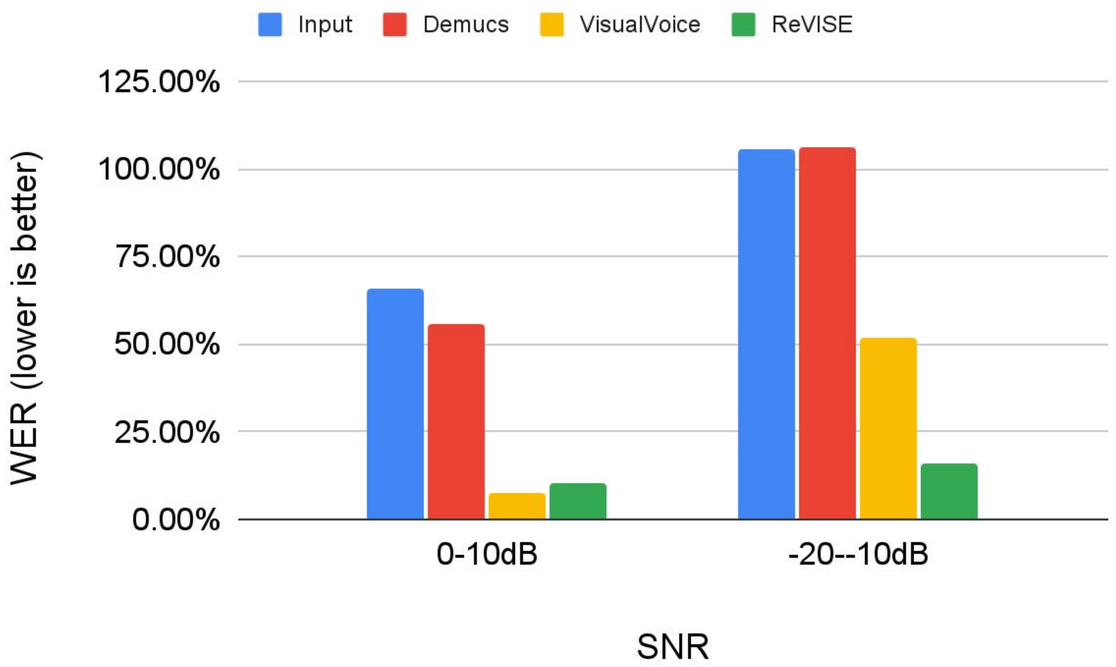
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

LRS3 (Lip-reading Sentences)

A clean dataset based on TED talk videos. It contains **433 hours** of audio-visual speech data and their corresponding text transcripts.

Separation





To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

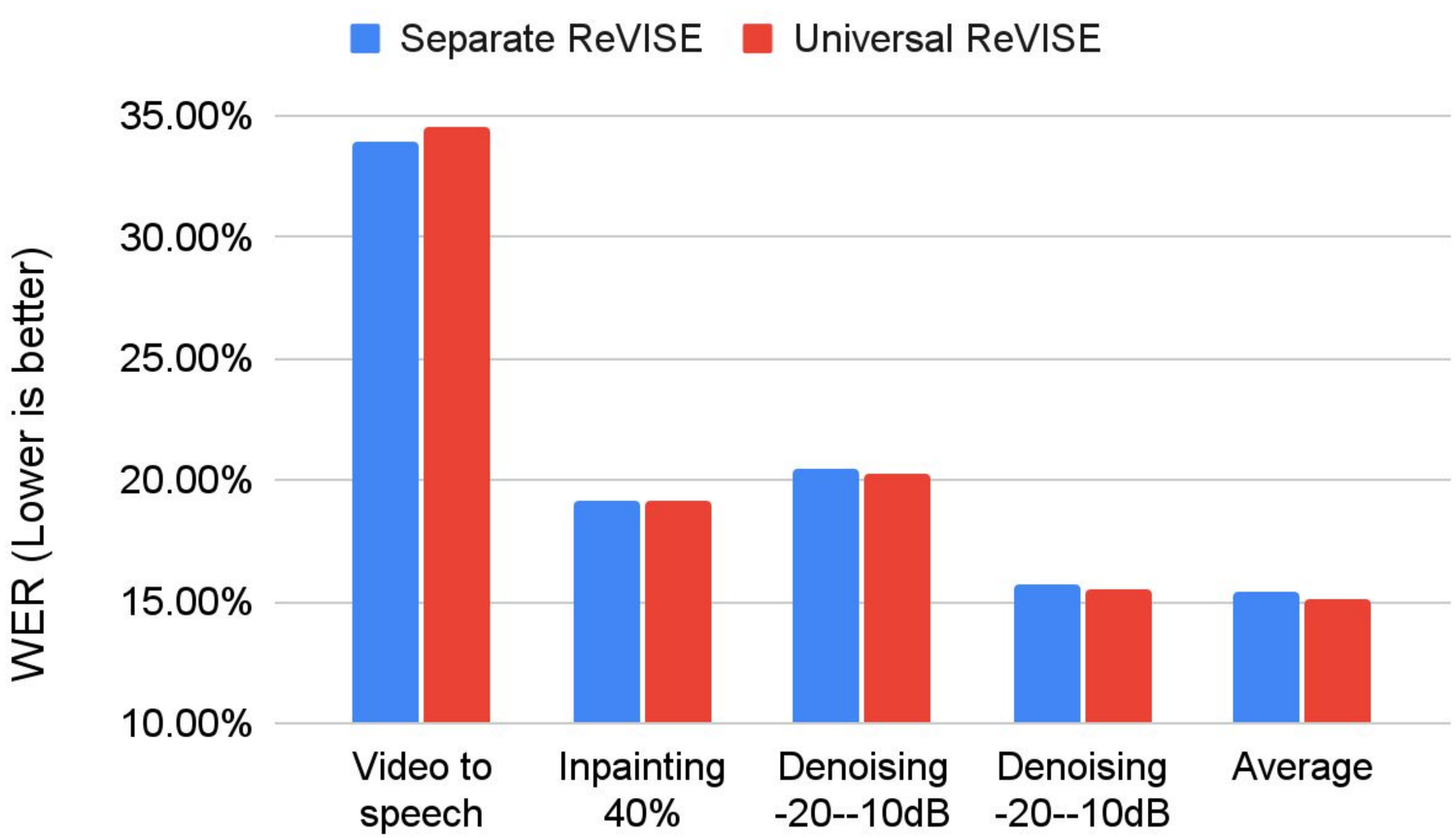
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

LRS3 (Lip-reading Sentences)

A clean dataset based on TED talk videos. It contains **433 hours** of audio-visual speech data and their corresponding text transcripts.

Universal Model



To learn more:

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

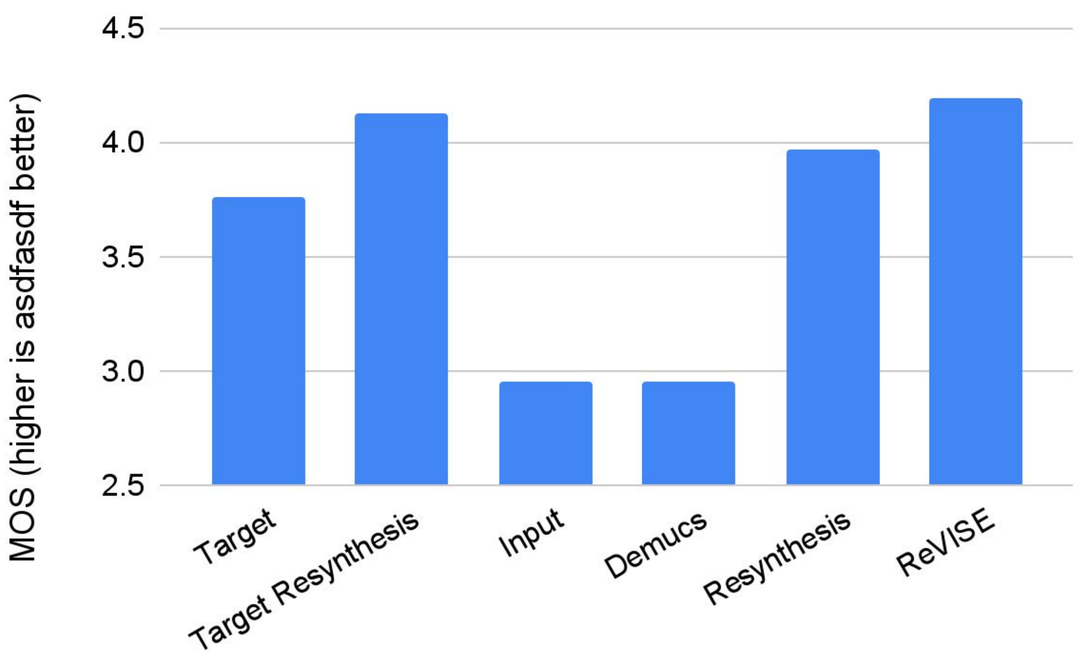
Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

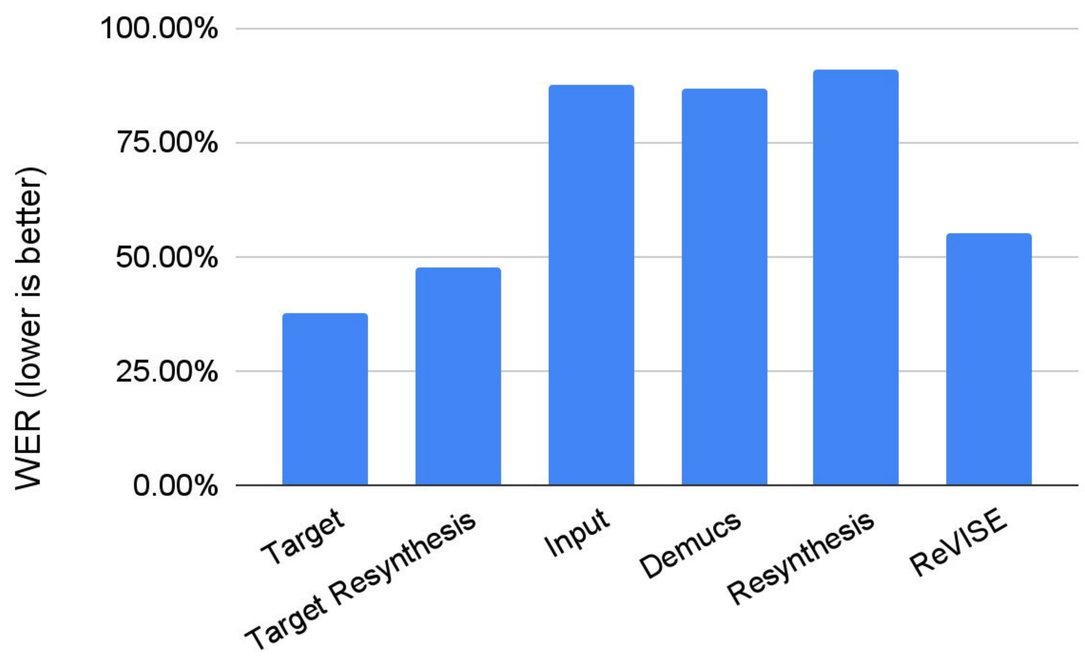
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

EasyCom

An audio-visual speech dataset addressing the cocktail party problem which contains clean close-talking recordings and noisy distant recordings with background noise, loud interfering speech, and room reverberation.
It contains only 1.6h of training data.

Separation





To learn more:

Meta AI

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Reference

To learn more:

Meta AI

ReVISE: Self-Supervised Speech **Re**synthesis with **V**isual **I**nput for Universal and Generalized **S**peech **E**nhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Input

To learn more:

Meta AI

ReVISE: Self-Supervised Speech **Re**synthesis with **V**isual Input for Universal and Generalized **S**peech **E**nhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Output

To learn more:

Meta AI

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Reference

To learn more:

Meta AI

ReVISE: Self-Supervised Speech **Re**synthesis with **V**isual **I**nput for Universal and Generalized **S**peech **E**nhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Input

To learn more:

Meta AI

ReVISE: Self-Supervised Speech **Re**synthesis with **V**isual **I**nput for Universal and Generalized **S**peech **E**nhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Output

To learn more:

Meta AI

ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement

Wei-Ning Hsu[1], Tal Remez[1], Bowen Shi[1,3], Jacob Donley[2], Yossi Adi[1,4]
[1] FAIR, Meta AI Research
[2] Meta Reality Labs Research
[3] Toyota Technological Institute at Chicago
[4] The Hebrew University of Jerusalem

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Limitations

While this paper studies universal enhancement, it only concerns audio but not video distortion. For future work, we hope to study more general multimodal audio-visual speech enhancement where an enhancement model can recover distortion in both modalities.

To learn more: