# Robust Automatic Speech Recognition By Integrating Speech Separation

**Peidong Wang**

**20210414**

# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

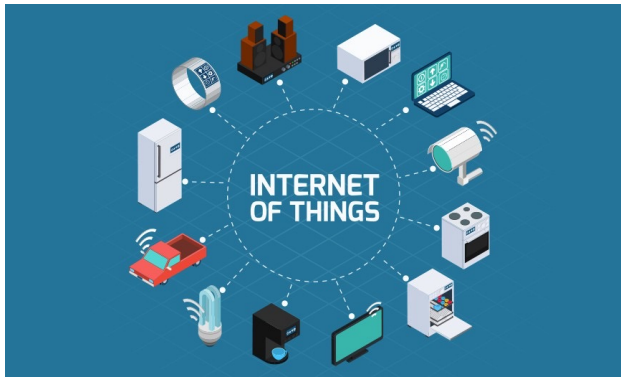
# Background

- **How should the physical world connect with the digital world?**
  - Mouse & keyboard


  - Touch screen

# Background

- **How should the physical world connect with the digital world? (cont'd)**
  - Speech interface?





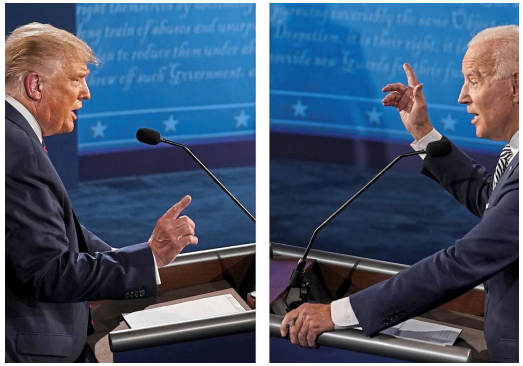
# Background

- **Automatic Speech Recognition (ASR)**
  - Convert speech to text

# Background

- **Robust ASR**
  - Noise

  - Multiple speakers



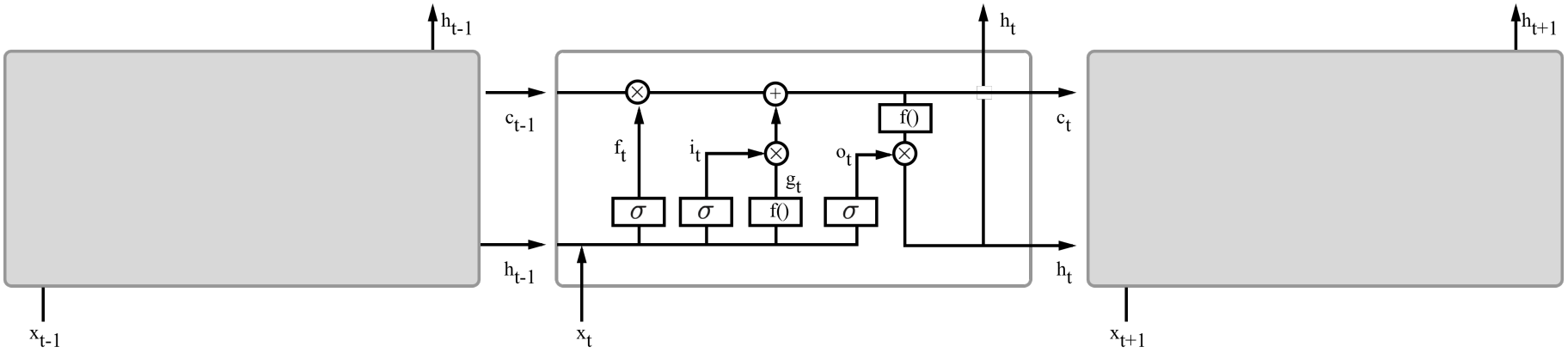
# Background

- **Robust ASR (cont'd)**
  - Based on acoustic environments:
    - Noisy speech
    - Overlapped speech
  - Based on the number of microphones:
    - Monaural (i.e. single-channel)
    - Multi-channel

# OUTLINE

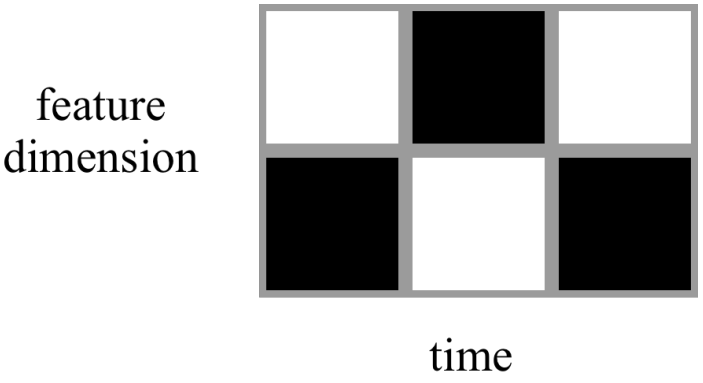
- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

- **Recurrent Neural Network Using LSTM Cells**
  - We use a long short-term memory (LSTM) based recurrent neural network (RNN) as the acoustic model
  - $\boldsymbol{x}_t$ and $\boldsymbol{h}_t$ denote the input and hidden vectors at step t; $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ are the input, forget, and output gates at step $t$; $\boldsymbol{g}_t$ is the vector of cell updates and $\boldsymbol{c}_t$ is the cell vector; $\sigma$ is the sigmoid function, $\otimes$ is the element-wise multiplication, and $f()$ is typically chosen to be tanh()



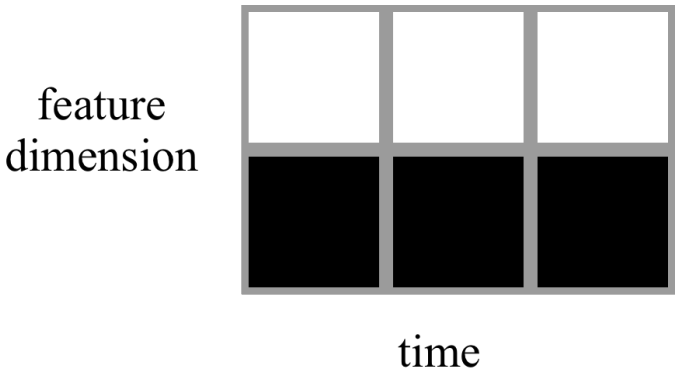
- **Conventional Dropout Method for RNNs**
  - Dropout functions for RNNs are sampled independently at each frame

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma(\mathbf{W}_i d_{xit}(\mathbf{x}_t) + \mathbf{U}_i d_{hit}(\mathbf{h}_{t-1})) \\ \sigma(\mathbf{W}_f d_{xft}(\mathbf{x}_t) + \mathbf{U}_f d_{hft}(\mathbf{h}_{t-1})) \\ \sigma(\mathbf{W}_o d_{xot}(\mathbf{x}_t) + \mathbf{U}_o d_{hot}(\mathbf{h}_{t-1})) \\ f(\mathbf{W}_g d_{xgt}(\mathbf{x}_t) + \mathbf{U}_g d_{hgt}(\mathbf{h}_{t-1})) \end{pmatrix}$$

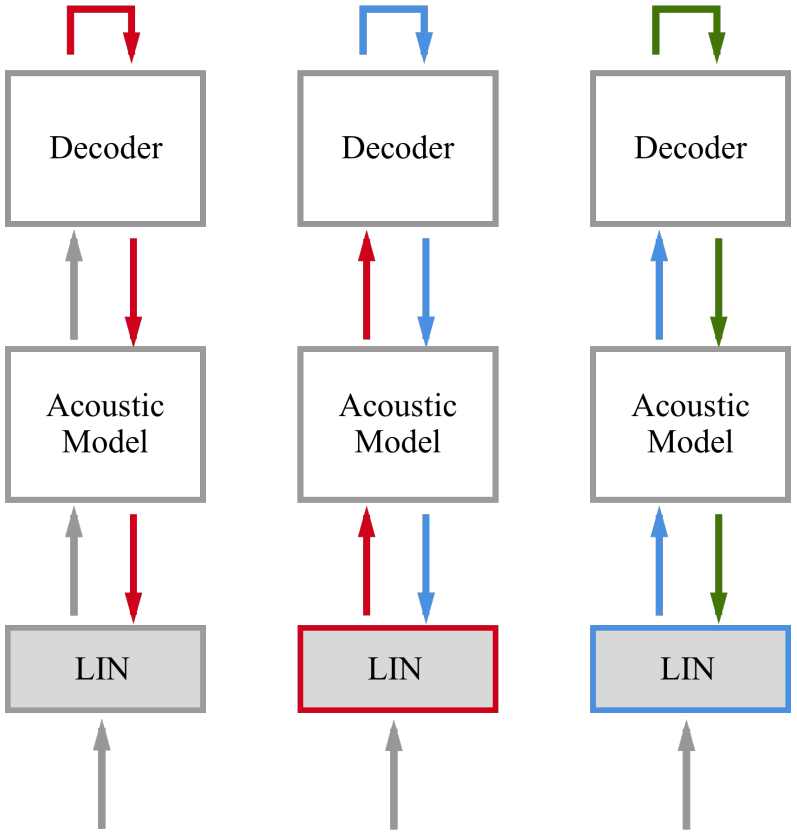


- **Utterance-Wise Recurrent Dropout for RNNs**
  - We propose utterance-wise recurrent dropout function, which is shared across different frames, for the hidden vectors in LSTMs
  - It aims to better exploit utterance-level information

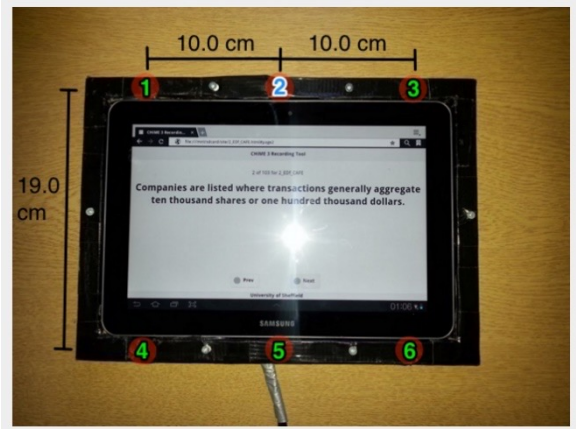
$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma(\mathbf{W}_i d_{xit}(\mathbf{x}_t) + \mathbf{U}_i d_{hi}(\mathbf{h}_{t-1})) \\ \sigma(\mathbf{W}_f d_{xft}(\mathbf{x}_t) + \mathbf{U}_f d_{hf}(\mathbf{h}_{t-1})) \\ \sigma(\mathbf{W}_o d_{xot}(\mathbf{x}_t) + \mathbf{U}_o d_{ho}(\mathbf{h}_{t-1})) \\ f(\mathbf{W}_g d_{xgt}(\mathbf{x}_t) + \mathbf{U}_g d_{hg}(\mathbf{h}_{t-1})) \end{pmatrix}$$



- **Iterative Speaker Adaptation**
  - Using decoding results as labels, the acoustic model can be adapted to specific speakers on the test set, attenuating the mismatch between training and test data
  - We apply unsupervised linear input network (LIN) based speaker adaptation
  - Iterate the speaker adaptation process by using the newly generated decoding result as the label for another adaptation iteration



- **Experimental Setup**
  - CHiME-4 corpus
  - Utterances are recorded by 6 microphones on a tablet
  - Training set:
    - 6 channels
    - Each channel contains 1600 real recorded utterances and 7138 simulated utterances
  - Test set:
    - 2640 utterances
    - Half real recorded and half simulated
  - Recording environments:
    - bus, cafeteria, pedestrian area, and street conjuncture





* [Images From the CHiME-4 website] *http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/overview.html*

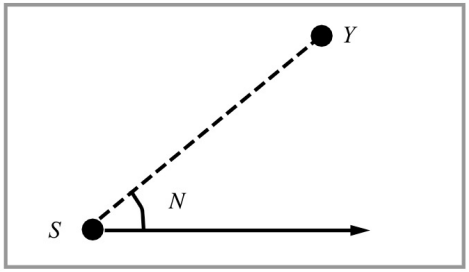
- **Evaluation Results**
  - *Baseline* and *Unconstrained* denote the baseline RNN language model and unconstrained language models, respectively
  - Our model outperforms the previous best model using the baseline RNN language model by 16% relatively in word error rate (WER)
  - It is even better than the best model using an unconstrained language model by 10% relatively

| systems | Baseline | | Unconstrained | |
|---|---|---|---|---|
| | simu | real | simu | real |
| Du *et al.*'16 | 13.62 | 11.15 | 11.81 | **9.15** |
| Heymann *et al.*'16 | 11.68 | **9.88** | 11.11 | 9.34 |
| Proposed | 11.14 | **8.28** | - | - |

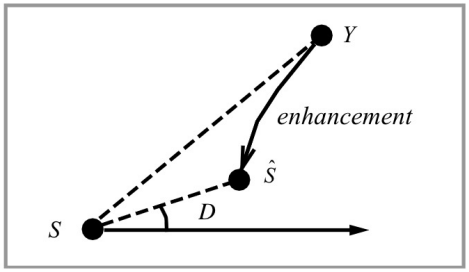
# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

- **The Distortion Problem**
  - Monaural speech enhancement may be useless or even harmful for robust ASR due to the introduction of distortions to speech signals
  - Compared with noisy speech, enhanced speech has a higher signal-to-noise ratio (SNR), but the noise type may be different

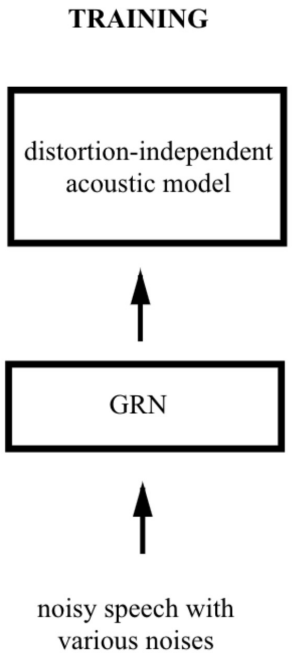



(a)

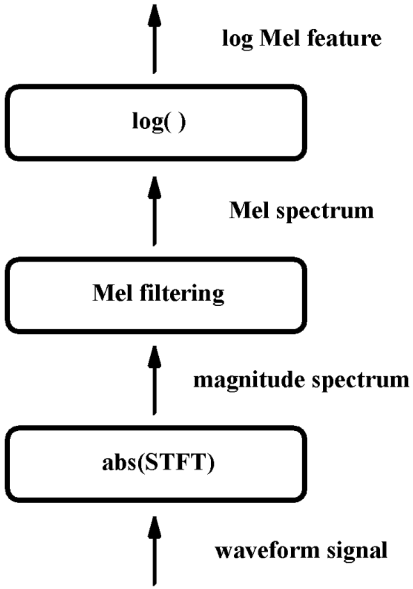



(b)

- **Distortion-Independent Acoustic Modeling**
  - Train the acoustic model using various types of enhanced speech
  - The enhanced speech is generated using noisy speech that contains various types of noises
  - GRN: gated residual network, a speech enhancement model



- **Enhanced Spectral Features for Distortion-Independent Acoustic Modeling**
  - ASR models typically use enhanced waveform signals as the input. We investigate various enhanced spectral features
  - Magnitude spectrum: skip the speech re-synthesis step in speech enhancement and part of the feature extraction step in ASR



- **Experimental Setup**
  - CHiME-2 corpus, whose training set includes 7138 reverberated utterances from 83 speakers and the test set comprises 330 noisy utterances from 12 other speakers
  - A noise database containing 10000 noises
- **Models**
  - The ASR backend is the same as the one described in the previous section

- **Evaluation Results**
  - Noise-dependent: the model is trained using only one type of noise and is tested on the same type of noise
  - Monaural speech enhancement is harmful for conventional noise-dependent acoustic model
  - Distortion-independent acoustic model benefits from enhanced speech, showing its ability to alleviate the distortion problem

| SNR | noise-dependent | | distortion-independent | |
|---|---|---|---|---|
| | w/o | w/ | w/o | w/ |
| 9dB | **5.49** | 5.81 | 7.42 | **4.54** |
| 6dB | **6.26** | 7.98 | 8.61 | **5.45** |
| 3dB | **6.78** | 8.33 | 10.01 | **6.20** |
| 0dB | **8.95** | 11.26 | 12.93 | **7.92** |
| -3dB | **9.98** | 14.48 | 14.85 | **9.43** |
| -6dB | **14.83** | 19.05 | 21.80 | **13.11** |
| avg | **8.7** | 11.2 | 12.6 | **7.8** |

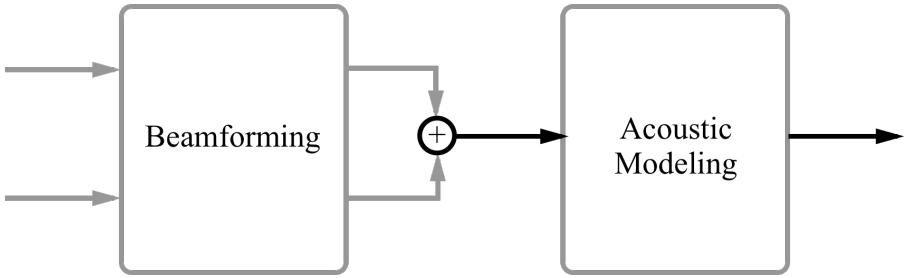
- **Evaluation Results (cont'd)**
  - We achieved 16% relative improvement over the previous best on CHiME-2

| systems | 9dB | 6dB | 3dB | 0dB | -3dB | -6dB | avg |
|---|---|---|---|---|---|---|---|
| Wang and Wang'16 | 6.61 | 6.86 | 8.67 | 10.39 | 13.02 | 18.23 | 10.6 |
| Plantinga *et al.*'18 | - | - | - | - | - | - | 9.3 |
| magnitude features + distortion-independent acoustic modeling | **4.54** | **5.45** | **6.20** | **7.92** | **9.43** | **13.11** | **7.8** |

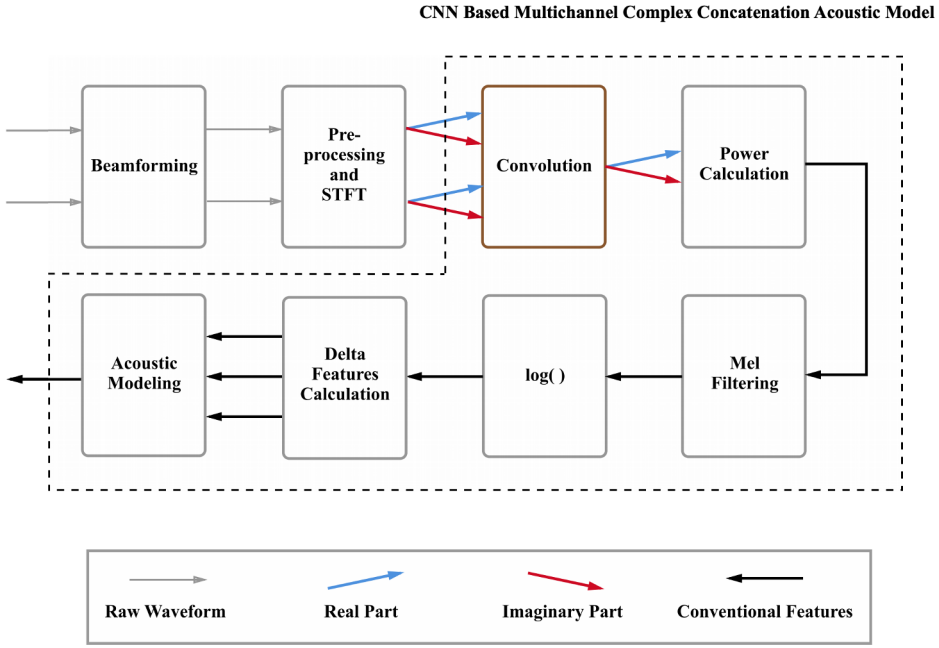
# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

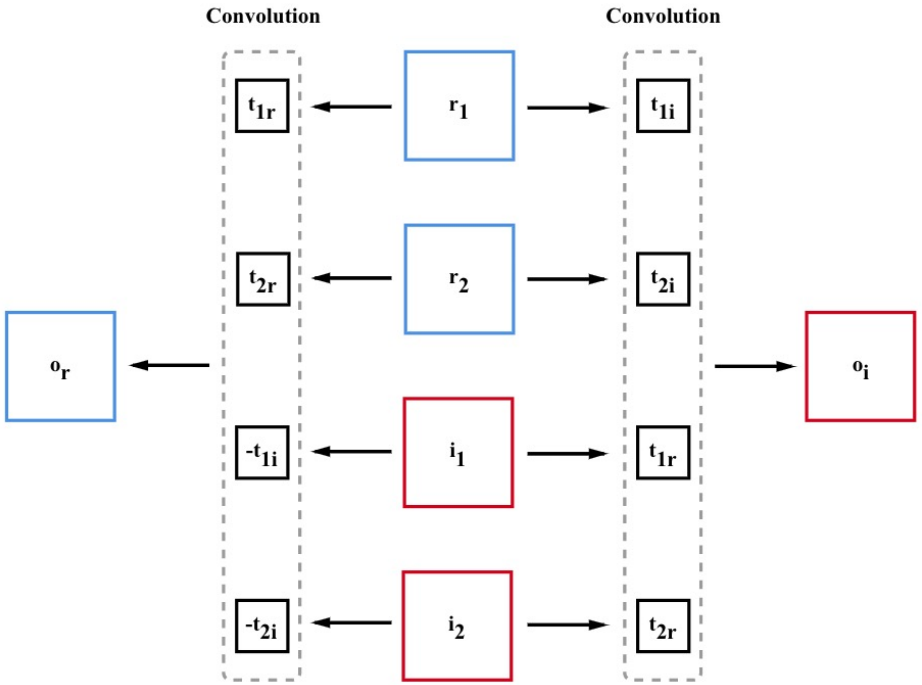
- **Conventional Method**
  - Multi-channel speech enhancement frontends sum the multiple spatially filtered signals into one channel
  - ASR uses the beamformed single-channel signal as input



- **Filter-and-Convolve: A CNN Based Complex Concatenation Acoustic Model**
  - Instead of summing the spatially filtered signals, we combine them using a learnable complex domain convolutional neural network (CNN)
  - STFT: short-time Fourier transform
  - Internal acoustic model: the ASR system described in the ASR Model section



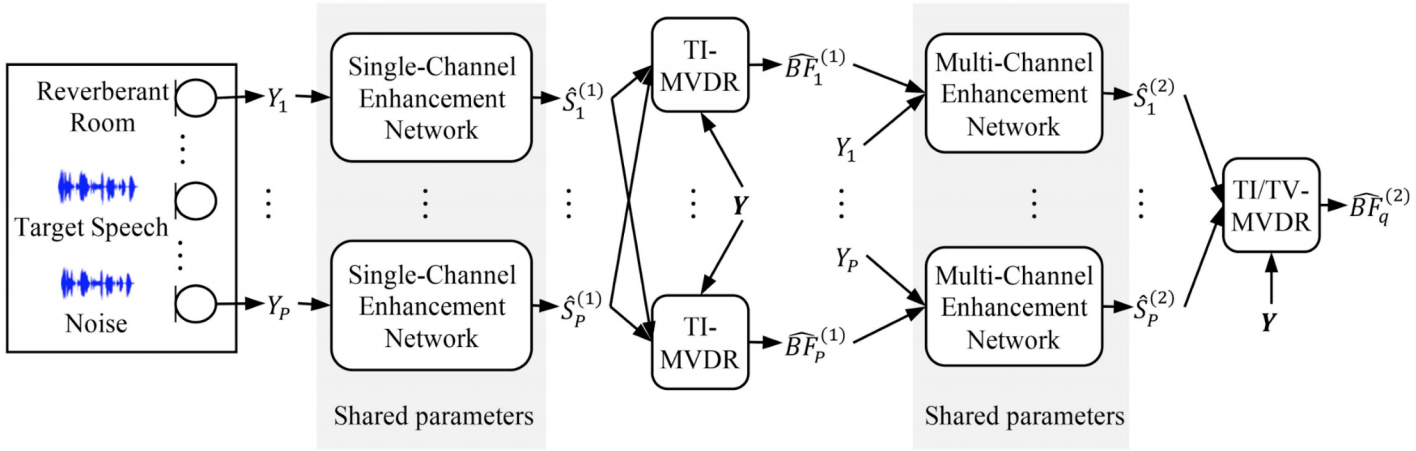
- **Filter-and-Convolve: A CNN Based Complex Concatenation Acoustic Model (cont'd)**
  - Complex-domain CNN is performed by dividing the input into real and imaginary parts
  - Experimental setup: the standard CHiME-4 corpus



- **Evaluation Results:**
  - With the learnable CNN layer, the results of BeamformIt and a minimum variance distortionless response (MVDR) beamformer are improved by 13% and 10% relatively

| Beamformer | BeamformIt | MVDR |
|---|---|---|
| Previous Best | 6.2 | 4.4 |
| Proposed | **5.4** | **4.0** |

- **Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR**
  - Two stage beamforming: first, extract spatial features; second, use spatial features together with spectral features as the input to the second speech enhancement model
  - Speech enhancement is performed in the complex domain



- **Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR (cont'd)**
  - The input and output to the speech enhancement model are both divided into real and imaginary parts
  - The loss function is a summation of real-imaginary loss and magnitude loss:

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \left\| \sqrt{\hat{R}_p^2 + \hat{I}_p^2} - |S_p| \right\|_1$$

$$\mathcal{L}_{\text{RI}} = \left\| \hat{R}_p - \text{Real}\,(S_p) \right\|_1 + \left\| \hat{I}_p - \text{Imag}\,(S_p) \right\|_1$$

- **Speech Enhancement Sound Demos**
  - The utterances are from CHiME-4 and are recorded on a bus
  - Mixture:


  - Six-channel beamformed:


- **Evaluation Results: Two-Channel**
  - Achieved a WER of 3.19% on the CHiME-4 two-channel evaluation set
  - Outperform the previous best by 18% relatively

| Approaches | Dev. Set | | Test Set | |
|---|---|---|---|---|
| | Simu. | Real | Simu. | Real |
| $\hat{BF}_q^{(2)}$ + Trigram | 5.32 | 5.03 | 6.85 | 7.72 |
| + Five-gram and RNNLM | 3.74 | 3.32 | 4.84 | 5.54 |
| + LSTMLM | 2.52 | 2.15 | 3.28 | 3.80 |
| + Iterative Speaker Adaptation | **2.17** | **1.99** | **2.53** | **3.19** |
| Baseline'16 | 3.94 | 2.85 | 5.03 | 5.40 |
| Du *et al.*'16 | 3.46 | 2.33 | 5.74 | 3.91 |

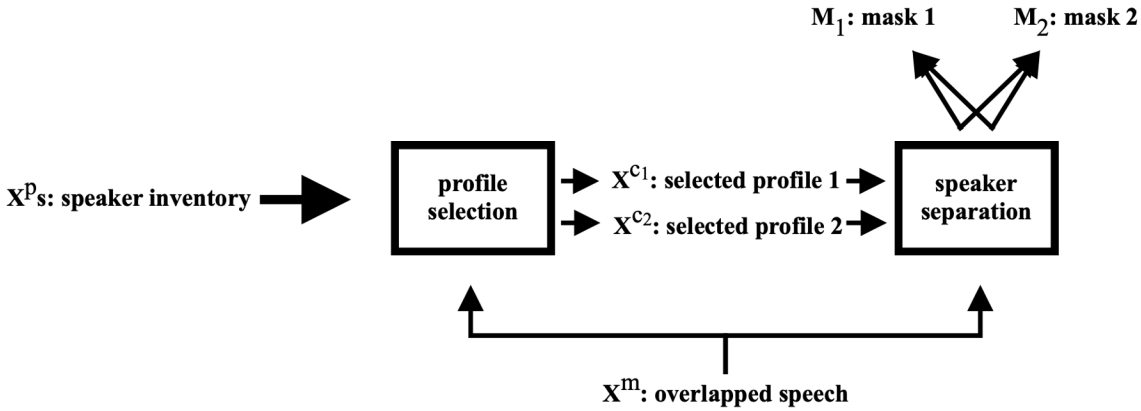
- **Evaluation Results: Six-Channel**
  - Achieved a WER of 1.99% on the CHiME-4 six-channel evaluation set
  - Outperform the previous best by 11% relatively

| Approaches | Dev. Set | | Test Set | |
|---|---|---|---|---|
| | Simu. | Real | Simu. | Real |
| $\hat{BF}_q^{(2)}$ + Trigram | 3.46 | 3.99 | 4.10 | 5.06 |
| + Five-gram and RNNLM | 2.28 | 2.57 | 2.71 | 3.59 |
| + LSTMLM | 1.39 | 1.66 | 1.76 | 2.32 |
| + Iterative Speaker Adaptation | **1.15** | **1.50** | **1.45** | **1.99** |
| Baseline'16 | 2.10 | 1.90 | 2.66 | 2.74 |
| Du *et al.*'16 | 1.78 | 1.69 | 2.12 | 2.24 |

# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

- **Speaker Separation Using Speaker Inventories (SSUSI)**
  - SSUSI uses two modules to perform speaker separation
  - The profile selection system selects relevant profiles from the speaker inventory
  - The speaker separation system uses selected profiles as additional information to separate the overlapped speech

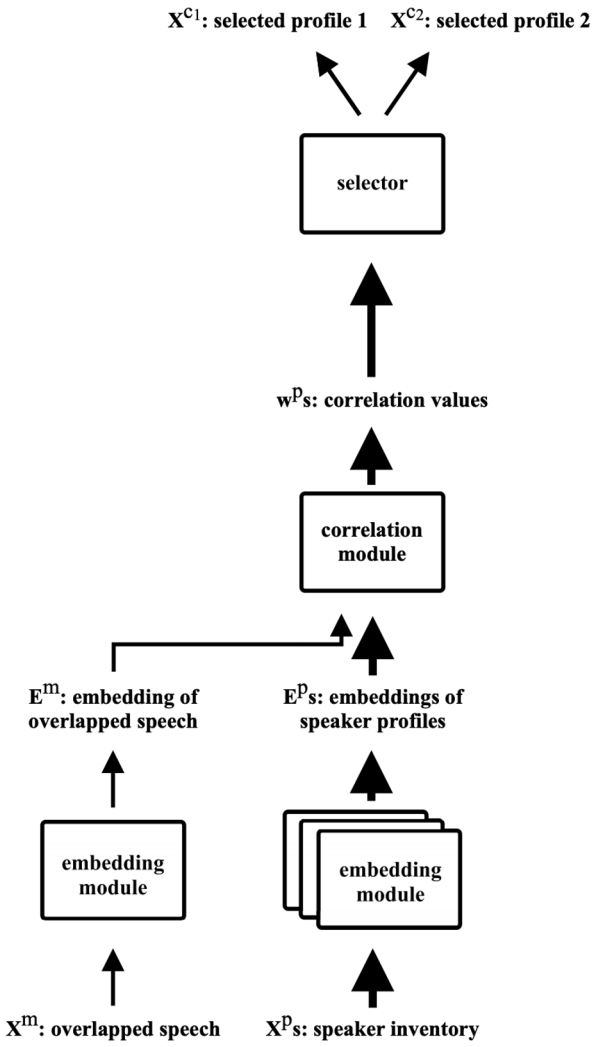


- **Profile Selection System**
  - The correlation module measures the correlations between the embedding of overlapped speech and those of speaker profiles
  - We denote the vector in $\boldsymbol{E}^m$ at time $i$ as $e_i^m$ and that in $\boldsymbol{E}^p$ at time $j$ as $e_j^p$ . The operations in the correlation module:

$$d_{i,j}^p = \boldsymbol{e}_i^m \cdot \boldsymbol{e}_j^p$$

$$w_{i,j}^p = \frac{\exp(d_{i,j}^p)}{\sum_{p \in \mathbf{P}} \sum_{j=1}^{T_p} \exp(d_{i,j}^p)}$$

  - Note that the denominator is a summation over both profile time steps $j$ and profiles $p$



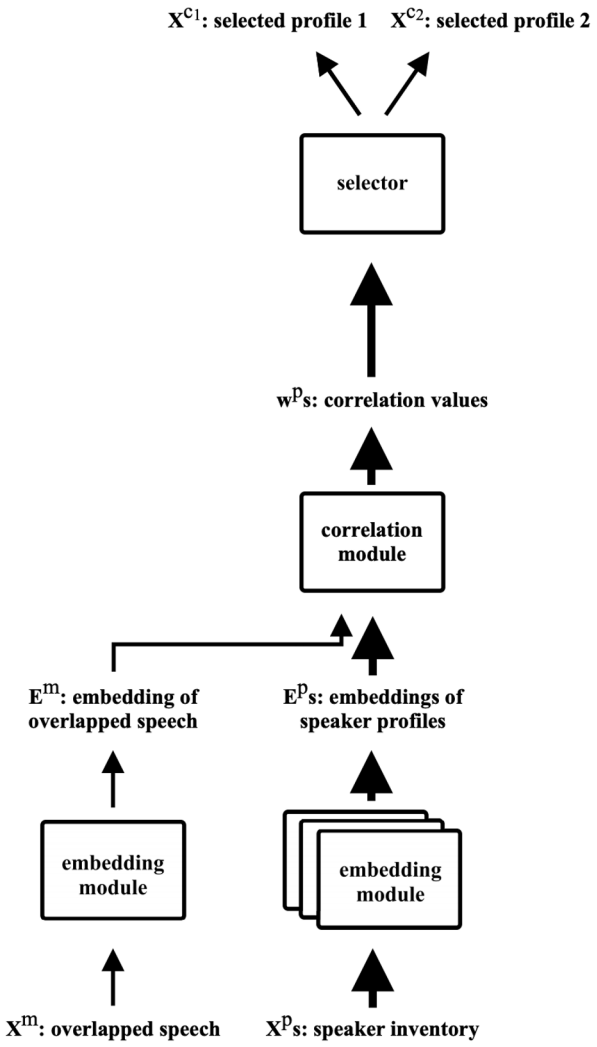
- **Profile Selection System (cont'd)**
  - The profile selector then calculates the average weight for each profile and selects two that have the first and second largest weights as the relevant profiles for the speaker separation system
  - Calculate average weight:

$$w^p = \frac{\sum_{i=1}^{T_m} \sum_{j=1}^{T_p} w_{i,j}^p}{T_m T_p}$$

  - Select relevant profiles:

$$c_1 = \arg\max_{p \in \mathbf{P}} \{w^p\}$$

$$c_2 = \arg\max_{p \in \mathbf{P} - \{c_1\}} \{w^p\}$$



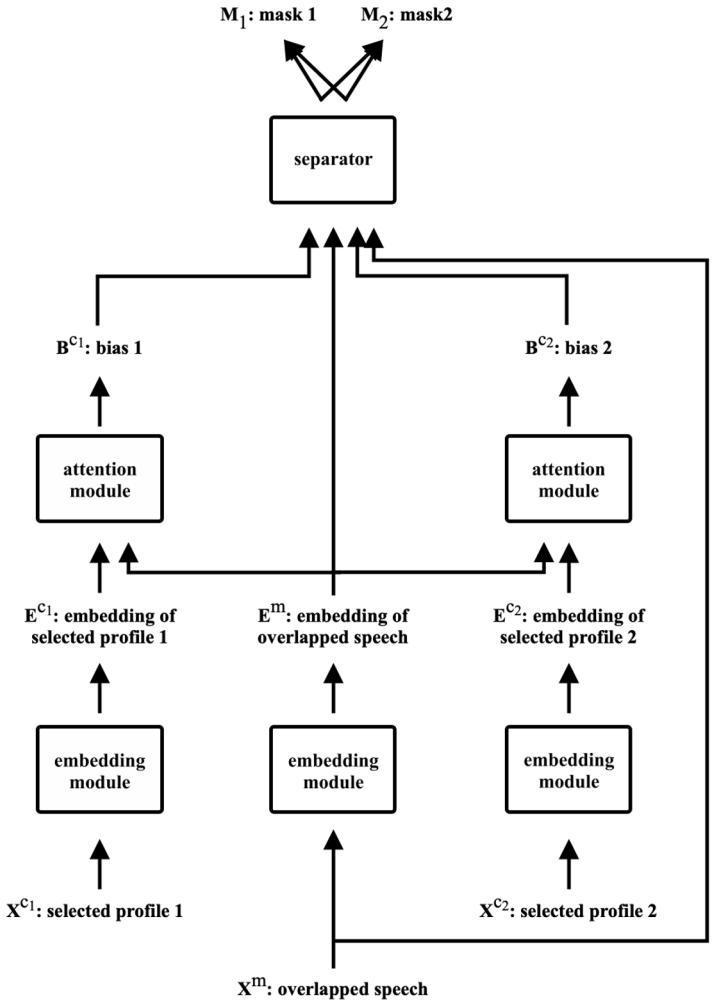
- **Speaker Separation System**
  - The attention mechanism for speaker separation is similar to the correlation calculation for profile selection:

$$d_{i,j}^{c_1} = \boldsymbol{e}_i^m \cdot \boldsymbol{e}_j^{c_1}$$

$$\alpha_{i,j}^{c_1} = \frac{\exp(d_{i,j}^{c_1})}{\sum_{j=1}^{T_{c_1}} \exp(d_{i,j}^{c_1})}$$

$$\boldsymbol{b}_i^{c_1} = \sum_{j=1}^{T_{c_1}} \alpha_{i,j}^{c_1} \boldsymbol{e}_j^{c_1}$$

  - Note that the attention matrix element softly aligns the embeddings of relevant profiles to that of the overlapped speech



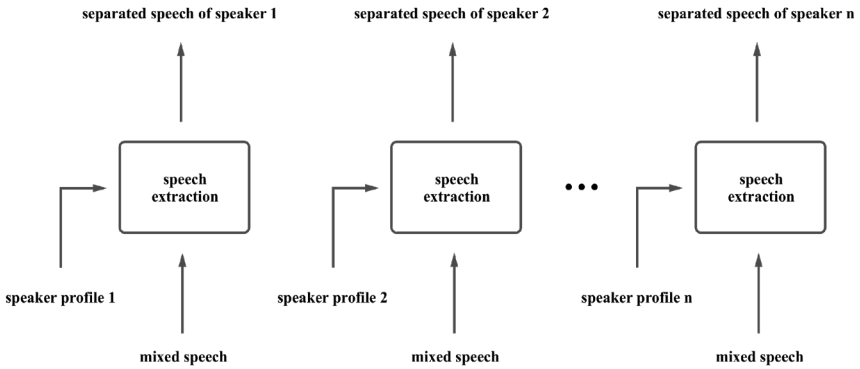
- **Experimental Setup**
  - LibriSpeech corpus
  - We generate the training set using the two clean training sets (i.e. train-clean-100 and train-clean-360) in LibriSpeech
  - At test time, the mixed speech is generated using the clean test set
  - There are 1172 speakers in the training set and 40 other speakers in the test set

- **Models**
  - We evaluate the ASR performance of separated speech using an acoustic model trained on the clean sets of LibriSpeech

- **Evaluation Results: Comparisons Between SSUSI and PIT**
  - SSUSI performs significantly better than permutation invariant training (PIT), which does not use speaker information
  - Even when there are 30 irrelevant profiles, SSUSI still yields a signal to distortion ratio (SDR) of 10.8 dB
  - For WERs, SSUSI outperforms PIT by 48% relatively when there is no irrelevant profile
  - In the case when there are 30 irrelevant profiles, the relative improvement is still 34%

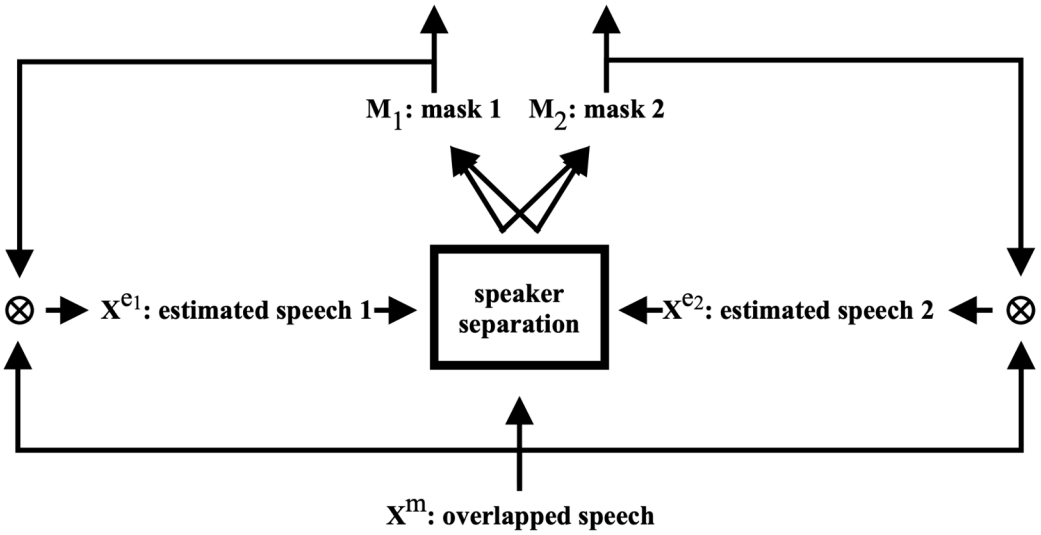
| method | # ir-profiles | SDR (dB) | WER (%) |
|---|---|---|---|
| PIT | - | 8.7 | 36.5 |
| SSUSI | 0 | 12.2 | 19.1 |
| | 6 | 11.5 | 21.8 |
| | 22 | 11.0 | 23.4 |
| | 30 | 10.8 | 24.1 |

- **Evaluation Results: Comparisons Between SSUSI and Speech Extraction**
  - Speech extraction generates a stream of separated speech for each candidate speaker
  - SSUSI achieves an improvement of more than 0.7 dB in SDR over speech extraction
  - For WER, the overall relative improvement is over 13%
  - In addition to separation accuracy, SSUSI improves the efficiency over the speech extraction system significantly
  - If there are 30 irrelevant profiles, the computation time reduction is about 70%




| method | # ir-profiles | SDR (dB) | WER (%) |
|---|---|---|---|
| Speech Extraction | 0 | 11.5 | 21.9 |
| | 1 | 11.1 | 23.3 |
| | 2 | 10.9 | 24.4 |
| SSUSI | 0 | 12.2 | 19.1 |
| | 1 | 12.0 | 19.9 |
| | 2 | 11.9 | 20.4 |

- **Speaker Separation Using Estimated Speech (SSUES)**
  - SSUES uses separated speech as speaker profiles for more iterations of speaker separation
  - The experimental setup is the same as that of SSUSI



- **Evaluation Results: SSUES**
  - With 30 irrelevant profiles, the SDR improvement of SSUSI + SSUES is 0.9 dB and the WER reduction is 16% relatively
  - Note that the SDR and WER results of SSUSI + SSUES match those of SSUSI with 2 irrelevant profiles (11.9 dB and 20.4%)
  - SSUES significantly improves PIT as well
  - For PIT, the WER improvement is 37% relatively

| method | # iter | SDR (dB) | WER (%) |
|---|---|---|---|
| SSUSI | - | 10.8 | 24.1 |
| + SSUES | 1 | 11.4 | 21.1 |
| | 2 | 11.6 | 20.4 |
| | 3 | 11.7 | 20.3 |

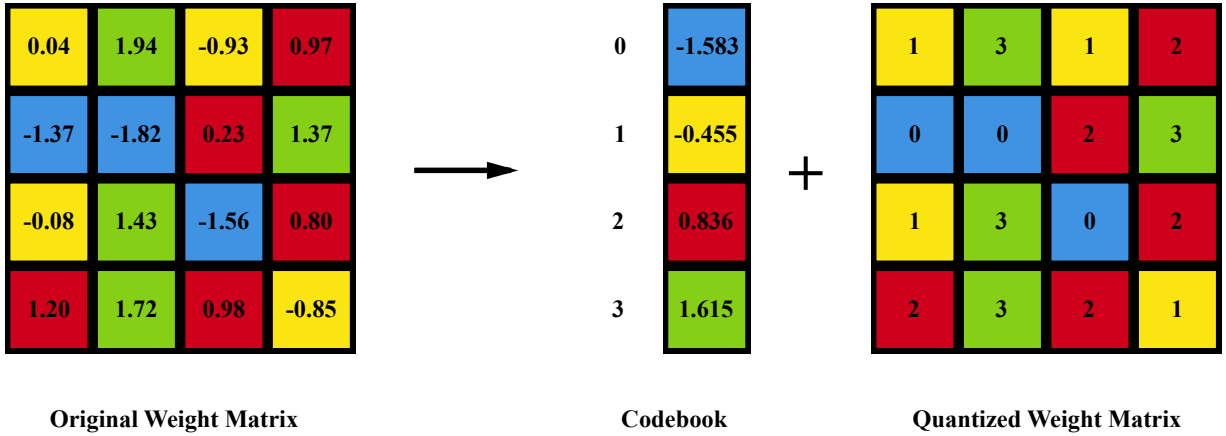
| method | # iter | SDR (dB) | WER (%) |
|---|---|---|---|
| PIT | - | 8.7 | 36.5 |
| + SSUES | 1 | 10.5 | 24.8 |
| | 2 | 10.8 | 23.2 |
| | 3 | 10.9 | 22.9 |

# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

- **Conventional Method: Weight Sharing**
  - Weight sharing clusters the weights in the matrix
  - It then uses the mean value of the cluster to replace the original weight values
  - The results are represented using a codebook and a quantized weight matrix (containing the indices of each weight in the codebook)

  - Weight sharing does not need to retrain the model

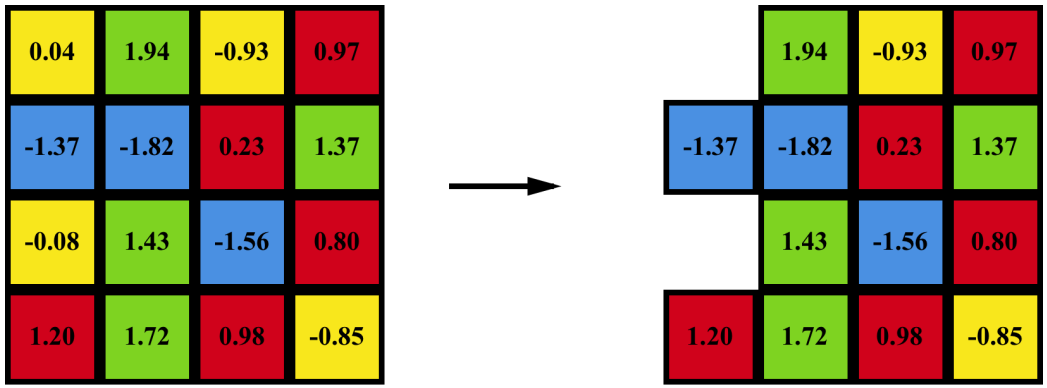


- **Limitation of Weight Sharing**
  - The table below is a comparison between 5-bit and 4-bit weight sharing of a conformer (a state-of-the-art end-to-end ASR model) on LibriSpeech
  - 5-bit corresponds to 32 clusters, and 4-bit is 16 clusters

  - The conformer can be compressed using 5-bit weight sharing without significant performance degradation
  - Its performance drops dramatically using 4-bit weight sharing

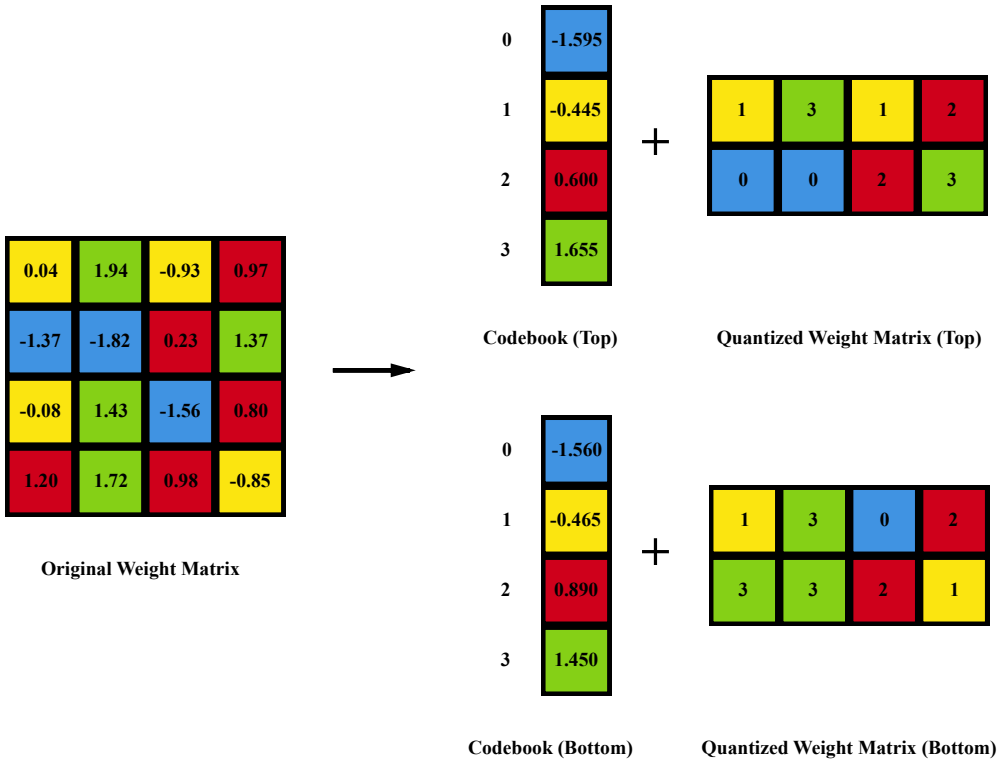
| Model | Size (MB) | WER | | | |
|---|---|---|---|---|---|
| | | dev-clean | dev-other | test-clean | test-other |
| conformer | 42.76 | 2.4 | 6.3 | 2.7 | 6.5 |
| 5-bit | 7.29 (6x) | 2.6 | 6.4 | 2.8 | 6.6 |
| 4-bit | 5.95 (7x) | 6.4 | 10.6 | 7.7 | 11.3 |

- **Proposed Method 1: Pruning without Retraining**
  - Remove the weights whose absolute values are close to zero but do not perform model retraining
  - The weights whose absolute values are close to zero may have a small impact on model performance. However, such small values may lead to bad clustering results since conventional weight sharing tries to cluster all weights
  - The dropout function used during training also ensures that removing part of the weights will not strongly influence the model performance

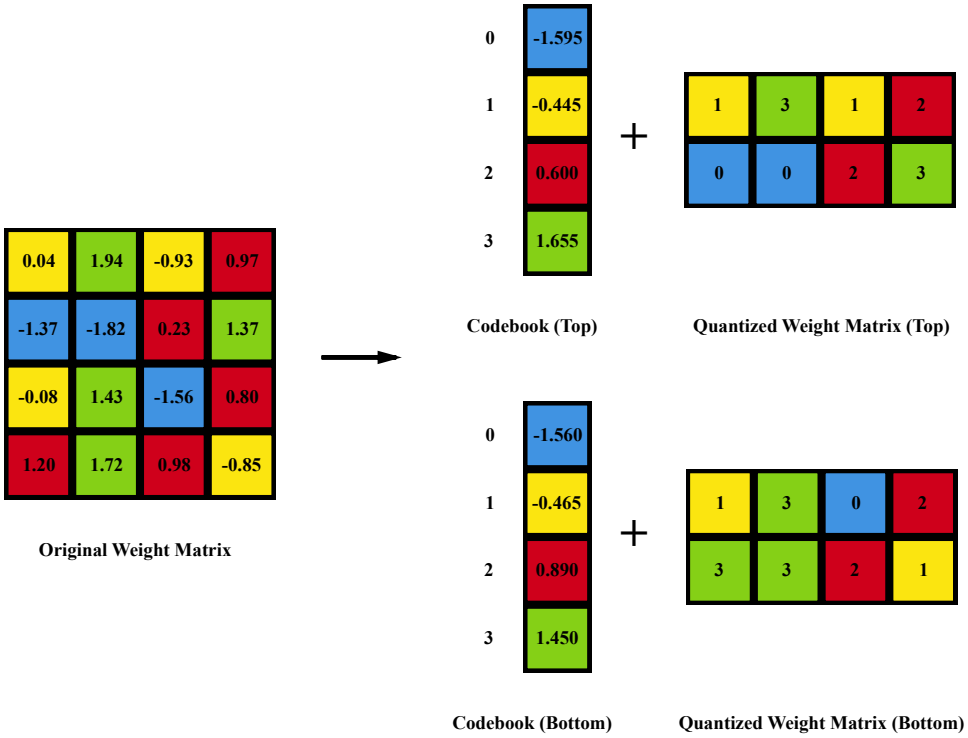



**Original Weight Matrix** **Pruned Weight Matrix**

- **Proposed Method 2: Submatrix Weight Sharing**
  - In the conformer model compressed by 4-bit weight sharing, the compressed model size is 5.95MB, in which only 0.033MB (i.e. 0.6%) corresponds to codebooks
  - Trade a small increase in codebook size for an improvement in performance



- **Proposed Method 2: Submatrix Weight Sharing (cont'd)**
  - Divide the original weight matrix into submatrices, perform weight sharing on each submatrix separately
  - Since the total number of elements that need to be clustered is smaller, the centroids of submatrix weight sharing may be closer to the original values



- **Proposed Method 3: Grid Search for Sensitivity Analysis**
  - Instead of a shared compression rate for all weights in the model, each weight matrix may need a different compression rate
  - Sensitivity analysis measures the sensitivity of each weight matrix to the model performance

  - Conventional sensitivity analysis is performed layer-by-layer
  - Grid search for sensitivity analysis is performed for all layers independently, resulting in a complete search space of all possible compression rate conditions

- **Experimental Setup**
  - Our experiments are conducted on the whole LibriSpeech corpus
  - The training set contains 960 hours of read English speech
  - All the training and evaluation pipelines are used the same way as the official recipe

  - We use a conformer as the ASR model
  - The number of parameters in the original model is about 10M

- **Evaluation Results: Submatrix Weight Sharing**
  - The number of submatrices in each weight matrix is 4
  - The proposed method improves the WER of 4-bit weight sharing by 52% relatively with only 0.1MB (i.e. 1.7%) increase in model size

| Model | Size (MB) | WER | | | |
|---|---|---|---|---|---|
| | | dev-clean | dev-other | test-clean | test-other |
| conformer | 42.76 | 2.4 | 6.3 | 2.7 | 6.5 |
| 5-bit | 7.29 (6x) | 2.6 | 6.4 | 2.8 | 6.6 |
| 4-bit | 5.95 (7x) | 6.4 | 10.6 | 7.7 | 11.3 |
| 4-bit (s) | 6.05 (7x) | 3.4 | 7.3 | 3.7 | 7.6 |

- **Evaluation Results**
  - Pruning without retraining not only significantly improves the WER (by 61% relatively) but also reduces the model size
  - Pruning and submatrix weight sharing can be combined to improve the WER: 4-bit (ps) achieves the same WER as 5-bit weight sharing on test-clean
  - Using grid search for sensitivity analysis, the model size can be further reduced to below 5MB (i.e. a 9-fold compression) with negligible performance degradation

| Model | Size (MB) | WER | | | |
|---|---|---|---|---|---|
| | | dev-clean | dev-other | test-clean | test-other |
| conformer | 42.76 | 2.4 | 6.3 | 2.7 | 6.5 |
| 5-bit | 7.29 (6x) | 2.6 | 6.4 | 2.8 | 6.6 |
| 4-bit | 5.95 (7x) | 6.4 | 10.6 | 7.7 | 11.3 |
| 4-bit (p) | 5.36 (8x) | 2.7 | 6.6 | 3.0 | 7.0 |
| 4-bit (ps) | 5.46 (8x) | 2.6 | 6.5 | 2.8 | 6.9 |
| 4-bit (psg) | 4.83 (9x) | 2.6 | 6.6 | 2.8 | 6.8 |

# OUTLINE

- **Background**
- **Robust ASR by Integrating Speech Separation**
  - **ASR Model**
  - **Monaural Speech Enhancement**
  - **Multi-Channel Speech Enhancement**
  - **Speaker Separation**
  - **ASR Model Compression**
- **Concluding Remarks**

- **Concluding Remarks**
  - Systematically approached robust ASR problem by integrating speech separation (including speech enhancement and speaker separation)
  - Proposed methods for single- and multi-channel speech enhancement and speaker separation for robust ASR
  - Proposed methods to compress ASR models
  - Advanced the state-of-the-arts on multiple evaluation corpora, including CHiME-2 and CHiME-4
  - Paved the road towards a better interaction between the physical world and the digital world

# Questions?

*Thank You!*