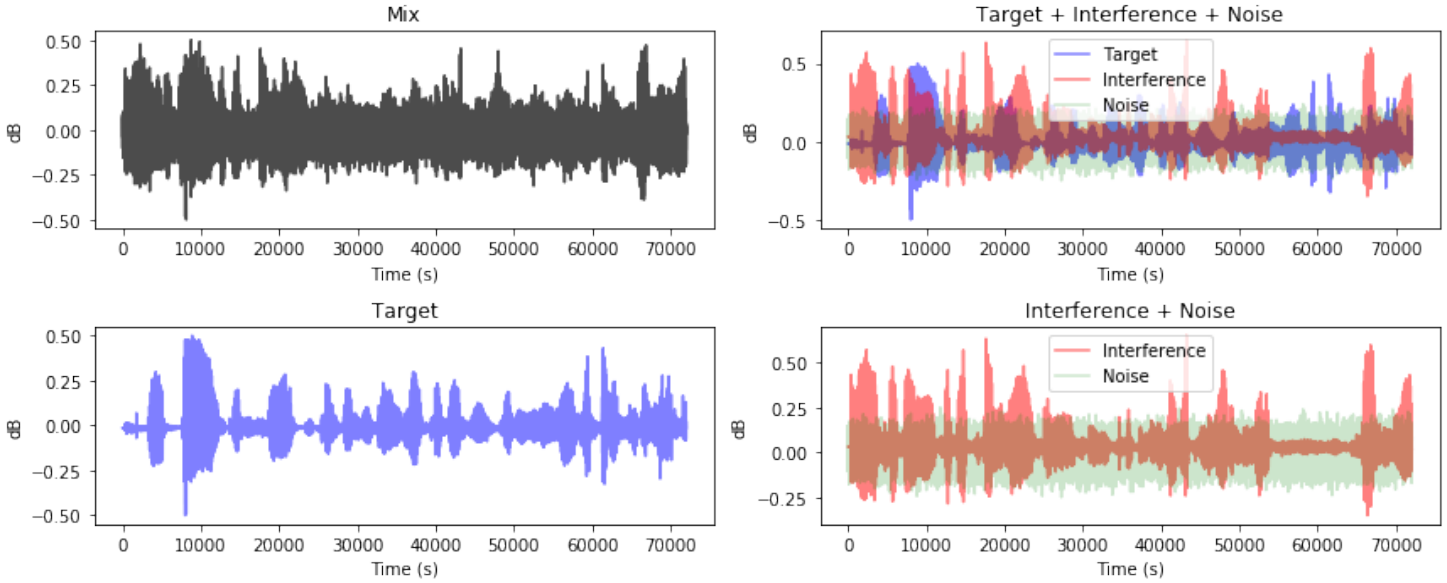
# AV(SE)$^2$: Audio-Visual Squeeze-Excite Speech Enhancement

Michael L. Iuzzolino
University of Colorado Boulder, Boulder, CO 80309, USA

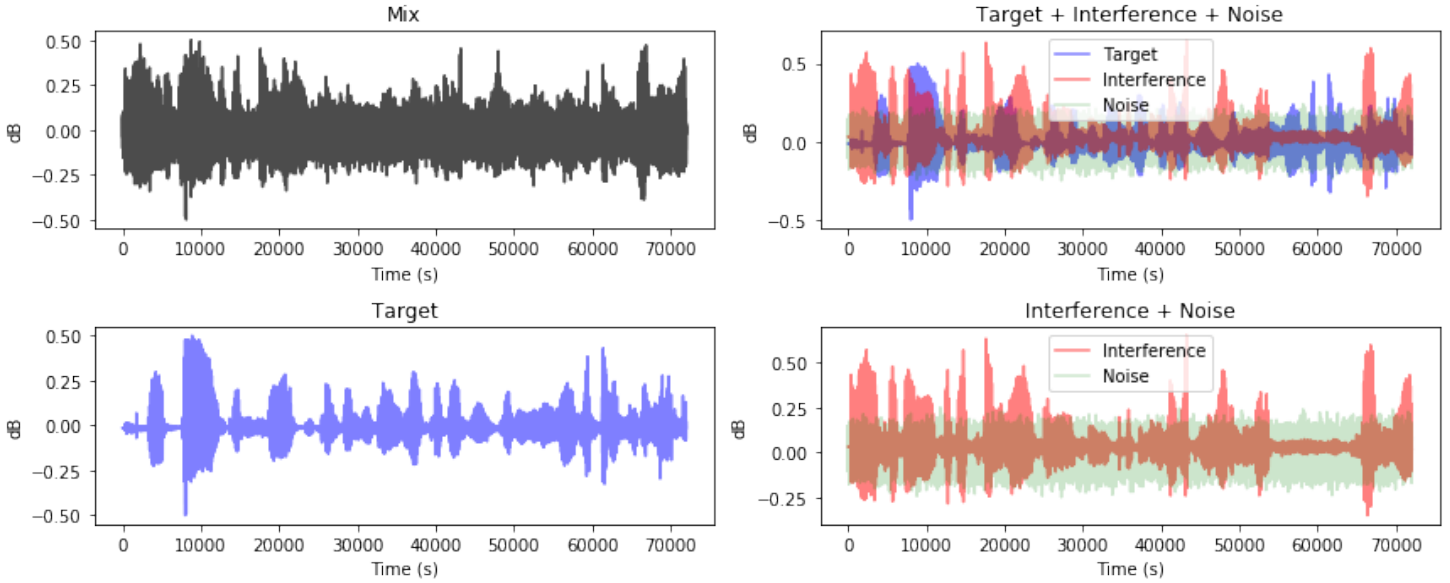
Kazuhito Koishida
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

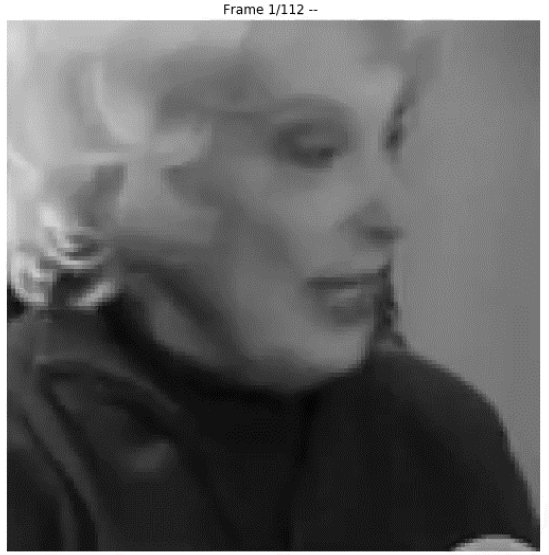

# Speech Enhancement




Audio Only Speech Enhancement

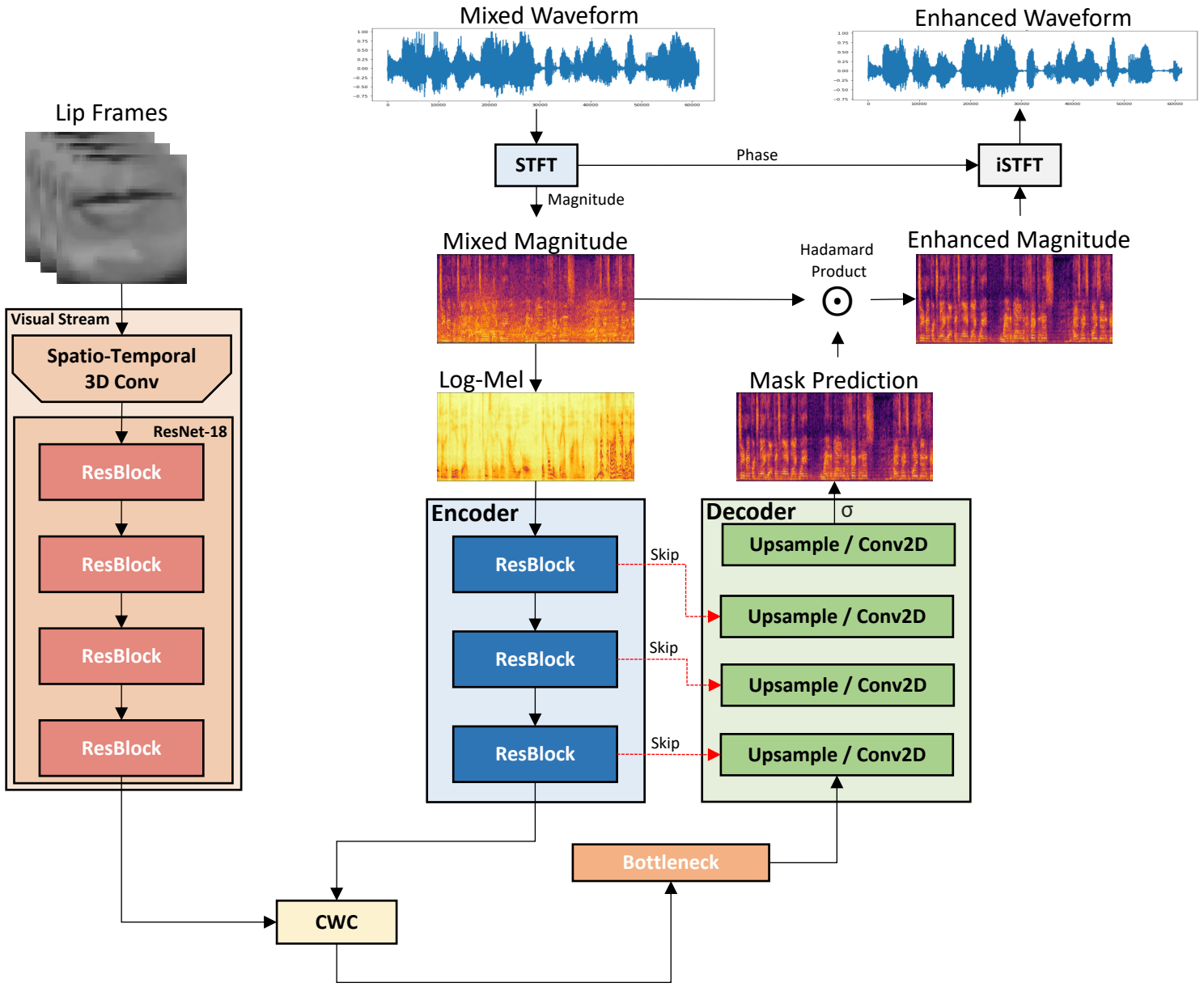
# Audio-Visual Speech Enhancement




Audio-Visual Speech Enhancement

# AV Fusion


Mixed Waveform
Enhanced Waveform
Lip Frames
STFT
Phase
iSTFT
Magnitude
Mixed Magnitude
Hadamard
Product
Enhanced Magnitude
Visual Stream
Spatio-Temporal
3D Conv
ResNet-18
ResBlock
ResBlock
ResBlock
ResBlock
Log-Mel
Mask Prediction
Encoder
Decoder
σ
ResBlock
ResBlock
ResBlock
Skip
Skip
Skip
Upsample / Conv2D
Upsample / Conv2D
Upsample / Conv2D
Upsample / Conv2D
Bottleneck
CWC

# Audio-Visual Information

## Cocktail Party Effect

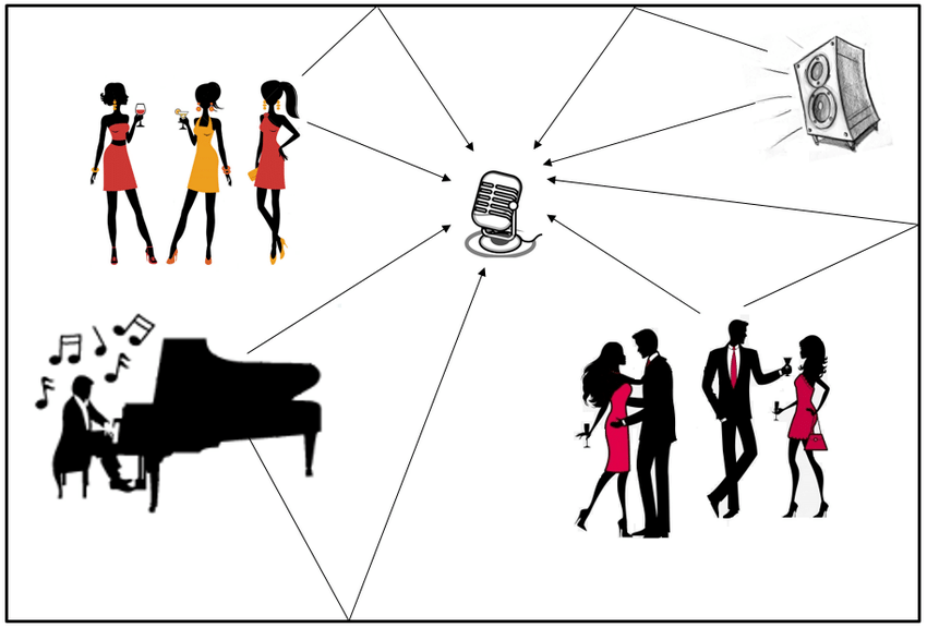


Figure taken from: Thanh, Duong Thi Hien. "Audio Source Separation exploiting NMF-based Generic Source Spectral Model"

**Cherry, C. (1953). Cocktail party problem. *Journal of the Acoustical Society of America*, *25*, 975-979.**

# Audio-Visual Information

## Cocktail Party Effect

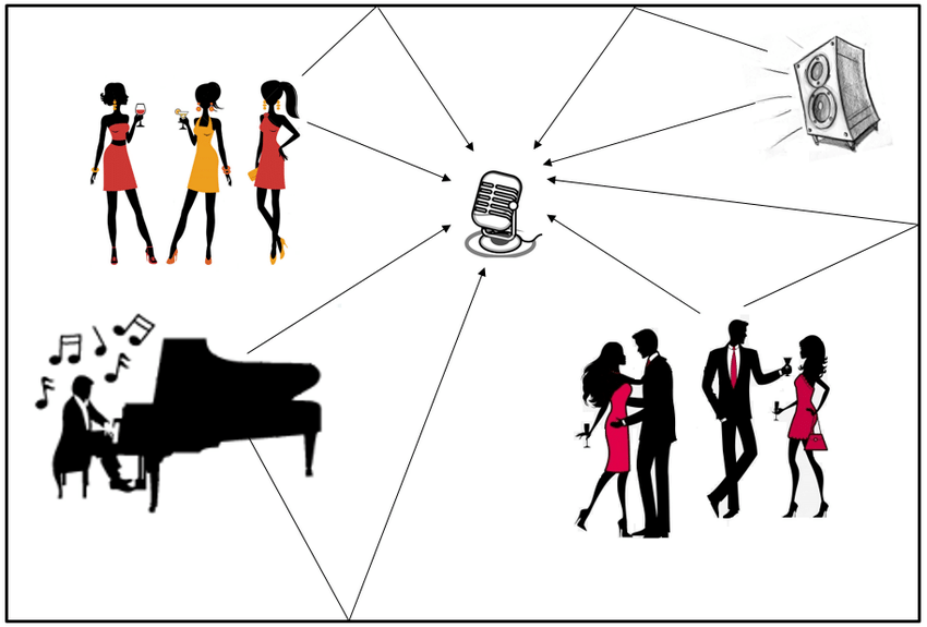


Figure taken from: Thanh, Duong Thi Hien. "Audio Source Separation exploiting NMF-based Generic Source Spectral Model"


Cherry, C. (1953). Cocktail party problem. *Journal of the Acoustical Society of America*, *25*, 975-979.

**Asif A Ghazanfar and Nikos K Logothetis, "Neuroperception: Facial expressions linked to monkey calls," Nature, vol. 423, no. 6943, pp. 937, 2003.**

**Sarah Partan and Peter Marler, "Communication goes multimodal," Science, vol. 283, no. 5406, pp. 1272–1273, 1999.**

**Candy Rowe, "Sound improves visual discrimination learning in avian predators," Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 269, no. 1498, pp. 1353–1357, 2002.**

# Audio-Visual Information

## Cocktail Party Effect

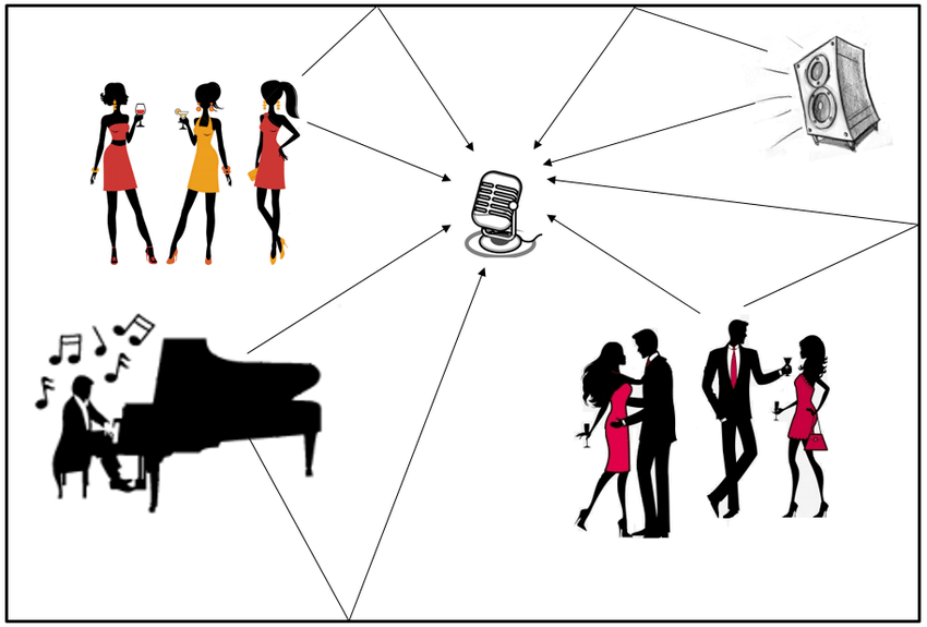


Figure taken from: Thanh, Duong Thi Hien. "Audio Source Separation exploiting NMF-based Generic Source Spectral Model"


Cherry, C. (1953). Cocktail party problem. *Journal of the Acoustical Society of America*, *25*, 975-979.

Asif A Ghazanfar and Nikos K Logothetis, “Neuroperception: Facial expressions linked to monkey calls,” Nature, vol. 423, no. 6943, pp. 937, 2003.

Sarah Partan and Peter Marler, “Communication goes multimodal,” Science, vol. 283, no. 5406, pp. 1272–1273, 1999.

Candy Rowe, “Sound improves visual discrimination learning in avian predators,” Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 269, no. 1498, pp. 1353–1357, 2002.

**Shahram Moradi, Bj¨orn Lidestam, and Jerker R¨onnberg, “Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy,” Frontiers in Psychology, vol. 4, pp. 359, 2013.**

**Hanne Stenzel, Jon Francombe, and Philip JB Jackson, “Limits of perceived audio-visual spatial coherence as defined by reaction time measurements,” Frontiers in neuroscience, vol. 13, pp. 451, 2019.**

# Dataset

## Target and Interference

VoxCeleb2

Over 1 million utterances for 6,112 celebrities, extracted from videos uploaded to YouTube

Sampled 20,000 videos for development set

+ 10,000 videos for test set

## Noise

CHiME3

HuNonspeech

NoiseX-92

PCC Data

Private Datasets

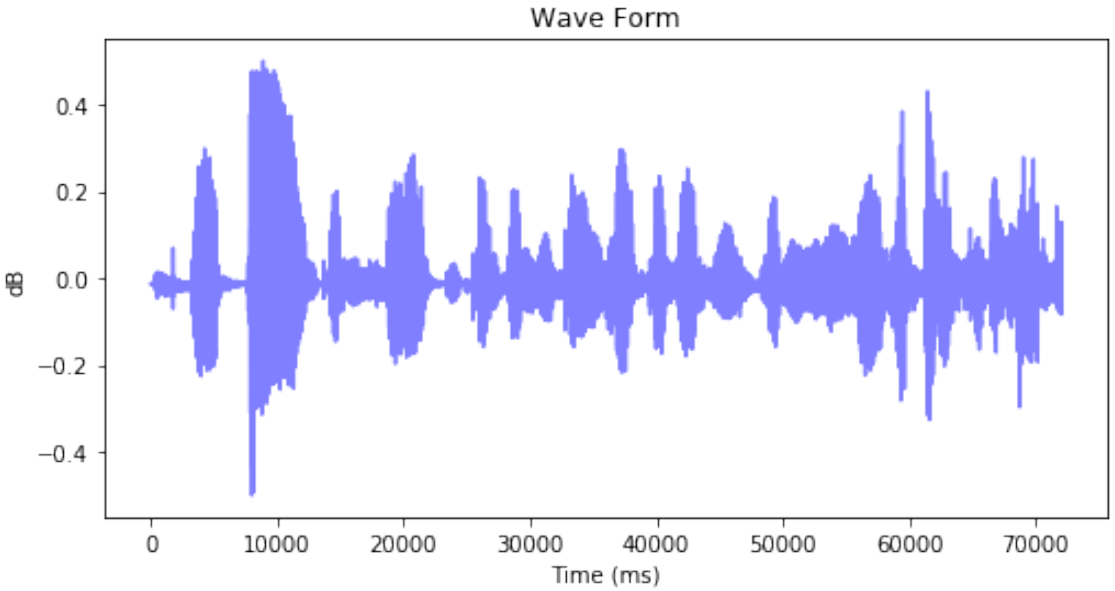
http://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

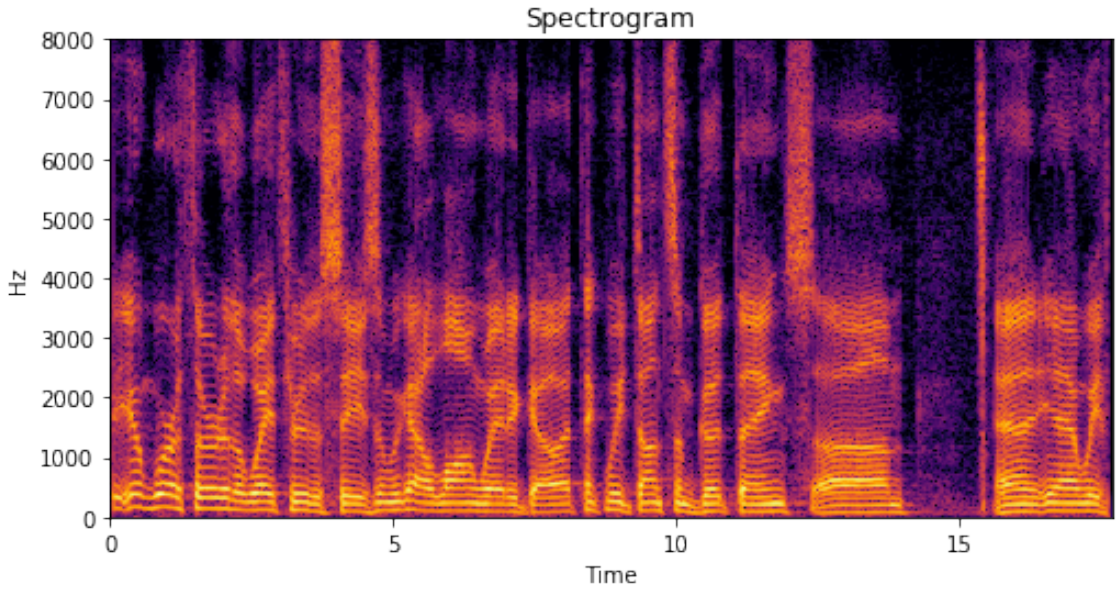
# Data Processing: Audio

iSTFT

Sampling Rate: 16kHz

Frequency Bins: 321



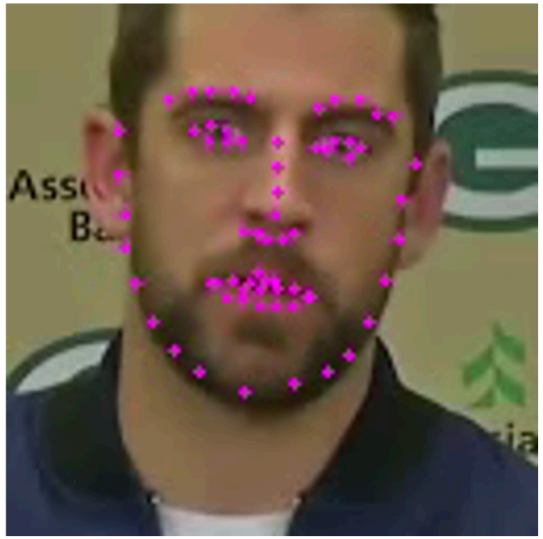




STFT

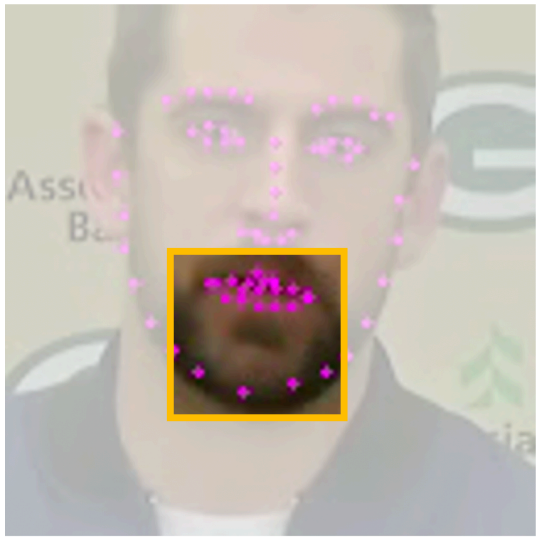
Window Size: 640

Hopsize : 160

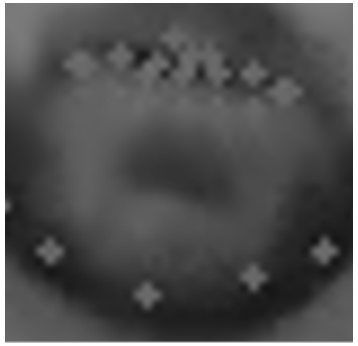
# Data Processing: Video



Facial Landmarks

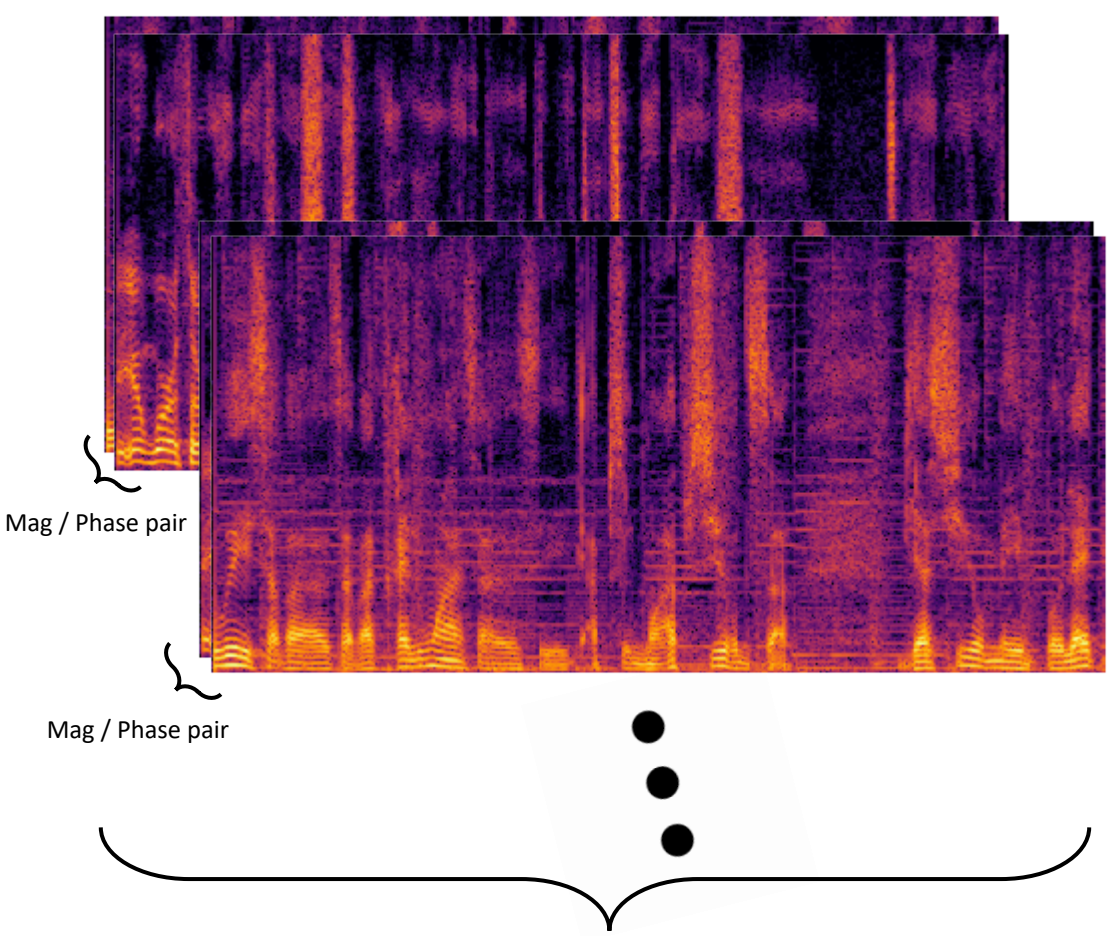


Mouth Landmarks for Cropping



Resized to (122x122)

# OFFLINE AV Data Processing





Spectrogram Set (*float16*)

Video Set (*uint8*)

**Training Set**
Utterances: ~843k
Memory: ~3.6 TB

**Validation Set**
Utterances: ~149k
Memory: ~640 GB

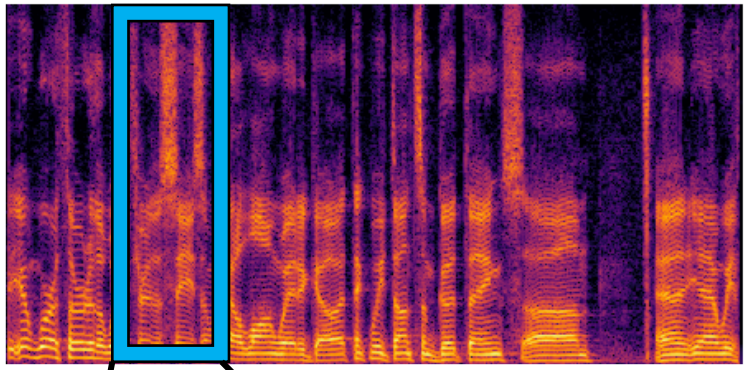
**Test Set**
Utterances: ~36k
Memory: ~126 GB

*Total Memory: ~4.4 TB*

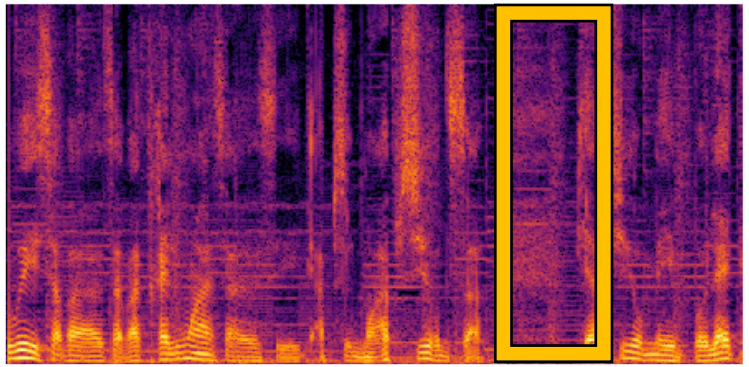
Training set's global mean / std approximated offline

# ONLINE Batch Generation

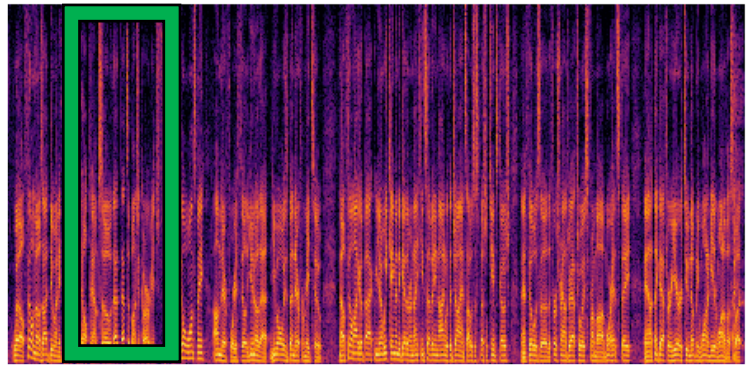
Target, spec



Noise, n

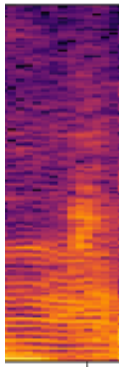


Non-Stationary Noise, nS



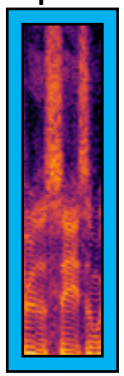
Lip Frames

mix




=

spec

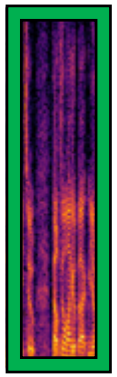


+

n

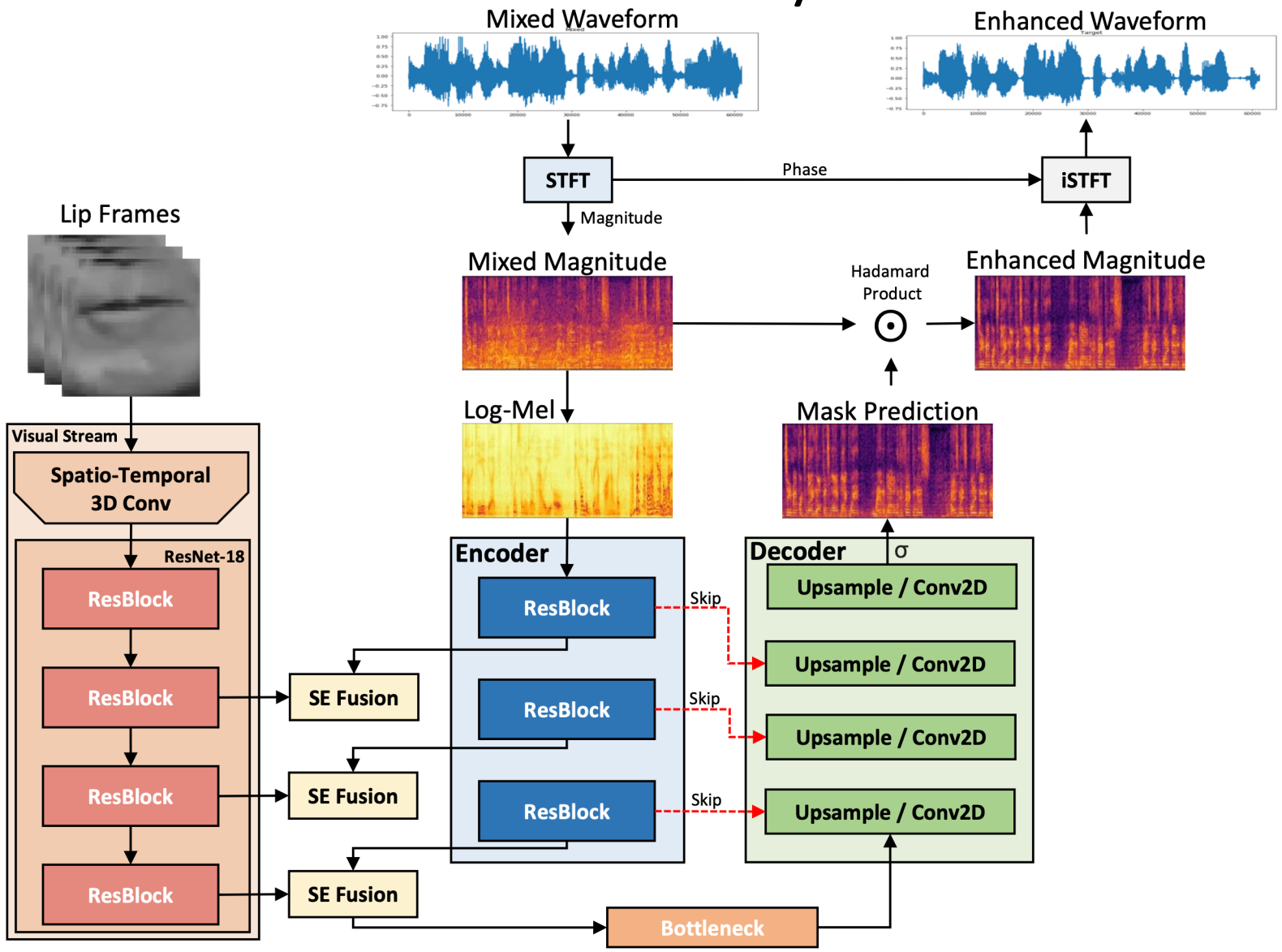
+

nS

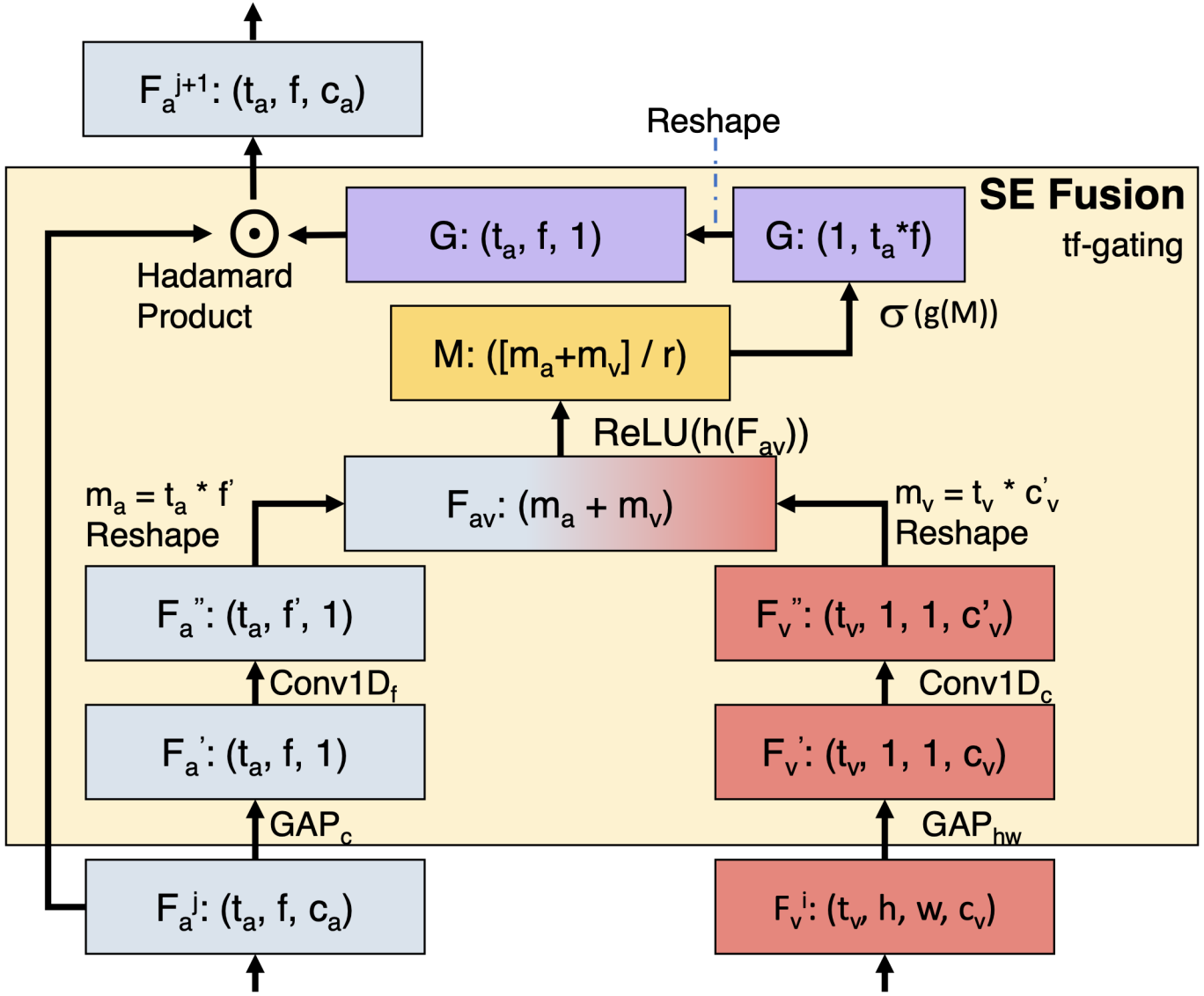


$$X_{mix} = X_{spec} + \alpha X_n + \beta X_{nS}$$

# AV Speech Enhancement System Overview



# Squeeze-Excite Fusion Block



# Experiments

Factor 1: Layer

Factor 2: Mode

Factor 3: Dimension

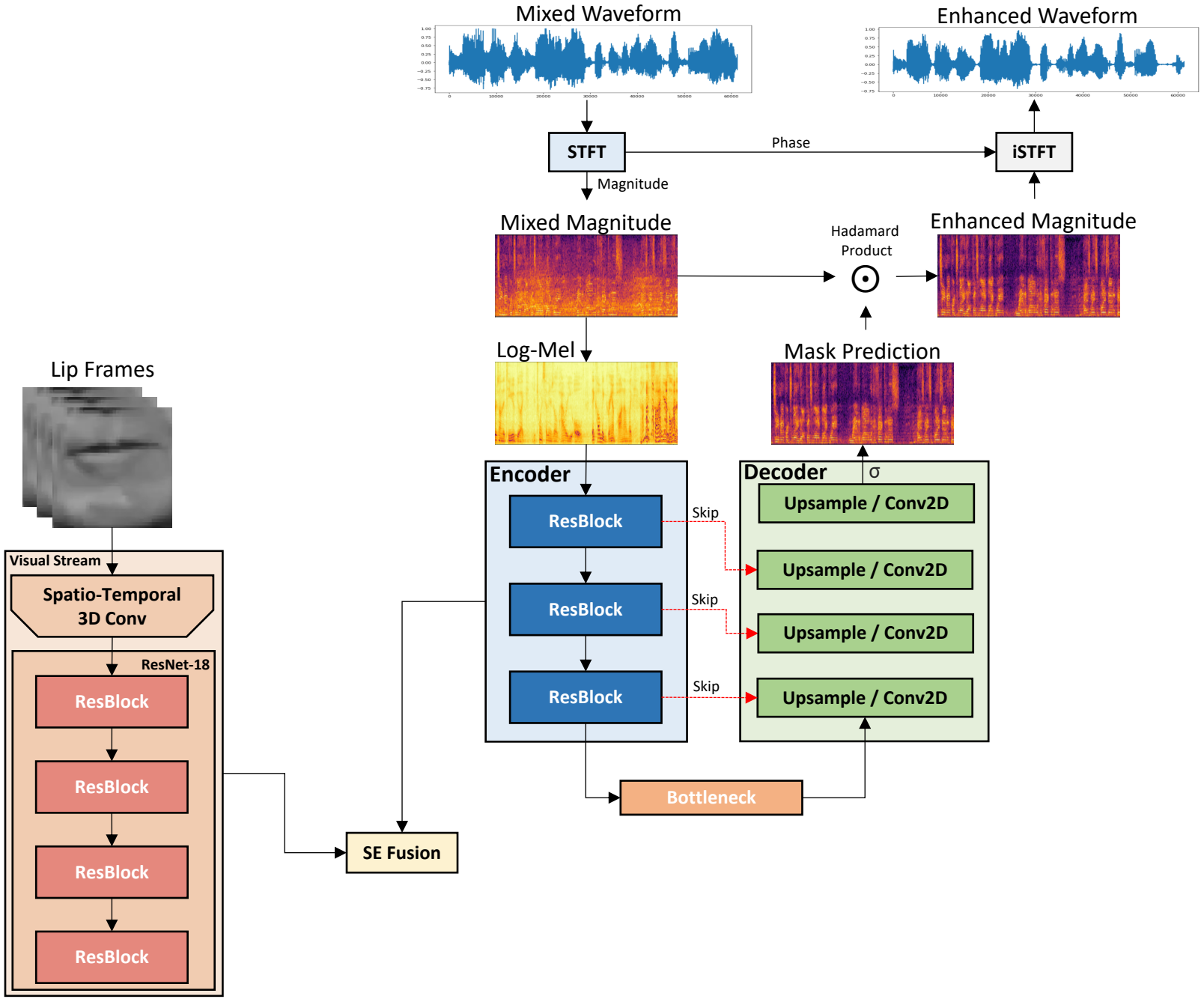
# Experiments

Factor 1: Layer

**Encoder only (E)**

Factor 2: Mode

Factor 3: Dimension



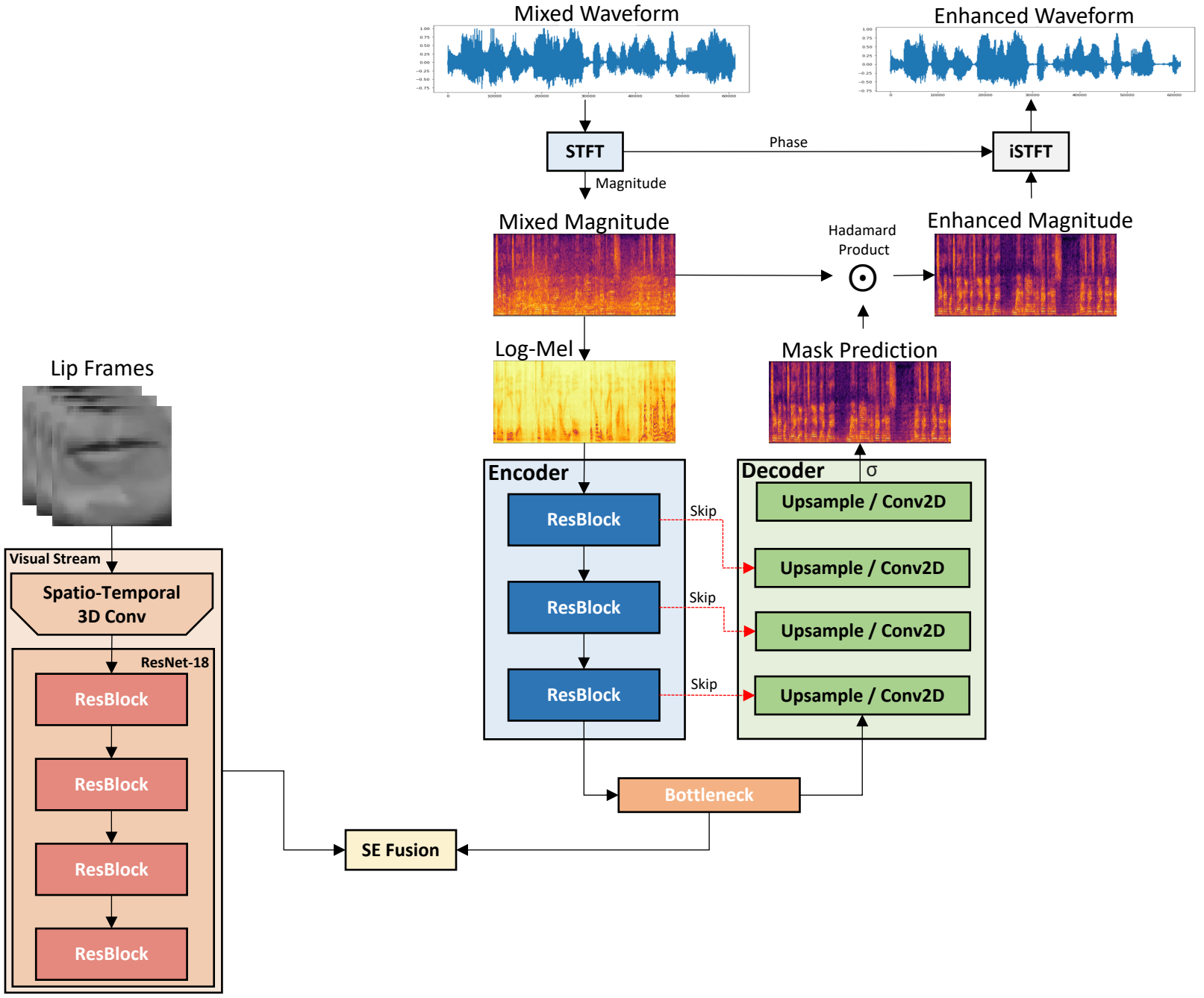
# Experiments

Factor 1: Layer

**Bottleneck (BN)**

Factor 2: Mode

Factor 3: Dimension



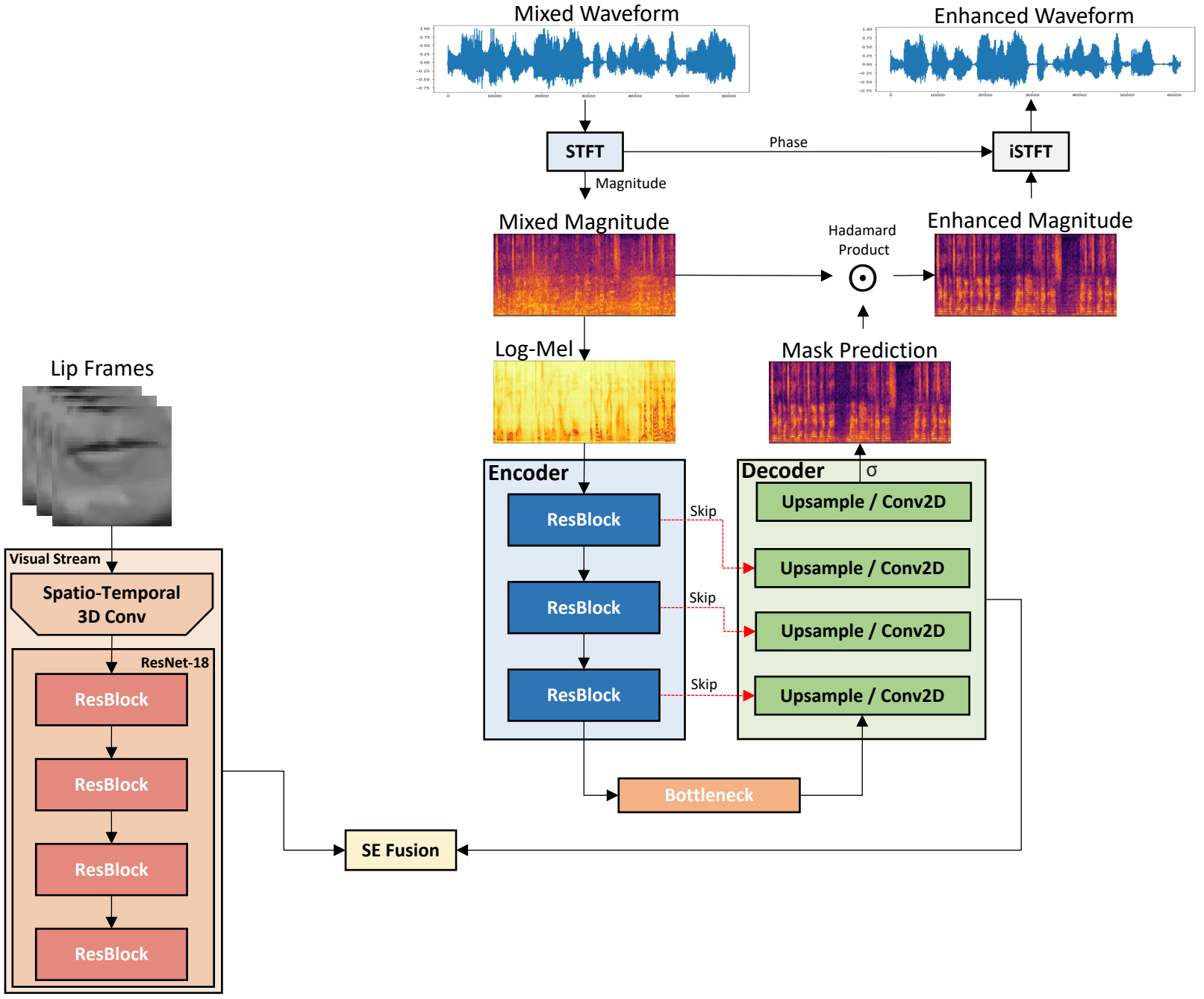
# Experiments

Factor 1: Layer

**Decoder only (D)**

Factor 2: Mode

Factor 3: Dimension



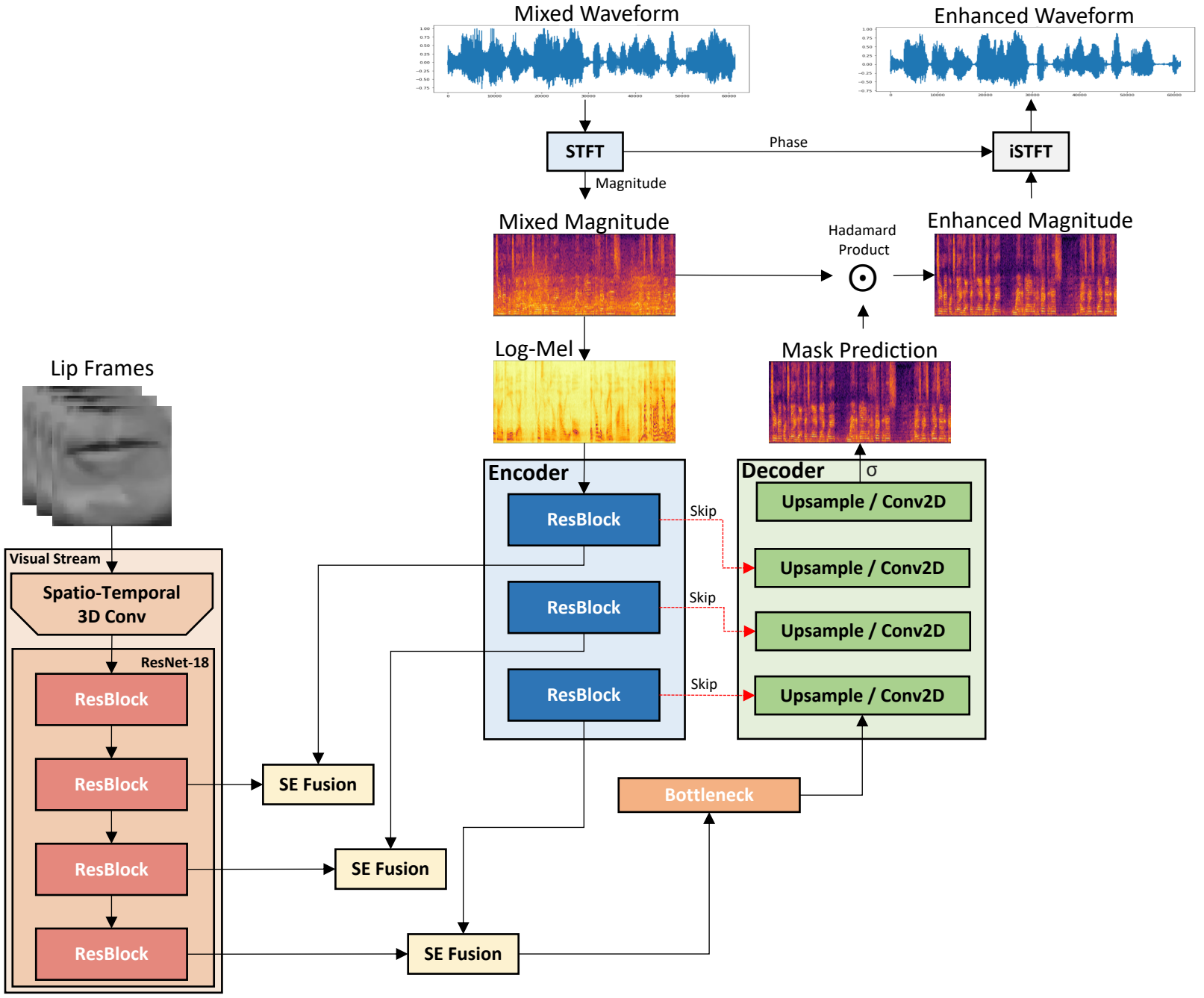
# Experiments

Factor 1: Layer

Factor 2: Mode

**1-to-1**

Factor 3: Dimension



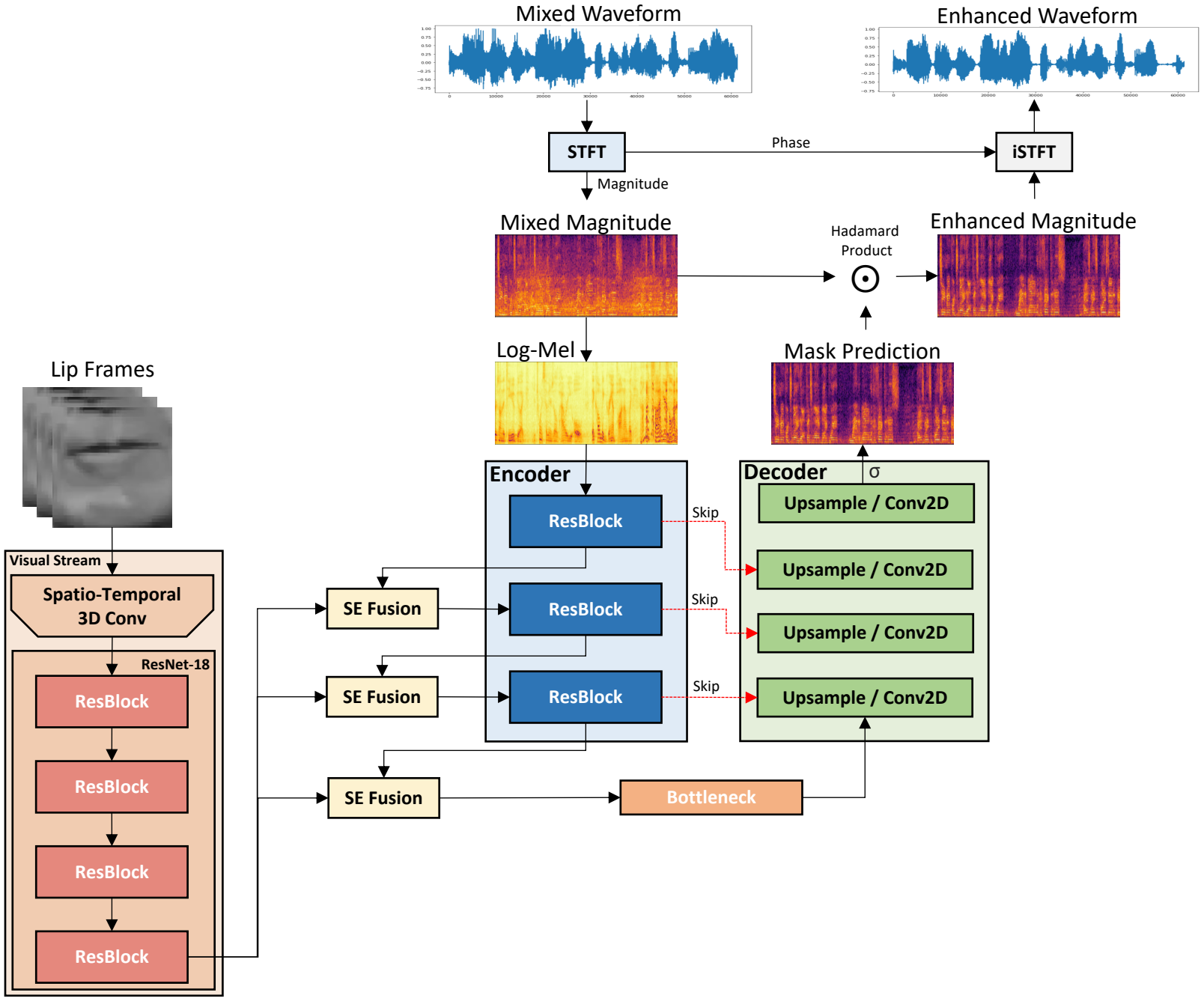
# Experiments

Factor 1: Layer

Factor 2: Mode

**$F_v^4$-to-all**

Factor 3: Dimension



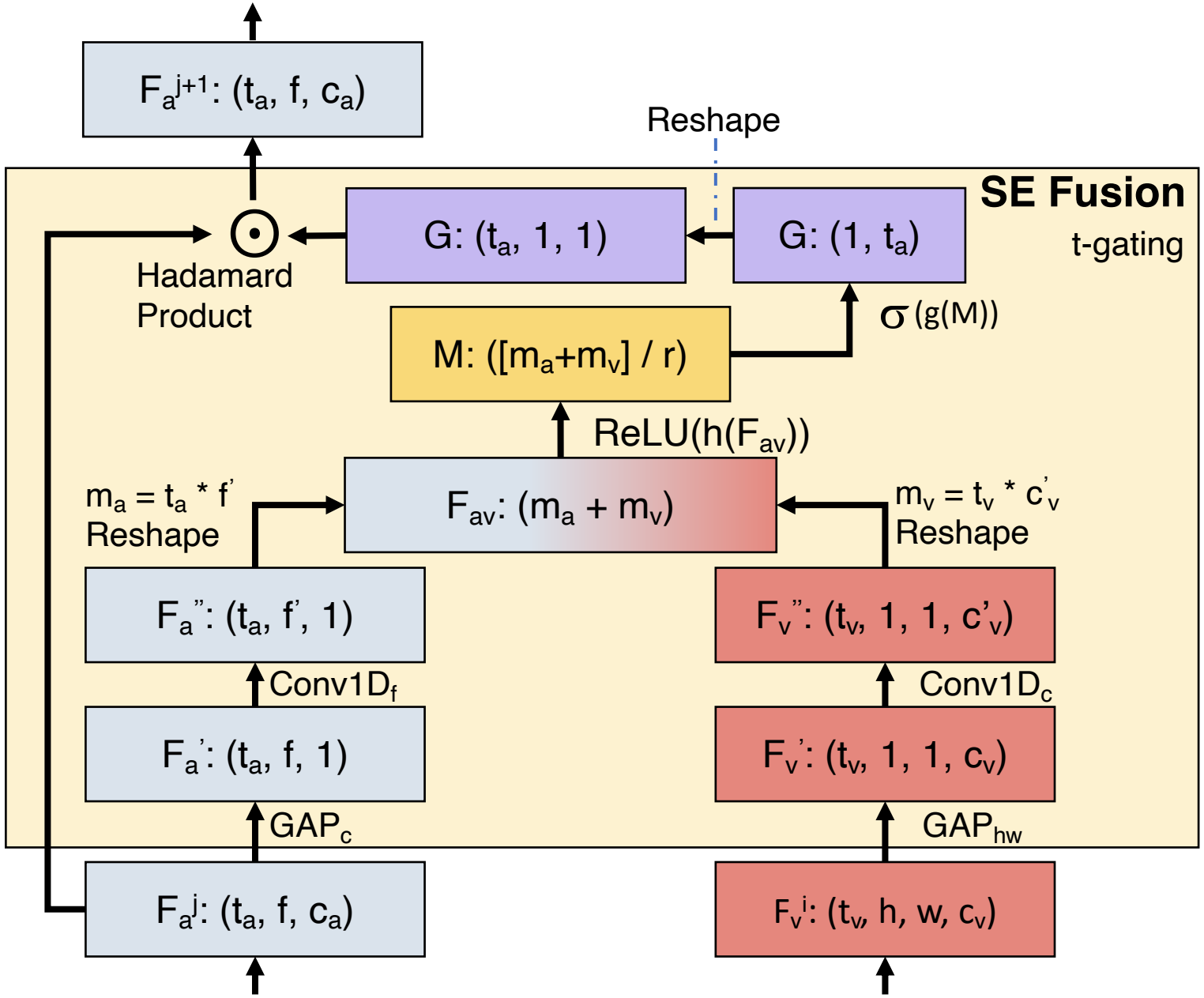
# Experiments

Factor 1: Layer

Factor 2: Mode

Factor 3: Dimension

**t-gating**



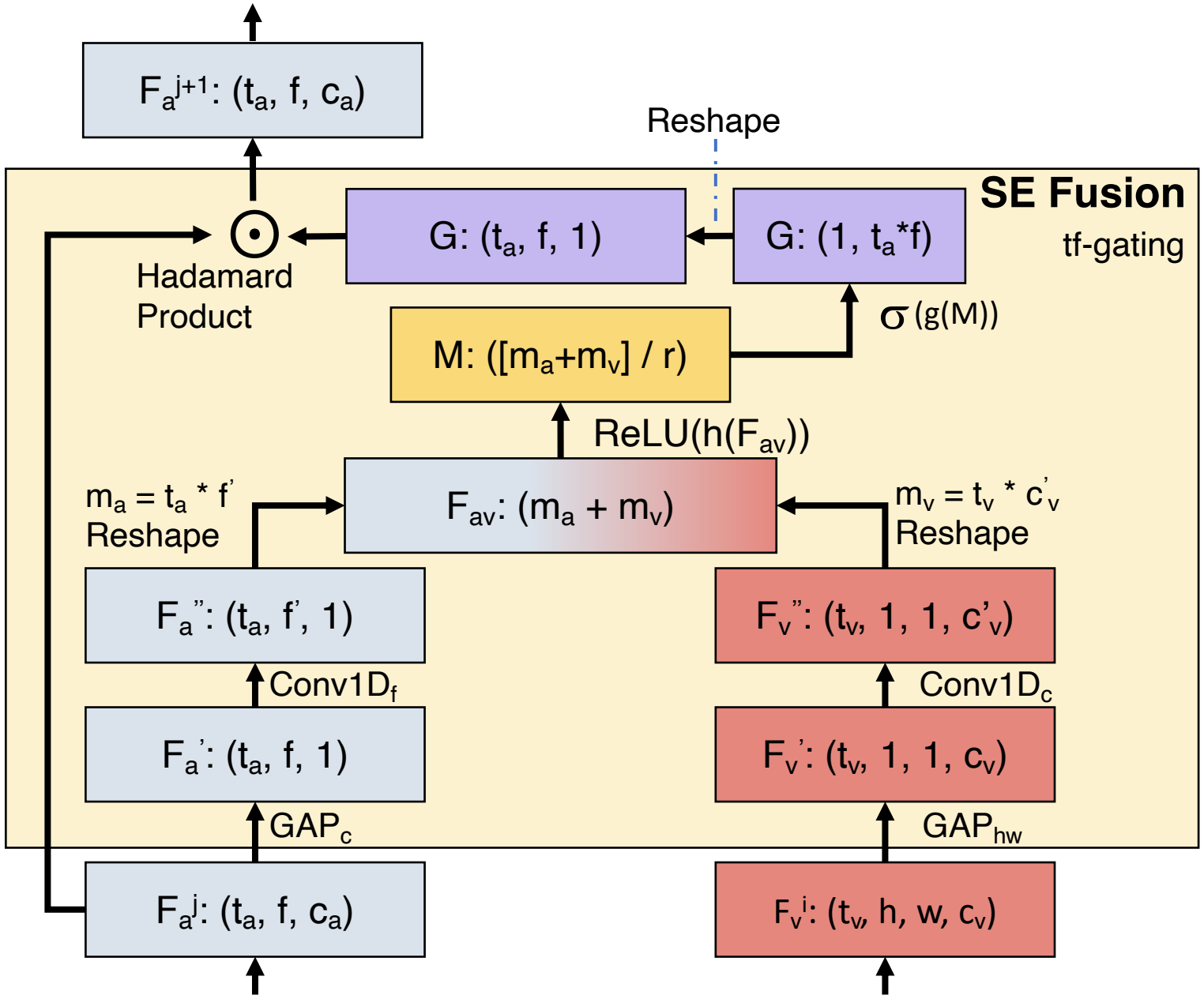
# Experiments

Factor 1: Layer

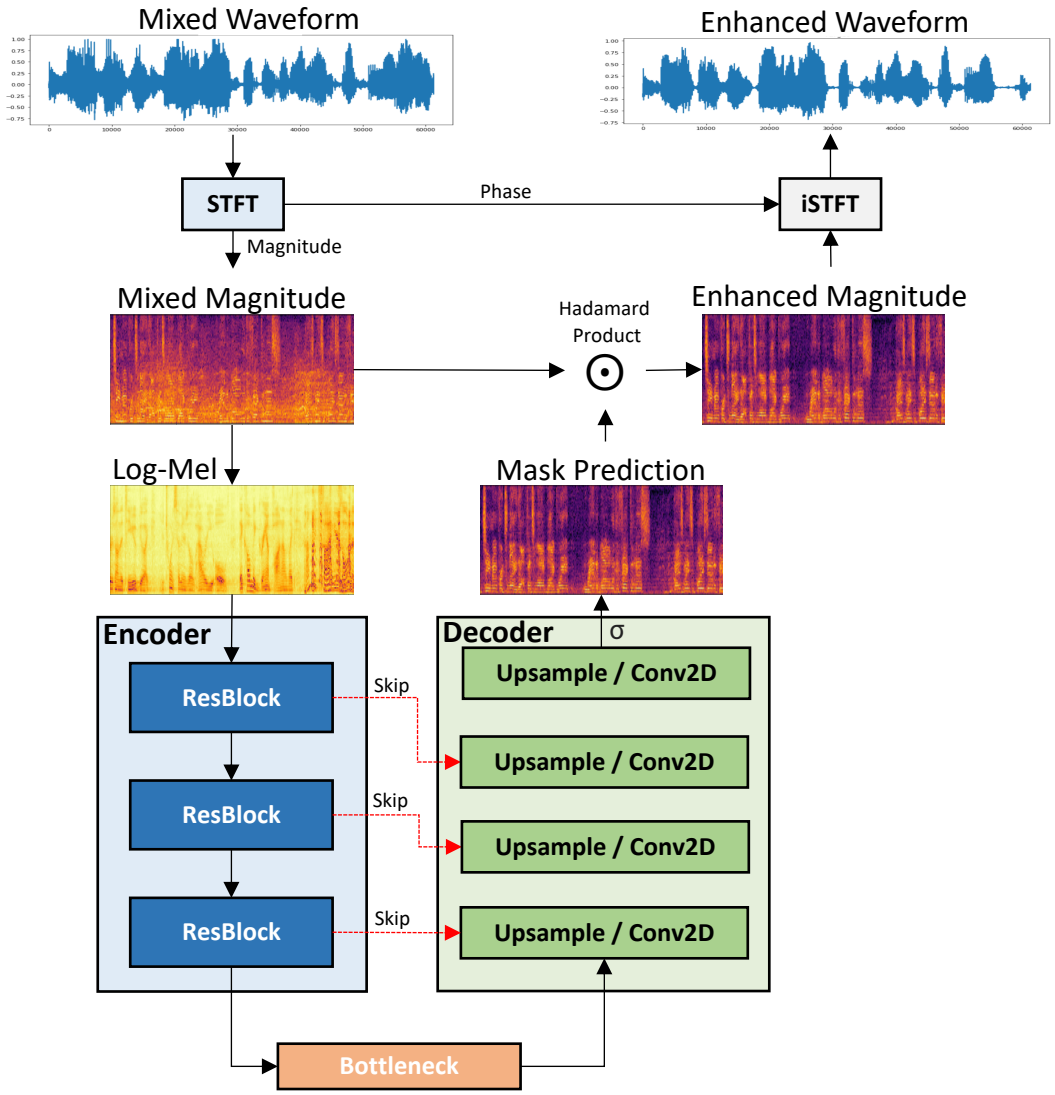
Factor 2: Mode

Factor 3: Dimension

**tf-gating**

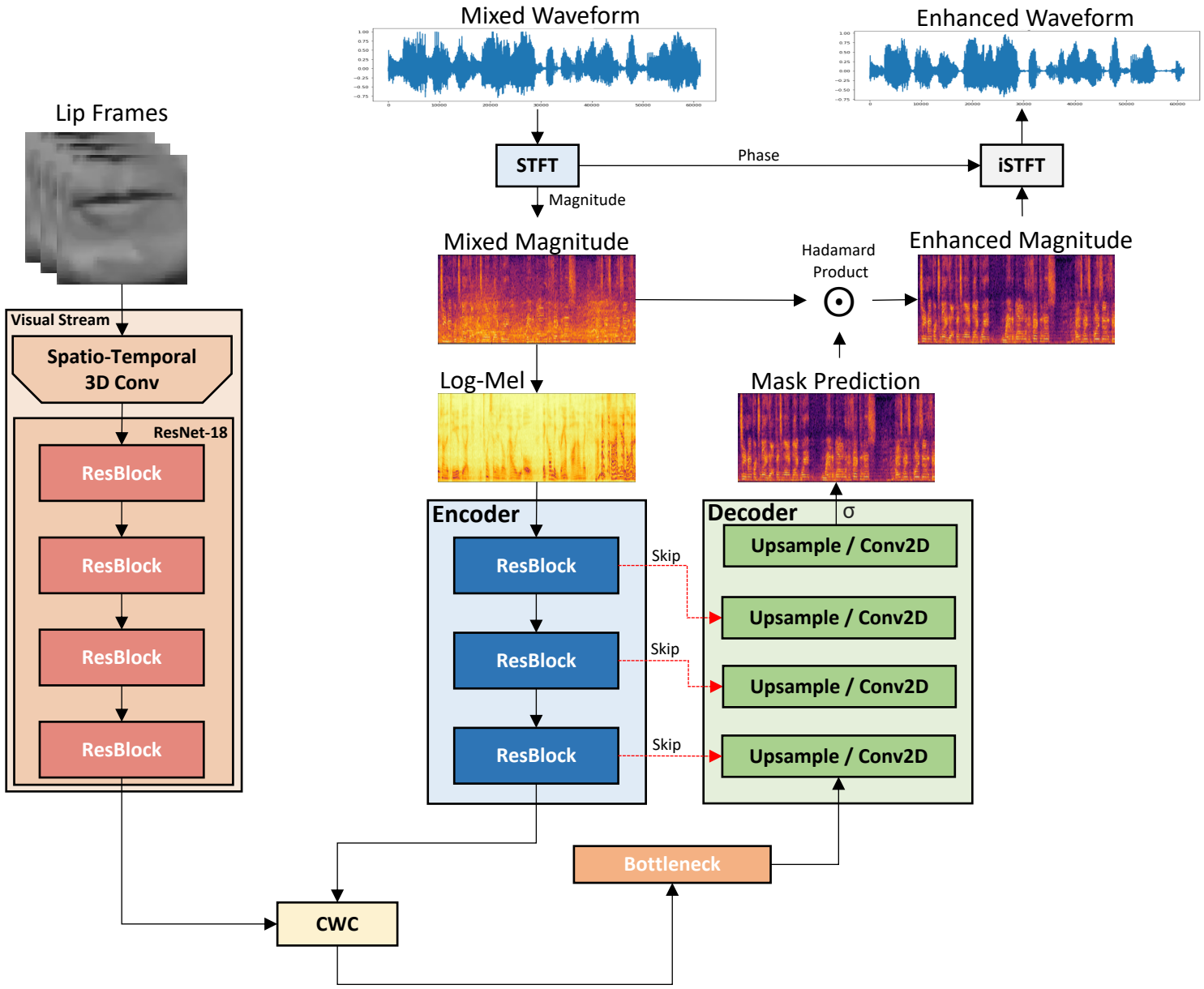


# Baselines




Audio-only Network

# Baselines




Audio-Visual Fusion
via Channel-wise Concatenation (CWC)

# Training

Batch Size: 4

Optimizer: Adam with initial alpha = 0.01

Loss Function:

$$\mathcal{L} = \left\| M \odot X_{mix} - X_{spec} \right\|_1$$

Validation loss monitored to avoid overfitting

Models trained ~140 on average

# Results

| Set | Layer | Mode | Dim | PESQ | STOI |
|---|---|---|---|---|---|
| Target | - | - | - | 4.644 | 1.000 |
| Mixed | - | - | - | 2.188 | 0.900 |
| AO Baseline | - | - | - | 2.788 | 0.930 |
| AV Baseline | BN | CWC | - | 2.917 | 0.933 |

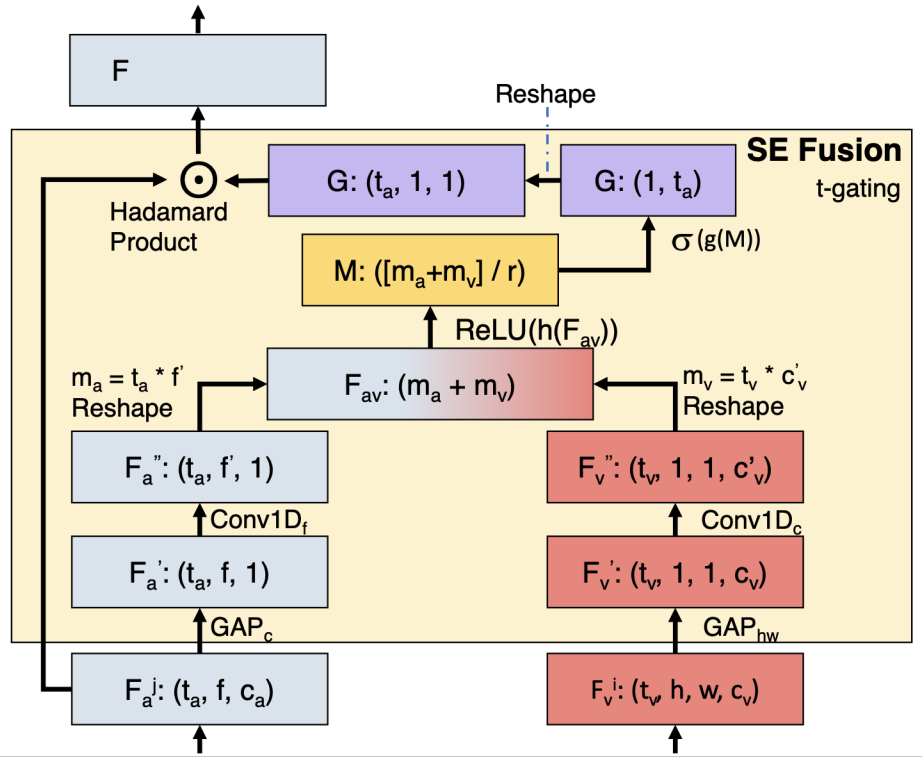
| Set | Layer | Mode | Dim | PESQ | STOI |
|---|---|---|---|---|---|
| AV(SE)$^2$-v1 | BN | $F_v^4$-to-all | c | 2.920 | 0.933 |
| AV(SE)$^2$-v2 | BN | $F_v^4$-to-all | t | **2.982** | **0.935** |
| AV(SE)$^2$-v3 | BN | $F_v^4$-to-all | tf | 2.907 | 0.934 |
| AV(SE)$^2$-v4 | BN | $F_v^4$-to-all | tc | 2.922 | 0.933 |
| AV(SE)$^2$-v5 | BN | $F_v^4$-to-all | tfc | 2.923 | 0.934 |
| AV(SE)$^2$-v6 | E | $F_v^4$-to-all | t | 2.891 | 0.932 |
| AV(SE)$^2$-v7 | E | $F_v^4$-to-all | tf | 2.888 | 0.930 |
| AV(SE)$^2$-v8 | E | $F_v^4$-to-all | tc | 2.862 | 0.934 |
| AV(SE)$^2$-v9 | E | 1-to-1 | t | 2.893 | 0.933 |
| AV(SE)$^2$-v10 | E | 1-to-1 | tf | 2.938 | 0.934 |
| AV(SE)$^2$-v11 | E | 1-to-1 | tc | 2.889 | 0.933 |
| AV(SE)$^2$-v12 | D | $F_v^4$-to-all | t | 2.782 | 0.923 |
| AV(SE)$^2$-v13 | D | $F_v^4$-to-all | tf | 2.837 | 0.930 |
| AV(SE)$^2$-v14 | D | $F_v^4$-to-all | tc | 2.802 | 0.928 |
| AV(SE)$^2$-v15 | D | 1-to-1 | t | 2.846 | 0.927 |
| AV(SE)$^2$-v16 | D | 1-to-1 | tf | 2.868 | 0.930 |
| AV(SE)$^2$-v17 | D | 1-to-1 | tc | 2.888 | 0.931 |

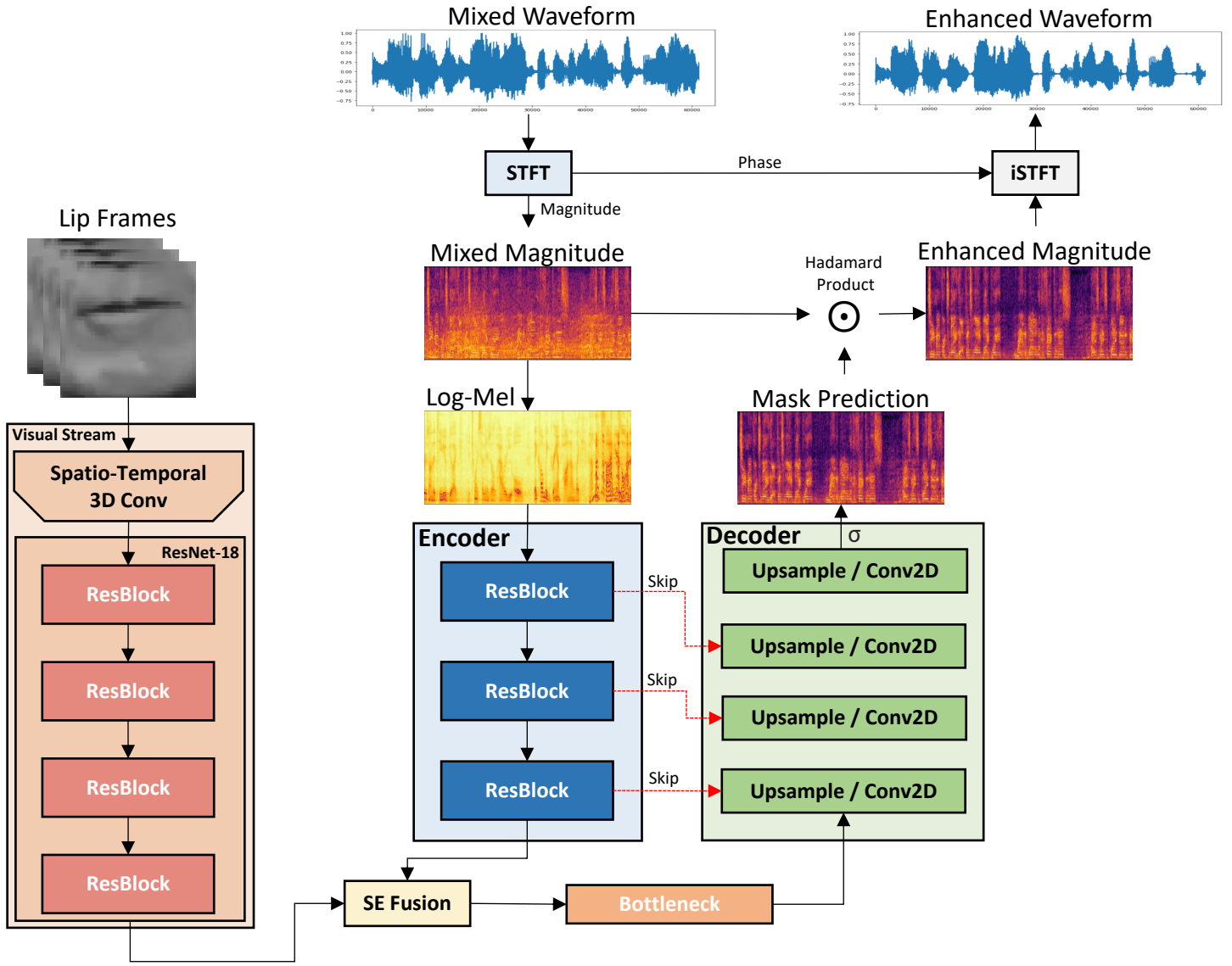
# Summary

- In this work, we presented a novel approach to audio-visual fusion via squeeze-excitation blocks
- Time-based recalibration at the bottleneck offers a significant improvement over CWC
- The improvements in objective measures are accompanied by a significant reduction in the number of model parameters

# Summary

- Best model



# Thank You!

## AV(SE)$^2$: Audio-Visual Squeeze-Excite Speech Enhancement

Michael L. Iuzzolino
Michael.iuzzolino@colorado.edu

Kazuhito Koishida
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

University of Colorado Boulder

Microsoft

# Thank You!

Reach out at michael.iuzzolino@colorado.edu