STANFORD

Task and Perception-aware Distributed Source Coding for Correlated Speech under Bandwidth-constrained Channels

SAGNIK BHATTACHARYA, MUHAMMAD AHMED MOHSIN, AHSAN BILAL, JOHN M. CIOFFI

MARCH 4, 2025

support acknowledgement:
[1]with Samsung, Intel, Ericsson

ERICSSON intel SAMSUNG ELECTRONICS

Challenges faced by current Autoencoder-based source coding techniques

1. **Dynamic bitrate adaptation**: Autoencoder-based source coding requires specifying a fixed autoencoder output dimension, followed by entropy coding. This is inefficient compared to variable dimension autoencoder output. But that requires training a new autoencoder every time the bitrate requirement (i.e., channel conditions) changes

2. Most source-coding methods consider the single-user problem. In a **multi-user scenario**, the correlation between various participating users could be utilized to achieve lower compression ratios.

3. **Task and perception-aware source coding:** Training end-to-end source coding + downstream task models, especially when incorporating a perception component, could provide lower compression ratios than training for traditional bit reconstruction loss

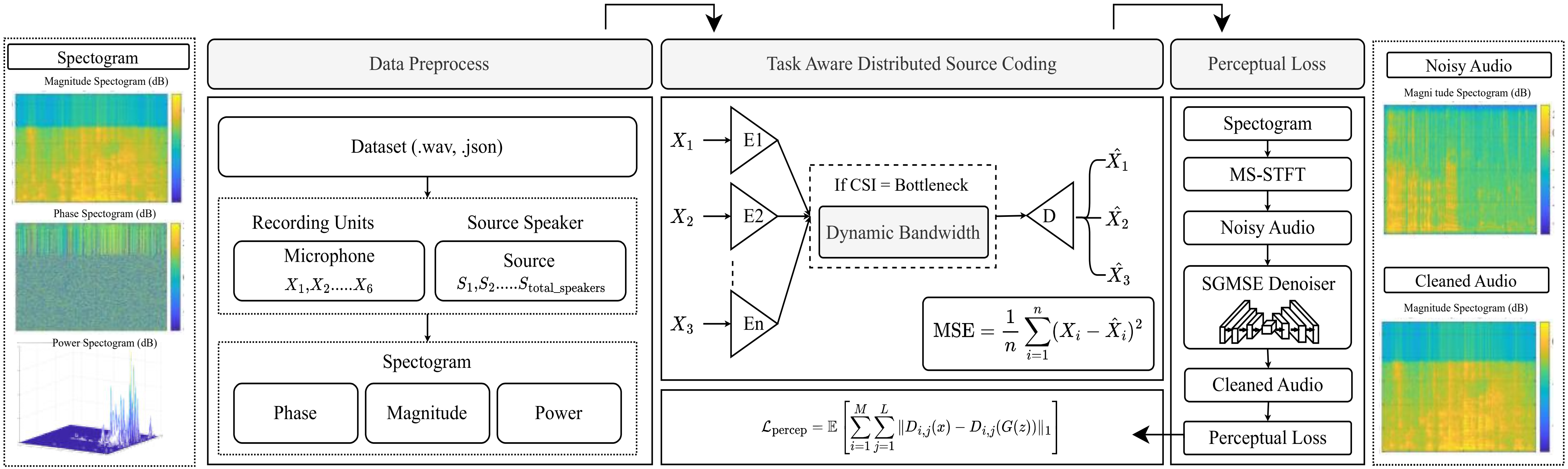
We build a pipeline incorporating all of the above.

Proposed Distributed Source Coding Pipeline

Channel State Information (CSI)-aware Dynamic Bitrate

Neural Distributed Principal Component Analysis (NDPCA)-aided Distributed Encoder

Task and Perception-aware Loss Function



CSI-aware Dynamic Bitrate

- Minimum number of bits per symbol required to maintain a quantization distortion $\leq D$ for signal of energy σ^2 is
  - $R(D) \leq \frac{1}{2}\log_2(\frac{\sigma^2}{D})$

- If encoder at source s at time t has dimension $l_{s,t}$, and the outputs are sent @ $\frac{1}{K}$ outputs/second, the bitrate of that source is

$$E_{s,t} = \frac{l_{s,t}\eta R(D)}{K} \leq \frac{l_{s,t}\eta \log_2(\frac{\sigma^2}{D})}{2K}$$

  where η is the number of symbols per floating point output.

- Therefore, the sum of bitrates is given by

$$\sum_s l_{s,t} \leq \frac{2C_t K}{\eta \log_2\left(\frac{\sigma^2}{D}\right)}$$

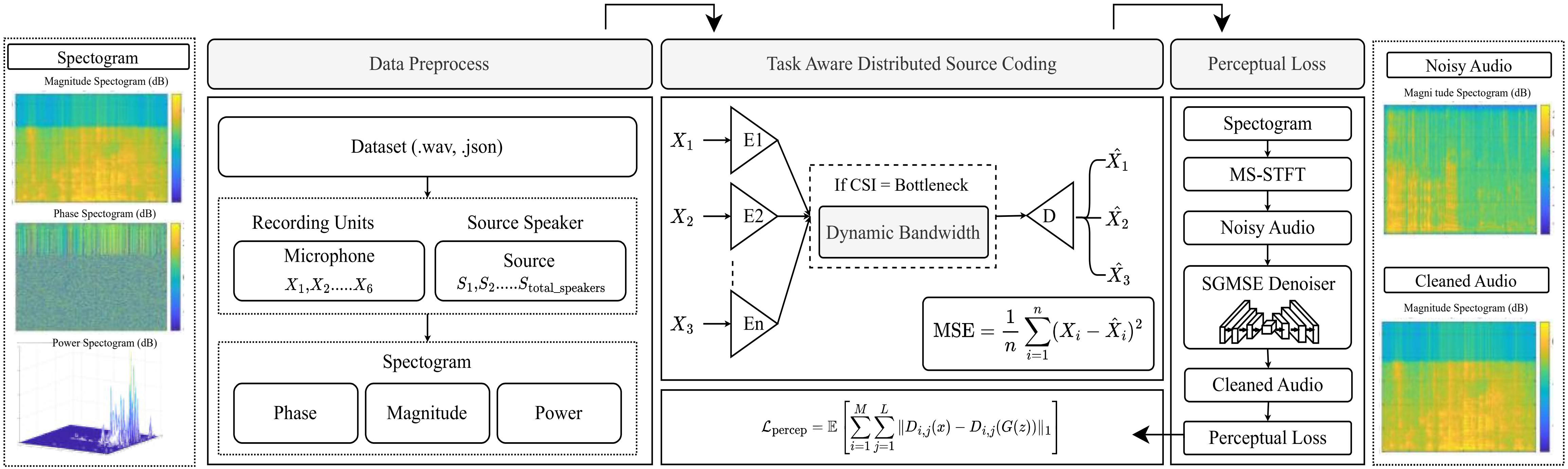
  where C_t is the total uplink channel capacity at time t

Proposed Distributed Source Coding Pipeline

CSI-aware Dynamic Bitrate

Neural Distributed Principal Component Analysis (NDPCA)-aided Distributed Encoder

Task and Perception-aware Loss Function

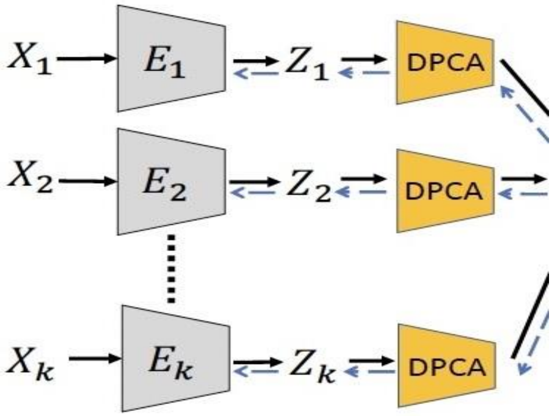


Neural Distributed Principal Component Analysis

- NDPCA combines two methods: **neural autoencoders** and **Distributed Principal Component Analysis (DPCA)**. It leverages neural networks to compress data non-linearly and uses DPCA to adjust this compression dynamically based on available bandwidth and data importance.

- It comprises two stages
  - **Neural Encoding:** Neural autoencoders first generate compact, fixed-size data representations.
  - **DPCA:** DPCA then reduces these representations further based on available bandwidth, prioritizing the most important features.

- It does not require retraining for varying bitrate requirements



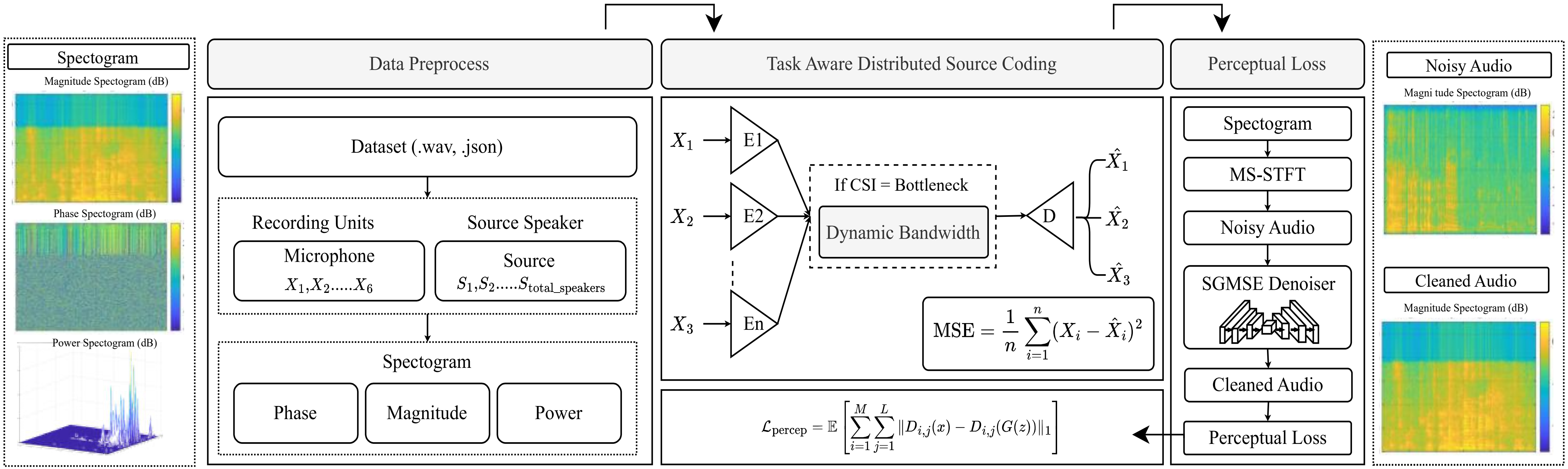
- Adopted from [Li et. al.]

Proposed Distributed Source Coding Pipeline

CSI-aware Dynamic Bitrate

NDPCA-aided Distributed Encoder

Task and Perception-aware Loss Function



Experiment Layout

• The experiment features multiple speakers engaged in conversational speech

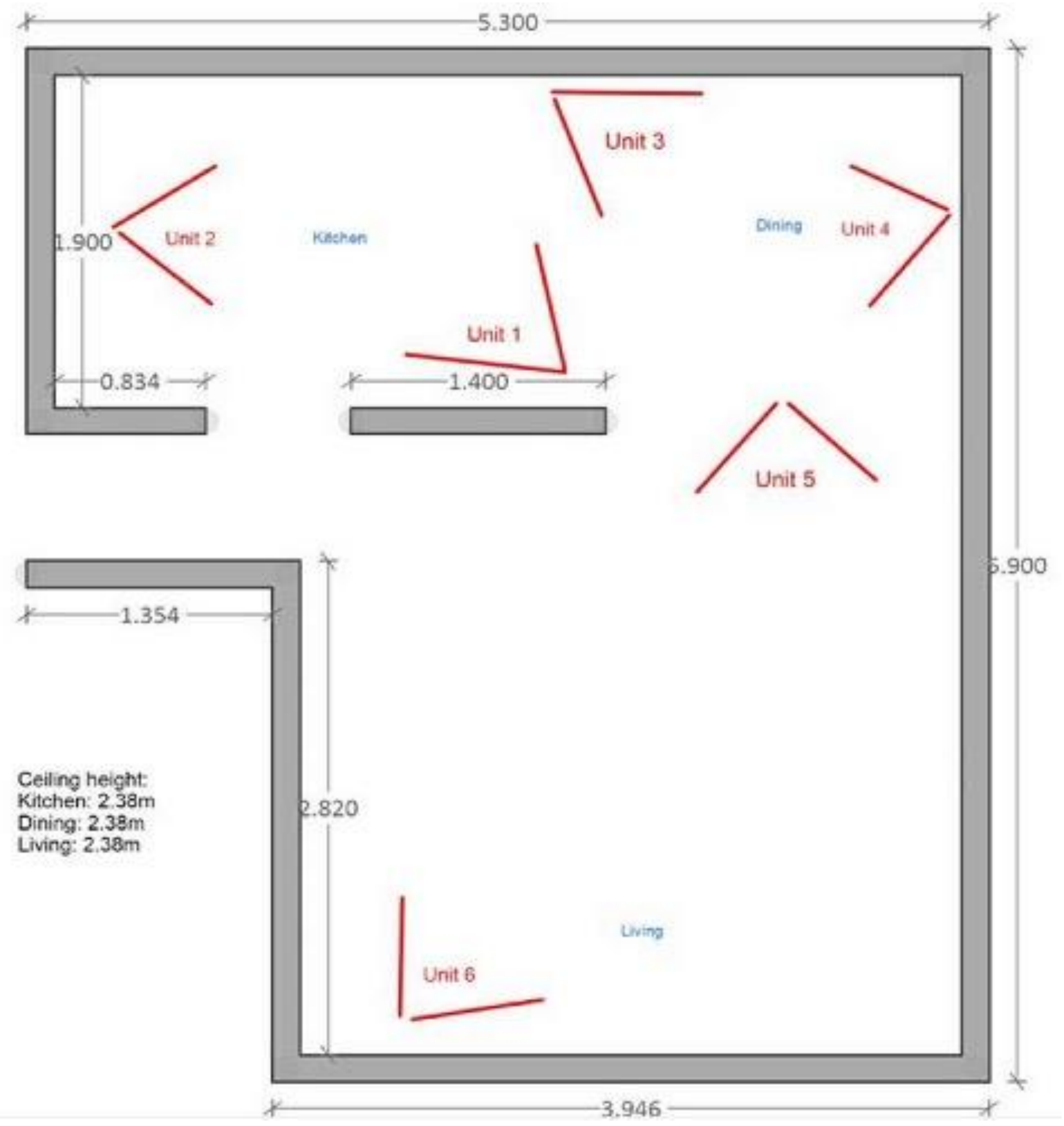
• Speakers are dispersed in the room, leading to overlapping and correlated audio signals.

• Six microphones are strategically placed around the environment to capture the speech from different vantage points.

• Each microphone's signal is partially correlated with the others because they record the same speakers from different positions.

- Local Encoding and Compression
  • Each microphone feed is passed through a local *NDPCA-aided* (Neural Distributed PCA) encoder.

- Bandwidth-Constrained Transmission
  • The compressed representations from each microphone are transmitted over bandwidth-limited channels.



Dataset and Preprocessing

Chime 6 Dataset

Consists of conversational speech audio with 8 speakers and 6 microphones spread across a room.

Audio Signal Preprocessing

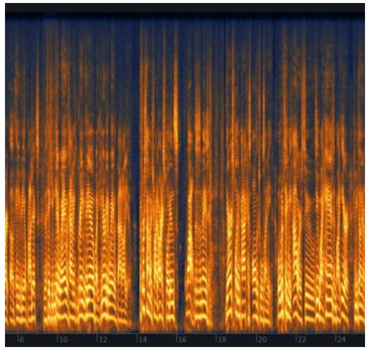
The audio signals were aligned, compensated for frame drops and clock skew, and distributed as WAV files with a sampling rate of 16 kHz.

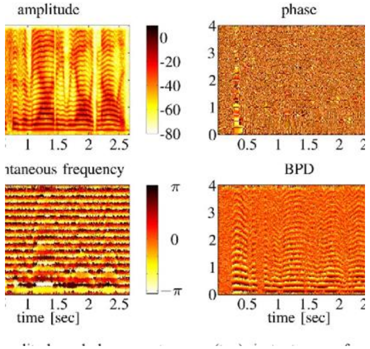
PKL Format Conversion

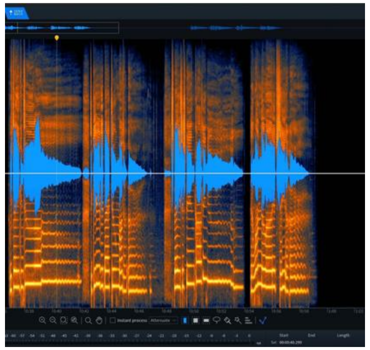
The extracted data was converted into .pkl format with magnitude, phase, and parameters across time.

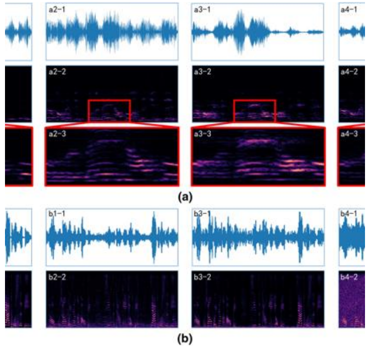
The Chime 6 dataset was preprocessed and converted into a PKL format to be used for further analysis and modeling.

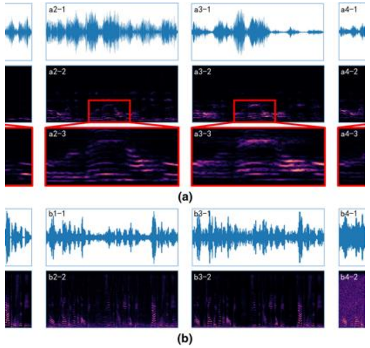
Spectrogram Distributions

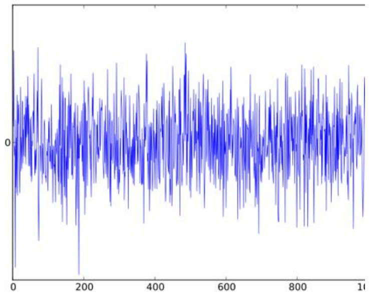












Clean Audio Magnitude Spectrogram

The clean audio magnitude spectrogram exhibits a clear time-frequency representation with consistent patterns, indicating the presence of tonal and structured features.

Clean Audio Phase Spectrogram

The clean audio phase spectrogram remains consistent with the underlying clean signal, demonstrating well-defined frequency components.

Clean Audio Power Spectrogram

The clean audio power spectrogram exhibits sharp and concentrated power peaks at specific frequencies, indicating well-defined tonal or harmonic components, with most of the power concentrated in the lower frequencies.

Noisy Audio Magnitude Spectrogram

The noisy audio magnitude spectrogram shows a broader distribution of energy across frequencies, with the magnitude reaching up to 20 dB, characteristic of added noise.

Noisy Audio Phase Spectrogram

The noisy audio phase spectrogram demonstrates less distinct frequency components, with additional energy spread across frequencies, indicating the presence of noise.

Noisy Audio Power Spectrogram

The noisy audio power spectrogram shows a broader distribution of power across the frequency spectrum, suggesting the presence of noise and increased power levels, especially in the high-frequency regions.

Task and Perception-aware Loss Function

Task-agnostic Loss

1. Mean Squared Error (MSE)
2. Cosine Similarity loss between NDPCA-encoded representation from all encoders,
3. Peak Signal-to-Noise Ratio (PSNR)

Downstream Speech Enhancement Task Loss

Pre-trained score-based Langevin diffusion model loss

(Diffusion model pre-trained for task, i.e., speech enhancement)

Perceptual Loss: Preserving Realism

Multi-Scale Short-Time Fourier Transform (MS-STFT) Discriminator to ensure the reconstructed speech sounds realistic.

The comprehensive loss function combines task-agnostic, task-specific, and perceptual objectives to achieve high-quality speech reconstruction and enhancement.

Baseline Models

The comprehensive loss function combines task-agnostic, task-specific, and perceptual objectives to achieve high-quality speech reconstruction and enhancement.

1. Joint Autoencoder (JAE)

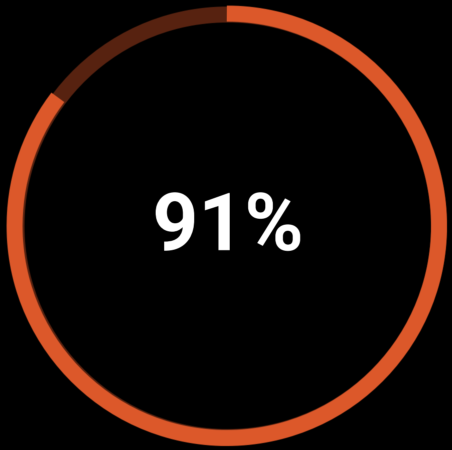
- Single Encoder, Single Decoder
  - All microphone signals are fused and fed into one large encoder, then reconstructed by one decoder.
- Upper Bound on Performance
  - Because it encodes all data at once, it can learn inter-microphone correlations directly.
  - Typically achieves the best possible (upper-bound) reconstruction quality.

2. Distributed E4D1 / E2D1

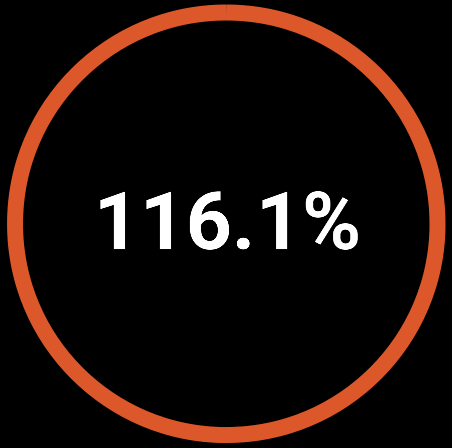
- Multiple Encoders, Single Decoder
  - Each microphone has its own local encoder (4 (2) encoders total, in the E4D1 (E2D1 setup).
  - The compressed streams then feed into a single decoder.
- Limited Exploitation of Correlation
  - Each encoder operates independently, so cross-microphone correlations are not fully leveraged.
  - Simpler to implement, but potentially less efficient in compression.

Results: Task-agnostic Settings

Relative PSNR values in task-agnostic settings (higher is better)



Proposed E2D1 / JAE Upper Bound



Proposed E2D1 / Baseline E2D1

Proposed E4D1 / Baseline E4D1

Results: Task and Perception-aware Settings

98.1%

Proposed E2D1 / JAE Upper Bound

146.9%

Proposed E2D1 / Baseline E2D1

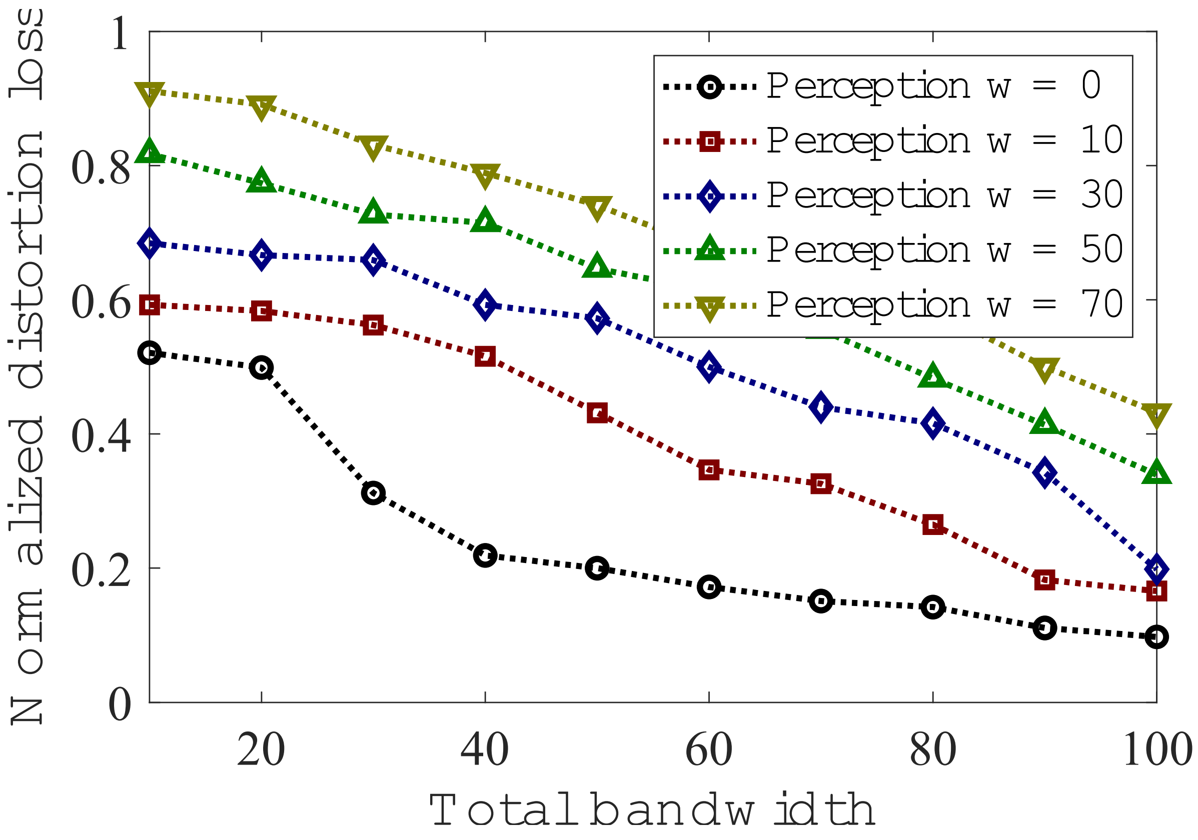
152.2%

Proposed E4D1 / Baseline E4D1

Rate-Distortion-Perception Trade-off

1. Higher weight assigned to perception loss leads to higher distortion at a given bandwidth

2. Optimizing for distortion minimization does not automatically lead to better perceptual quality – there is a distortion-perception trade-off under finite bandwidth



Conclusion

Leveraging Correlations

The NDPCA-aided architecture effectively leverages the correlations between the multiple speech sources to achieve better compression and reconstruction performance.

Significant Performance Gains

The proposed algorithm outperforms baseline distributed methods by 16.1% and 18.8% in PSNR for task-agnostic settings, and up to 52.2% in speech enhancement tasks.

Dynamic Bitrate Adaptation

The distributed PCA encoder allows for dynamic bitrate adaptation based on available wireless channel bandwidth, without the need for retraining the model.

Approaching Upper Bound

The NDPCA-aided algorithms approach the performance of the upper bound Joint E1D1 case, demonstrating the effectiveness of the proposed approach.

The proposed NDPCA-aided distributed source coding algorithm effectively addresses the challenges of transmitting correlated speech signals from multiple microphones, achieving significant performance improvements over baseline methods while enabling dynamic bitrate adaptation.