30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

NIH National Institutes of Health
Turning Discovery Into Health

Separation of Emotional and Reconstruction Embeddings on Ladder Network to Improve Speech Emotion Recognition Robustness in Noisy Conditions

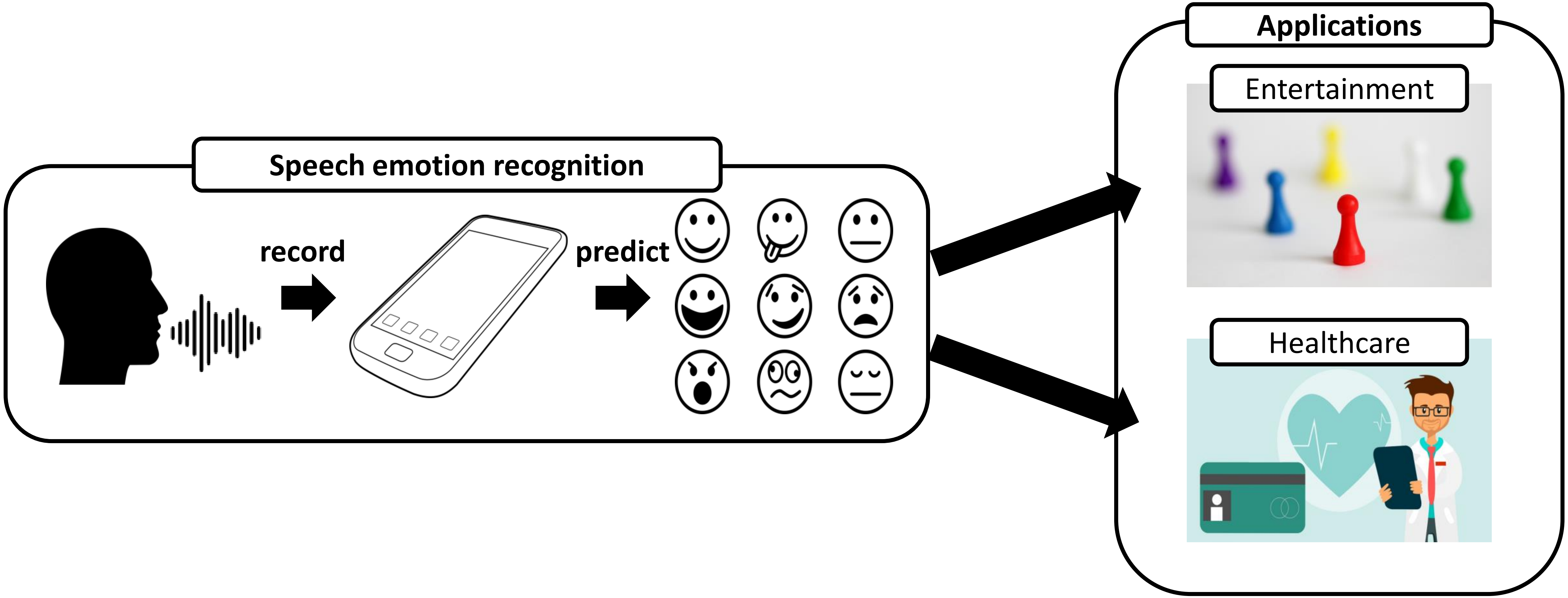
Seong-Gyun Leem, Daniel Fulford,
Jukka-Pekka Onnela, David Gard, and Carlos Busso

UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

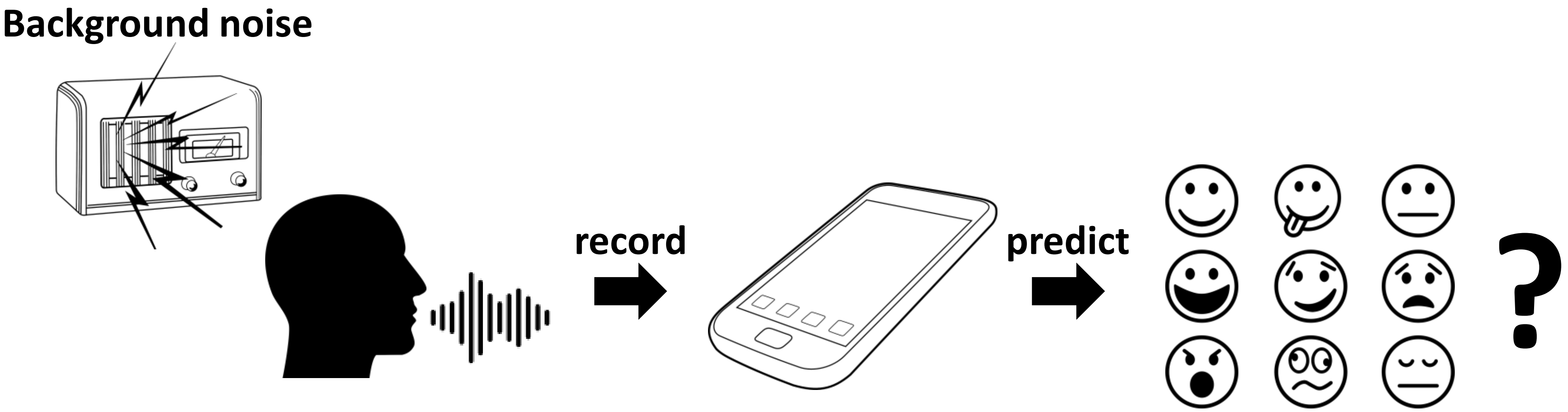
UT Dallas MSP Multimodal Signal Processing Laboratory

Speech emotion recognition (SER) in real-world applications



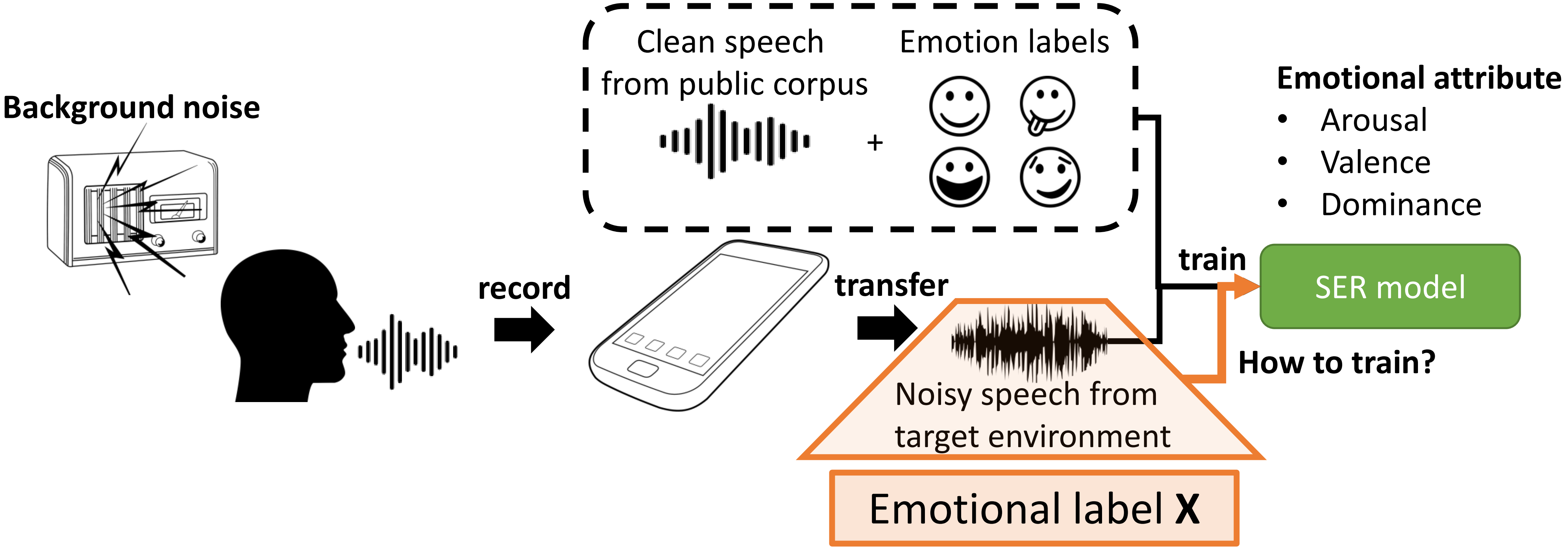
Speech emotion recognition (SER) in real-world applications

- **Needs to be robust against background noise**
  - Speech can be acquired from **unconstrained noisy environment**
  - Background noise can degrade the performance of SER system



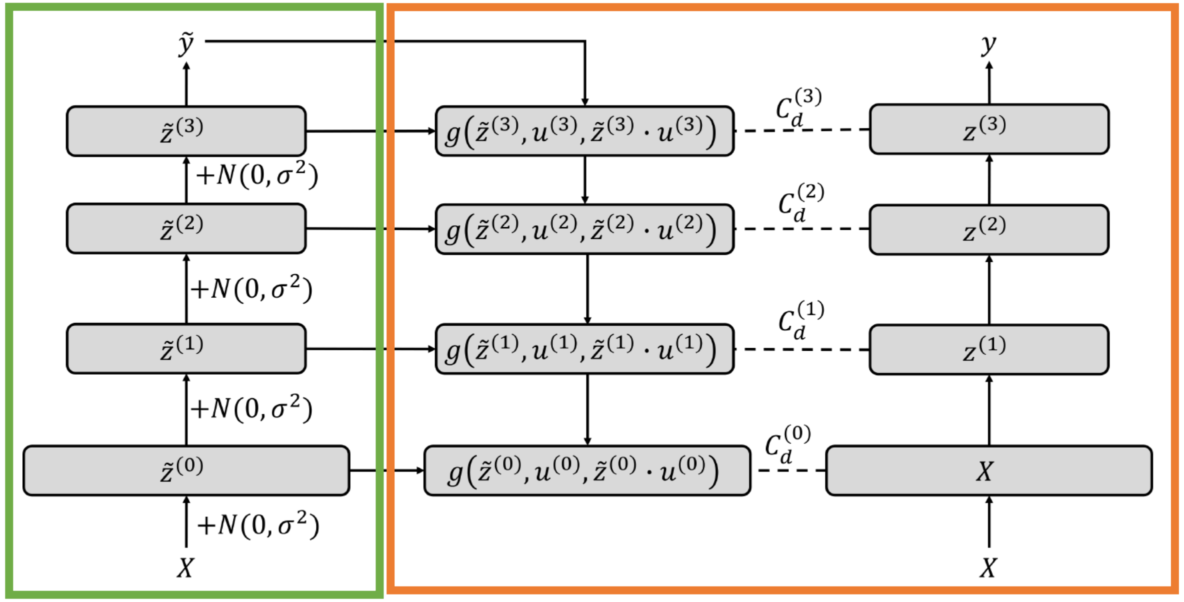
Semi-supervised learning for noise robust SER

- **Usage scenario**



Ladder network-based SER

- **Strengths**
  - It does not require emotional labels for target domain recordings
  - It can minimize train/test mismatch
- **Training**
  - Prediction task
    - Predict an emotional label by using labeled set
  - Reconstruction task
    - Reconstruct clean representations for each hidden layer

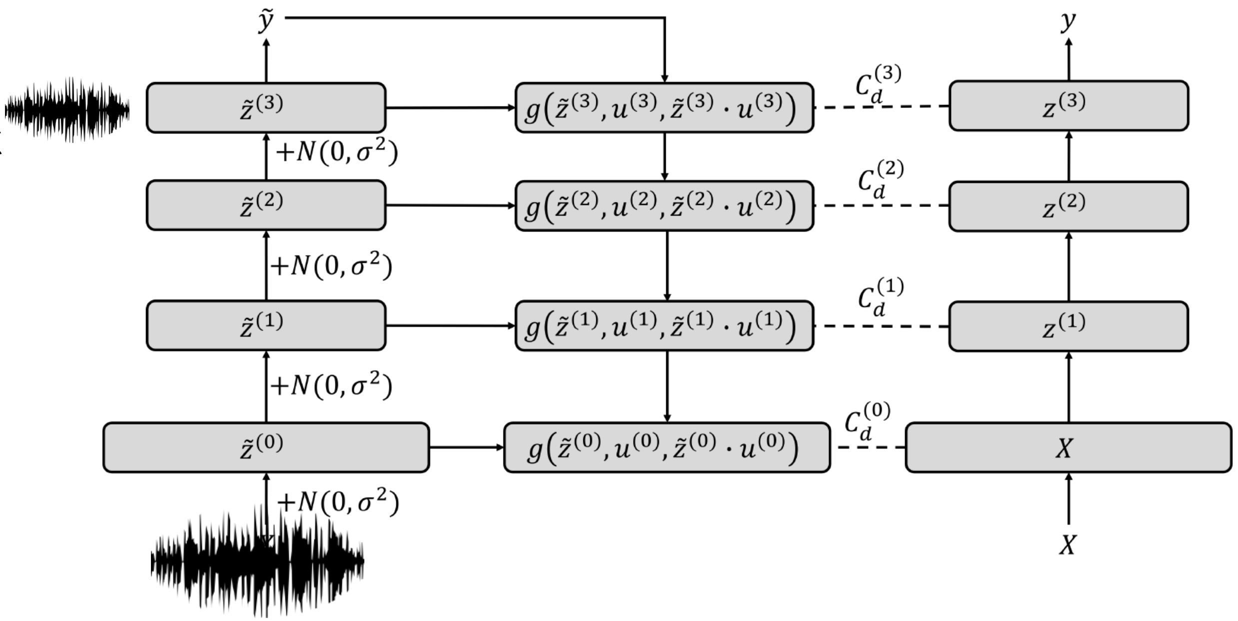


Parthasarathy, Srinivas, and Carlos Busso. "Semi-supervised speech emotion recognition with ladder networks." IEEE/ACM transactions on audio, speech, and language processing 28 (2020): 2697-2709.

Ladder network for noise robust SER

- **Problem**
  - Audio samples contain complex background noises
    - It can disrupt the emotion prediction task

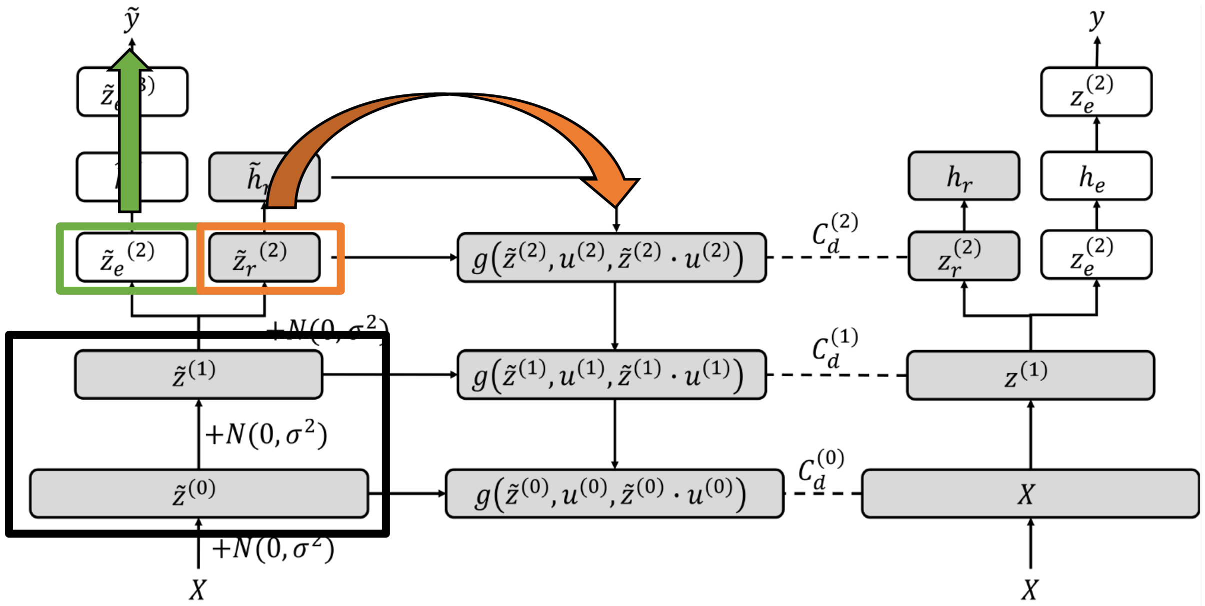
Information of noise
for the reconstruction task



Noisy speech

Decoupled ladder network (DLN)

- **Solution**
  - Decouple last hidden layer into emotion and reconstruction embedding
- **Reconstruction embedding**
  - Reconstruction task
- **Emotion embedding**
  - Prediction task
- **Lower layers**
  - Prediction + reconstruction task



Decoupled ladder network architecture

Decoupled ladder network (DLN)

- **Loss function**

$$C_{DLN} = \boxed{C_p\left(y, \widetilde{h_e^{(L)}}\right)} + \boxed{\sum_{l=0}^{L-2} \lambda^l \times C_r^l\left(\hat{z}_{BN}^{(l)}, z^{(l)}\right) + \lambda^{L-1} \times C_r^{L-1}\left(\hat{z}_{BN}^{(L-1)}, z_r\right)}$$

Prediction loss (for emotional attributes)

Reconstruction loss

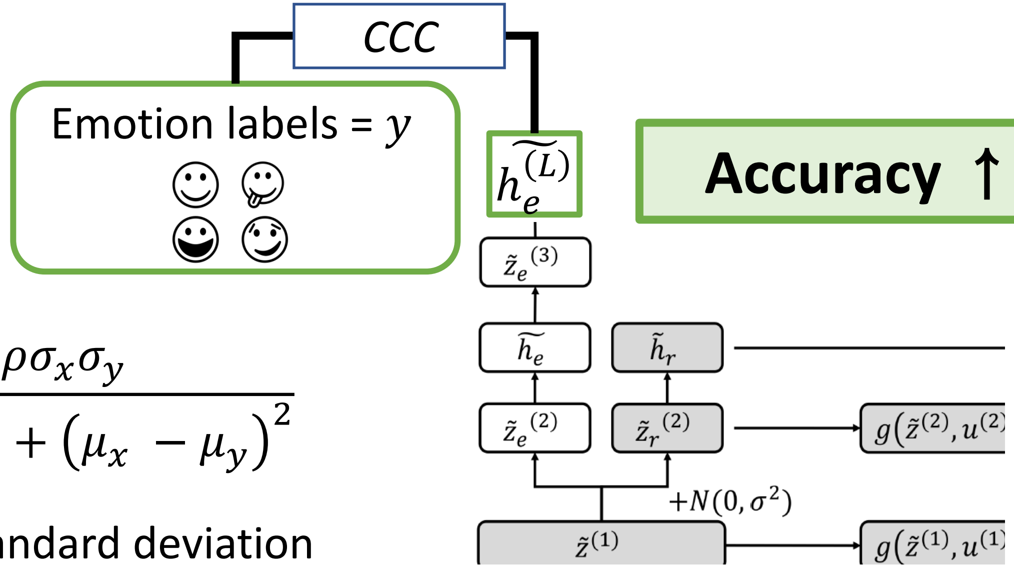
$$1 - \mathrm{CCC}\left(y, \widetilde{h_e^{(L)}}\right)$$

$$CCC(x, y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

σ_x, σ_y : standard deviation

μ_x, μ_y : mean

ρ: correlation coefficient



Accuracy ↑

Decoupled ladder network (DLN)

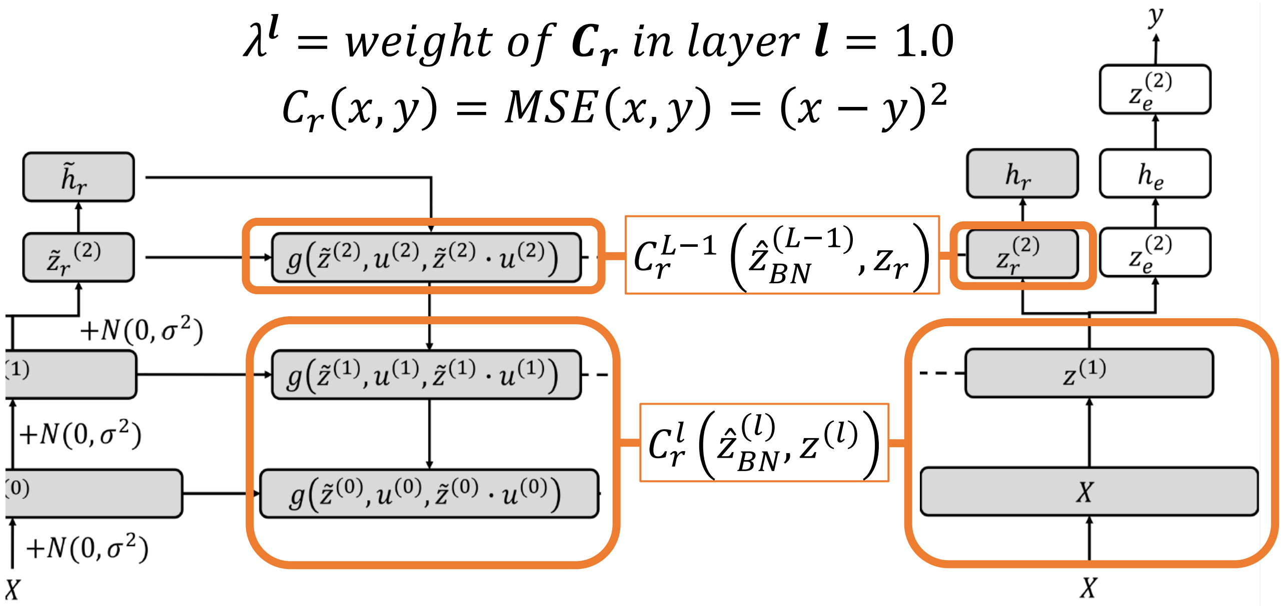
- **Loss function**

$$C_{DLN} = \boxed{C_p\left(y, \widetilde{h_e^{(L)}}\right)} + \boxed{\sum_{l=0}^{L-2} \lambda^l \times C_r^l\left(\hat{z}_{BN}^{(l)}, z^{(l)}\right) + \lambda^{L-1} \times C_r^{L-1}\left(\hat{z}_{BN}^{(L-1)}, z_r\right)}$$

Prediction loss **Reconstruction loss**

$$\lambda^l = weight\ of\ \boldsymbol{C_r}\ in\ layer\ \boldsymbol{l} = 1.0$$

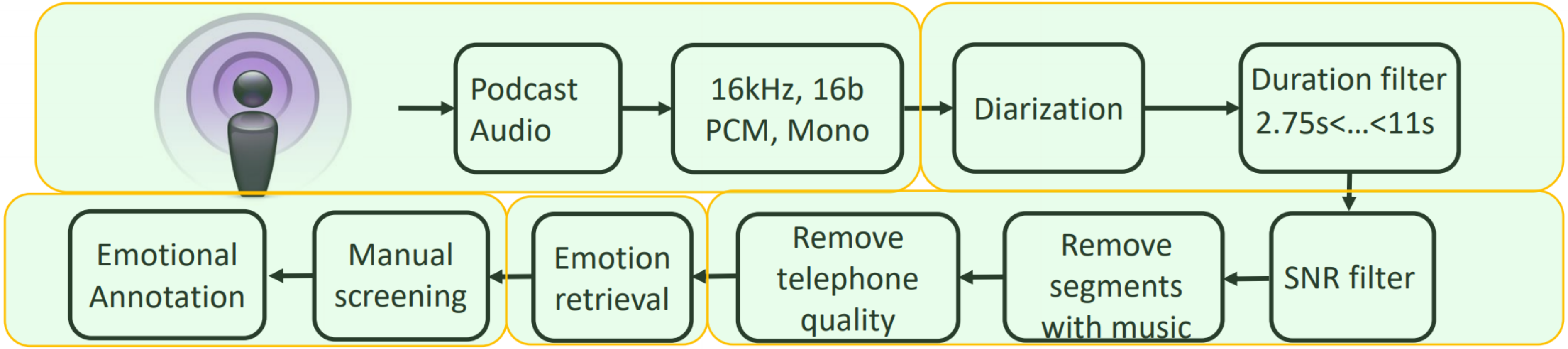
$$C_r(x, y) = MSE(x, y) = (x - y)^2$$



Difference ↓

The MSP-Podcast corpus (v1.8)

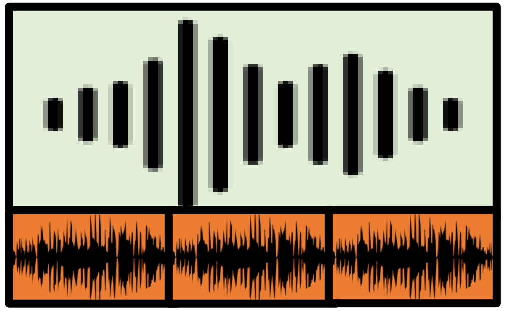
- **Spontaneous emotional speech dataset**
  - Podcast recordings are collected **(> 113 hours)**
- **Clean speech dataset**
  - SNR is above 20dB



Noisy version of the MSP-Podcast corpus

- **Noisy speech used in previous studies**
  - Noisy speech had been artificially synthesized in previous works
- **Limitation**
  - Not enough to simulate actual recording conditions

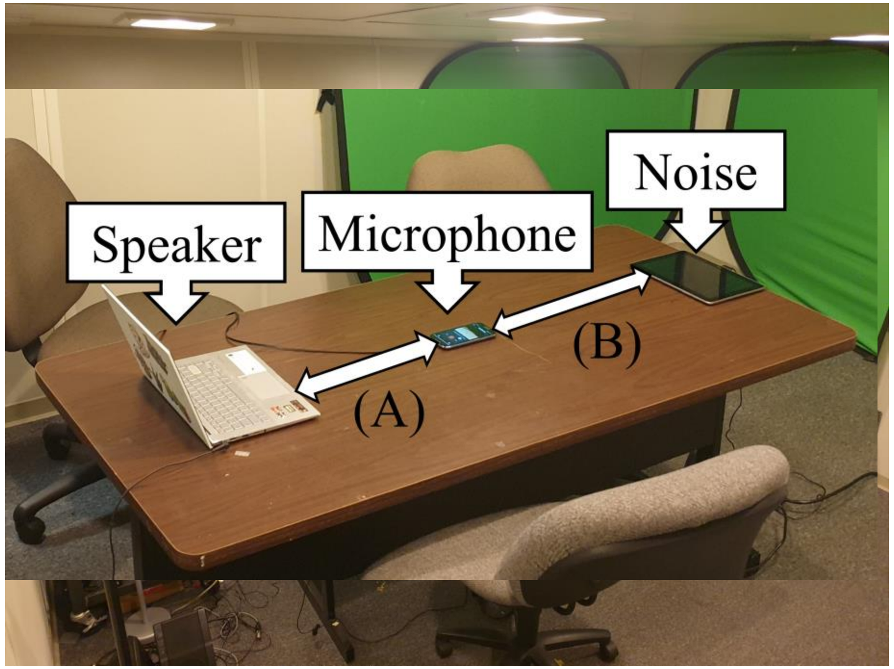
Fixed noise



Repeated noises

Noisy version of the MSP-Podcast corpus

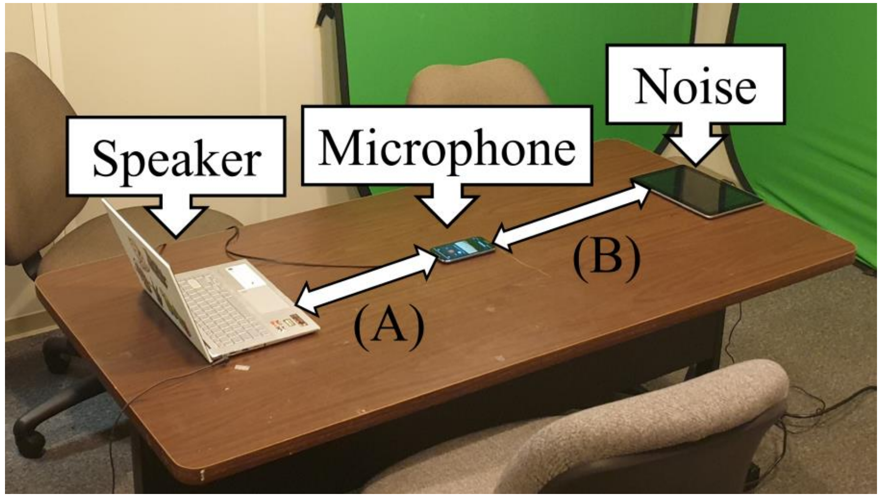
- **Solution**
  - Simultaneously playing the MSP-Podcast corpus and noise sound
  - Recording it with smartphone
- **Radio shows without copyright (noise)**
  - Simulating non-stational background noise
    - Human voice, musical sound, and sound effect



Noisy version of the MSP-Podcast corpus

- **Settings for each recording conditions**
  - 10dB, 5dB, 0dB conditions are collected

Recording condition	**(A)** (inch)	**(B)** (inch)	**Estimated SNR** (dB)
10dB	5	35	11.06
5dB	10	30	4.34
0dB	15	25	0.15



- **Emotional labels**
  - Noise sound is not related to the emotion
  - Emotional labels can be transferred from the MSP-Podcast corpus

Experiment setting

- **Data preparation**
  - MSP-Podcast v1.8 (clean speech set)
  - Noisy version of the MSP-Podcast corpus (noisy speech set)

Condition	Training	Development	Test	Unlabeled
Clean	44,879	7,800	15,326	43,361
Noisy (10dB, 5dB, 0dB)	-	-	15,326	43,361

  - Recording condition between the test set and the unlabeled training set is matched

- **Acoustic features**
  - 6,373 dimensions of 2013 ComParE feature set is used

Experiment setting

- **Baseline models**
  - Dense network
    - Model cannot use unlabeled set during training
  - Ladder network
    - Its last hidden layer is not separated into emotion and reconstruction embedding
  - All hyperparameters for the training and the number of layers, nodes are same as decoupled ladder network

Result

- **Concordance correlation coefficient (CCC)**
  - Average CCC over 20 trials

Task	Arousal				Valence				Dominance			
SNR	Clean	10dB	5dB	0dB	clean	10dB	5dB	0dB	clean	10dB	5dB	0dB
Dense network	**0.631**	0.248	0.229	0.192	**0.296**	0.151	0.120	0.104	**0.562**	0.253	0.252	0.215
Ladder network	0.627	0.438	0.424	0.364	0.280	0.146	0.129	0.111	0.545	0.381	0.385	0.339
Decoupled ladder network	0.625	**0.488**	**0.460**	**0.402**	0.283	**0.160**	0.126	0.114	0.556	**0.450**	**0.436**	**0.397**

Performance ↑

Performance ↓

- **Noisier speech shows lower performance than cleaner speech**
  - Background noise evokes detrimental effects on emotion prediction
- **Ladder network shows better performance in noisy conditions than dense network**
  - Semi-supervised learning can improve the robustness against the noise

Result

Task	Arousal				Valence				Dominance			
SNR	Clean	10dB	5dB	0dB	clean	10dB	5dB	0dB	clean	10dB	5dB	0dB
Dense network	**0.631**	0.248	0.229	0.192	**0.296**	0.151	0.120	0.104	**0.562**	0.253	0.252	0.215
Ladder network	0.627	0.438	0.424	0.364	0.280	0.146	0.129	0.111	0.545	0.381	0.385	0.339
Decoupled ladder network	0.625	**0.488**	**0.460**	**0.402**	0.283	**0.160**	0.126	0.114	0.556	**0.450**	**0.436**	**0.397**

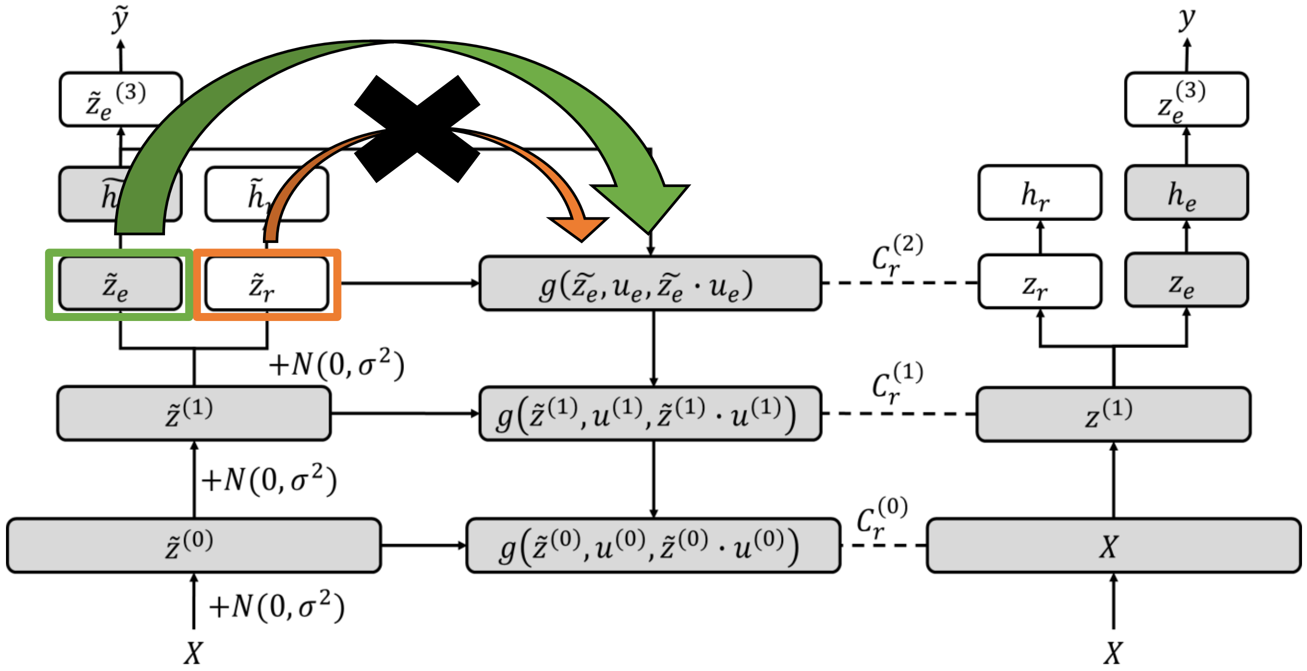
Performance ↑

No clear improvements

- **Decoupled ladder network improves the ladder network**
  - Arousal: 11.4% (10dB), 8.4% (5dB), 10.2% (0dB) ↑
  - Dominance: 17.1% (10dB), 13.2% (5dB), 7.0% (0dB) ↑

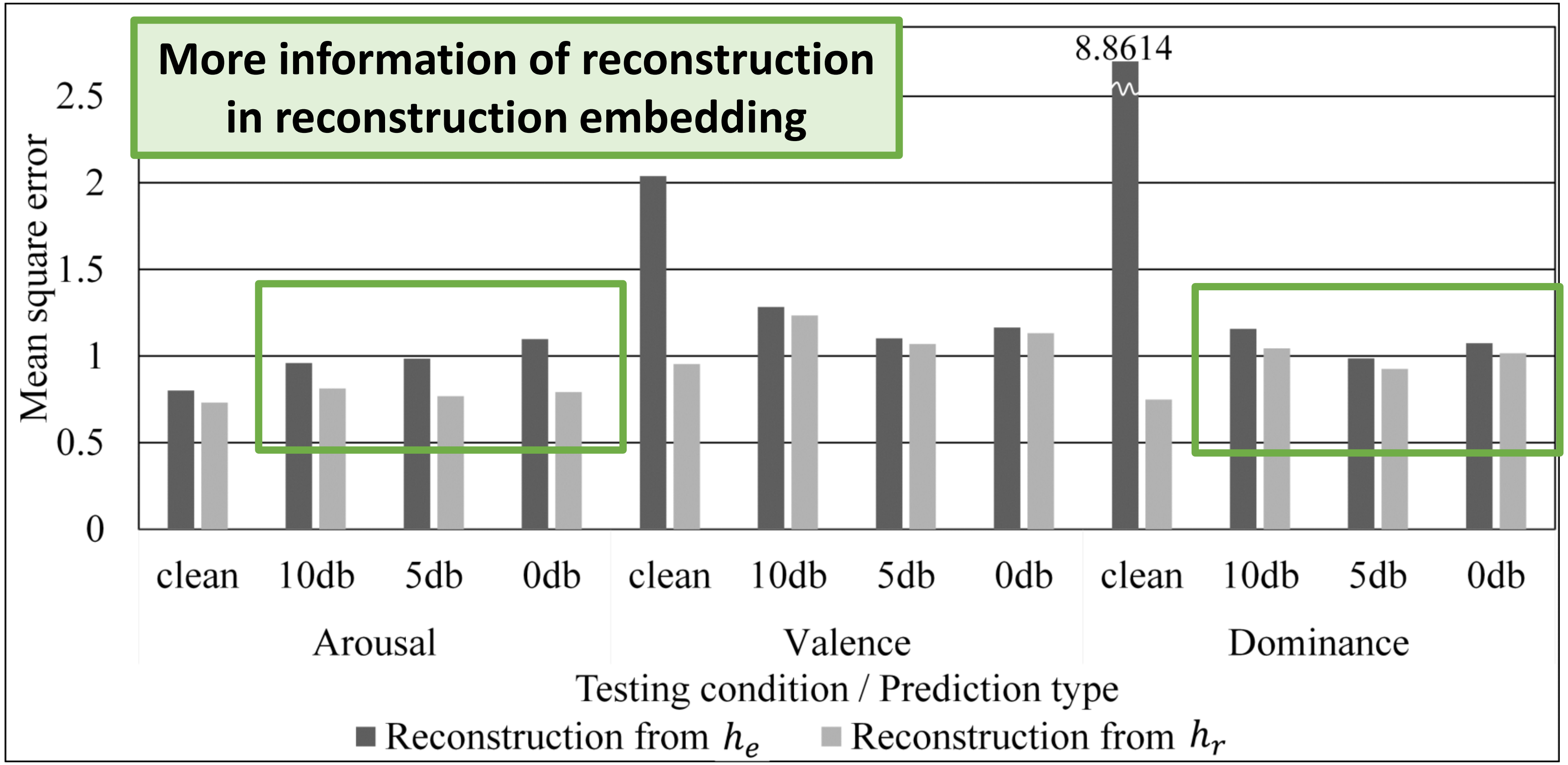
Analysis on separating the embedding

- **Reconstruction by using emotion embedding**
  - Emotion embedding is fed into the highest layer of decoder
  - Loss of using emotion embedding > Loss of using reconstruction embedding



Analysis on separating the embedding

- **Reconstruction loss**

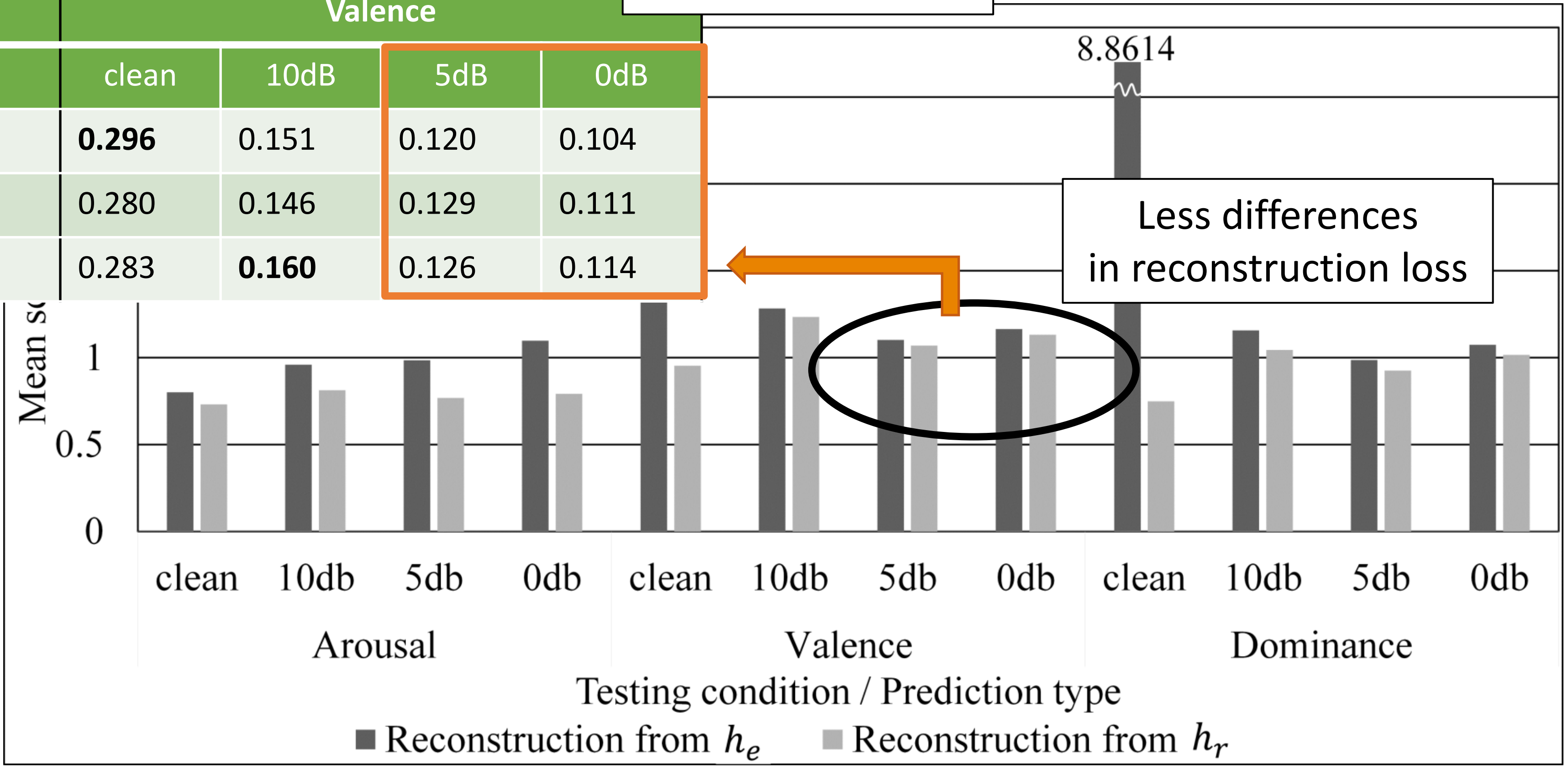


Analysis on separating the embedding

- **Reconstruction loss**

Task	Valence			
SNR	clean	10dB	5dB	0dB
Dense network	**0.296**	0.151	0.120	0.104
Ladder network	0.280	0.146	0.129	0.111
Decoupled ladder network	0.283	**0.160**	0.126	0.114

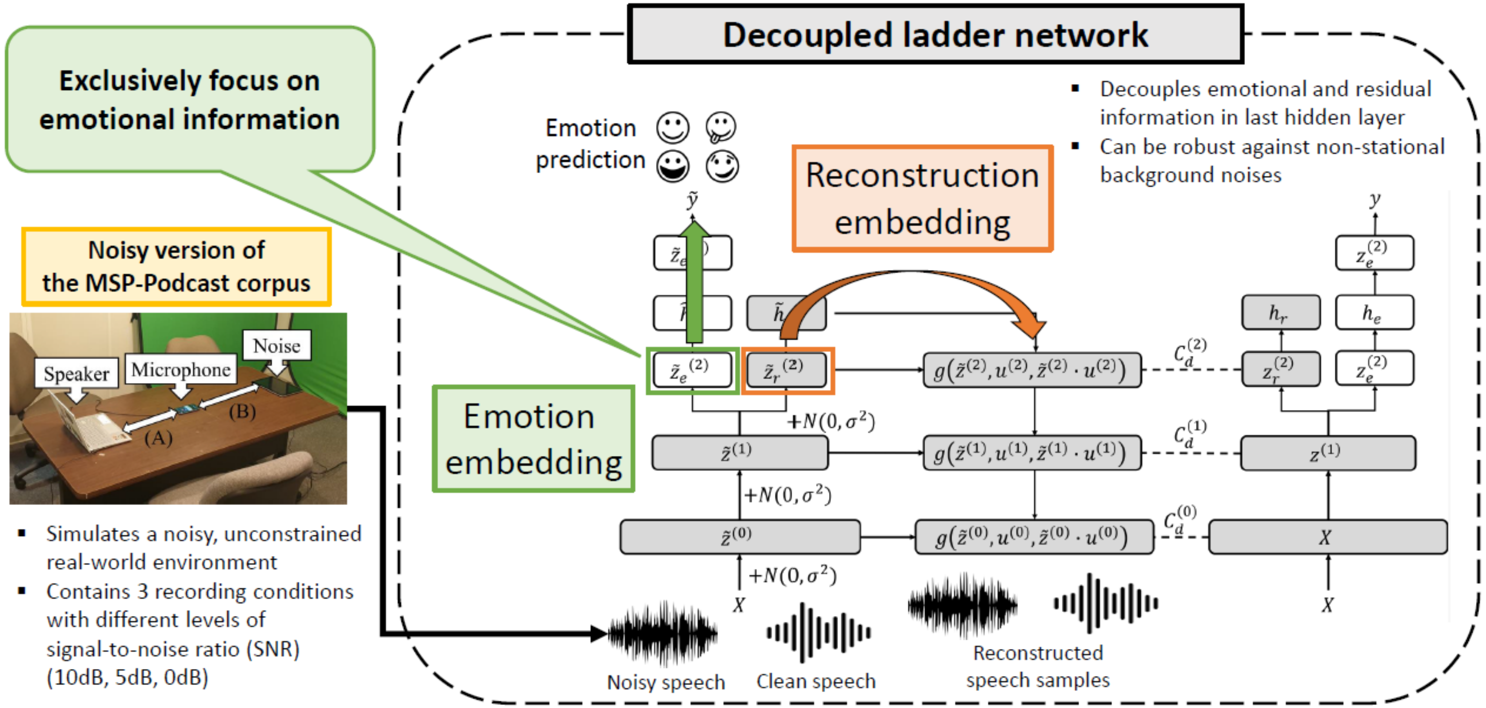
Less improvement in DLN



Less differences in reconstruction loss

Conclusion

- **Decouple ladder network**
  - Decouples the emotional and residual information to improve performance in noisy conditions
- **Noisy version of the MSP-Podcast corpus**
  - Simulates noisy, unconstrained recording environment.



Release of the MSP-Podcast corpus

- **Academic license**
  - ➢ Federal Demonstration Partnership(FDP) Data Transfer and Use Agreement
  - ➢ Free access to corpus
- **Commercial license**
  - ➢ Commercial license through UT Dallas
- **Plan to release the noisy version of the MSP-Podcast corpus**

https://msp.utdallas.edu

Thank you

- **This study was supported by NIH under grant 1R01MH122367-01.**

- **Questions or Contact: Seong-Gyun Leem**
  - SeongGyun.Leem@UTDallas.edu