Selective Acoustic Feature Enhancement for Speech Emotion Recognition with Noisy Speech

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

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

Published on IEEE/ACM Transactions on Audio, Speech, and Language Processing

ICASSP 2024 KOREA

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

National Institutes of Health

Introduction

Background:

- SER model requires multiple types of acoustic features
  - **Each feature has a difference in noise robustness**
  - **Only using noise-robust features improves an SER performance under noisy conditions** [Leem et al. ,2022]

Our work:

- Keep the noise-robust features
- **Enhance the noise-sensitive features**

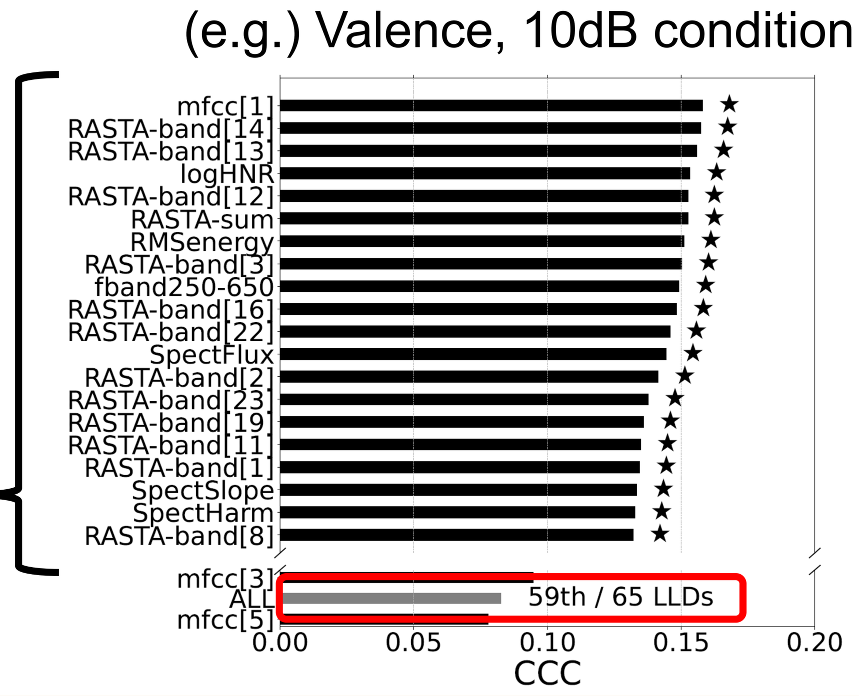
Single feature assessment

1. Train each probe model by using a **single clean LLD**
2. Evaluate performance with **a single clean/noisy LLD**
3. **Rank features based on the following criteria:**

1- Performance	2- Robustness	3- Joint
CCC_{noisy}	$\frac{CCC_{noisy} - CCC_{clean}}{CCC_{clean}}$	0.5*Performance + 0.5*Robustness

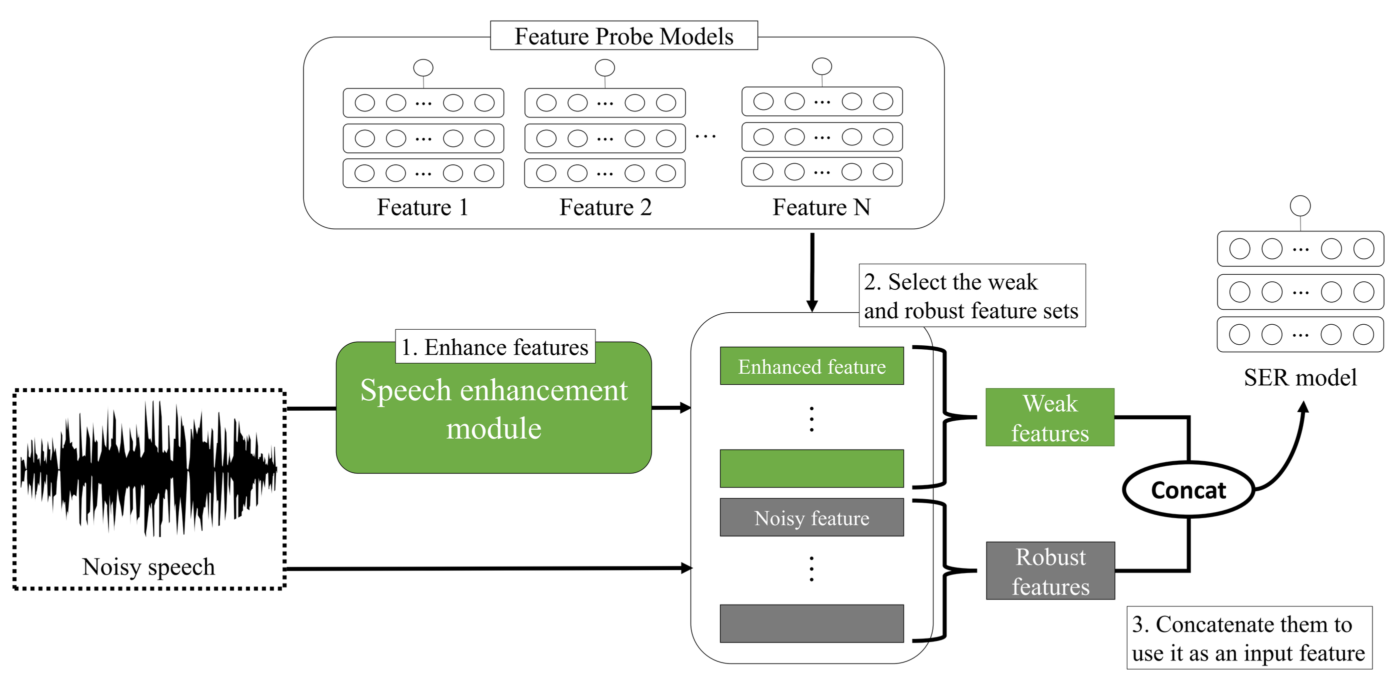
Some features perform better than using all features in noisy condition!

(e.g.) Valence, 10dB condition

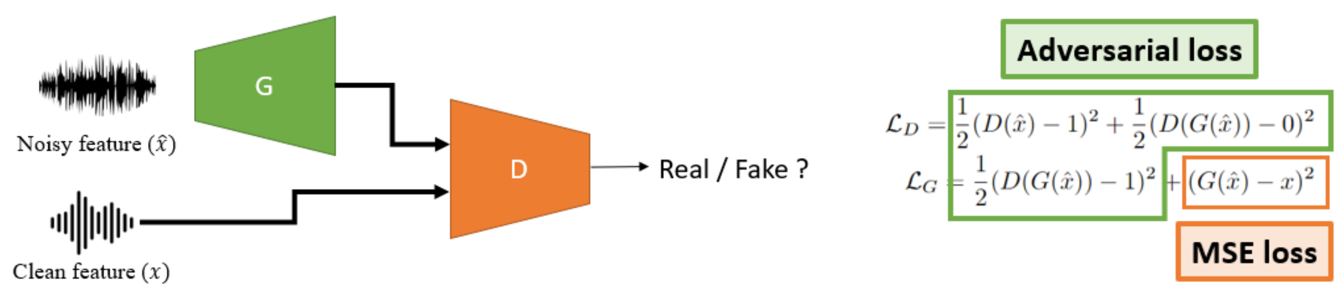


Proposed Method

Selective feature enhancement



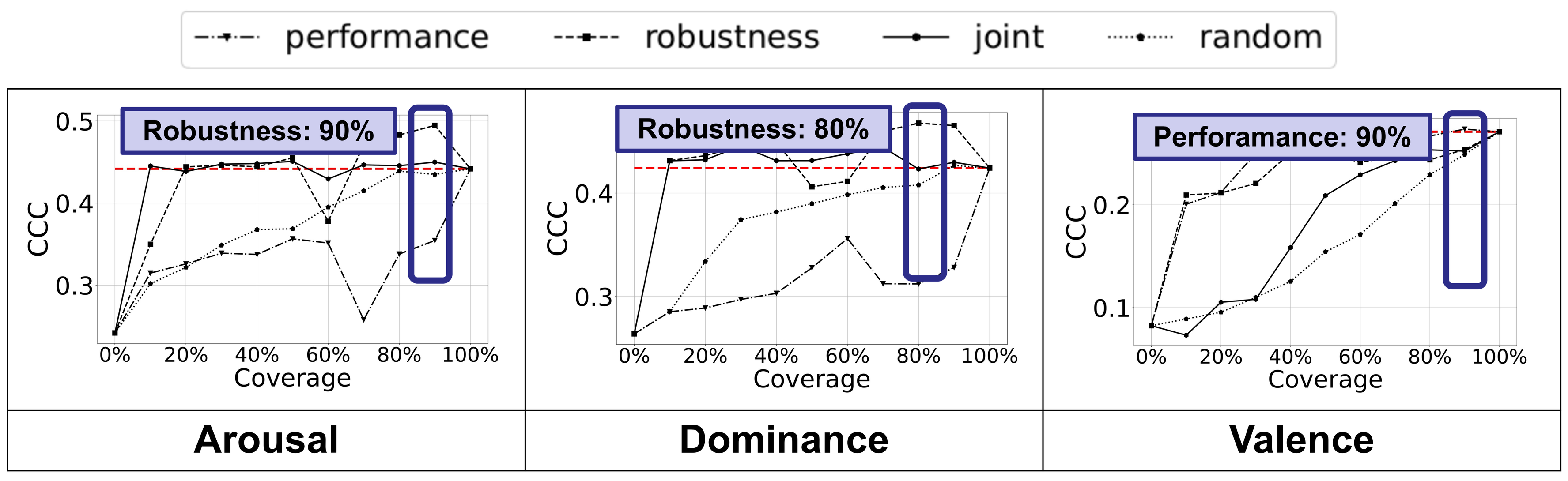
GAN-based feature enhancement



- Generator: 4 layers of 512 bidirectional gated recurrent unit (GRU) + linear output layer
- Discriminator: 3 layers of 32 bidirectional GRU + sigmoid output layer

Robust feature set selection

1. **Rank each feature** based on the proposed metric
2. **Add LLDs in increments of 10%** from the bottom to the top to define a weak feature set
   - Consider the rest of the features as robust features
3. **Check SER performance with original robust features + enhanced weak features**
   - (e.g.) 80% coverage = 80% features are enhanced + 20% features are kept
4. **Select the best feature set based on the development set analysis**
   - (e.g.) 10dB condition



SER performance in noisy condition:
enhancing weak features ≥ enhancing all the features

Experiment Settings

Data preparation

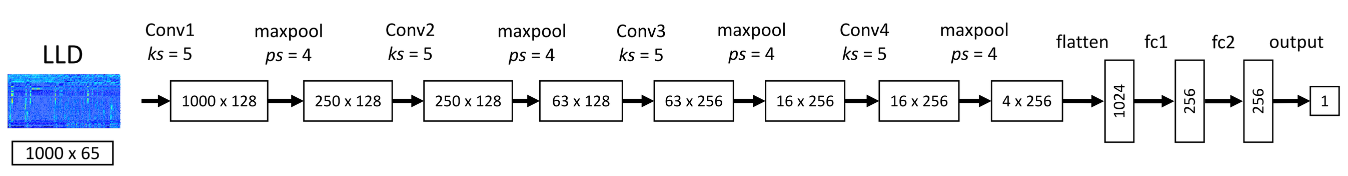
- Use the MSP-Podcast corpus (v1.8) as a clean speech set
- Noisy version of the corpus by directly recording the emotional speech with non-stational radio noise
  - We collect 10dB, 5dB, and 0dB conditions

Acoustic features

- Interspeech 2013 Computational Paralinguistic Challenge feature set
- 65 LLDs in the set

Emotion Recognition Framework

- Predict the emotional attribute scores
- Use multitask learning approach during training [Parthasarathy, 2017]



Conclusion

- Our selective feature enhancement approach can improve the prediction of emotional attribute scores under noisy conditions
- Feature-based enhancement approach leads to clear improvements for the top-performing features, which compensate for other features when all the LLDs are combined.
- Some features lead to lower SER performance after they are enhanced by the feature-based enhancement model
  - **Not all the features need to be enhanced**

Results

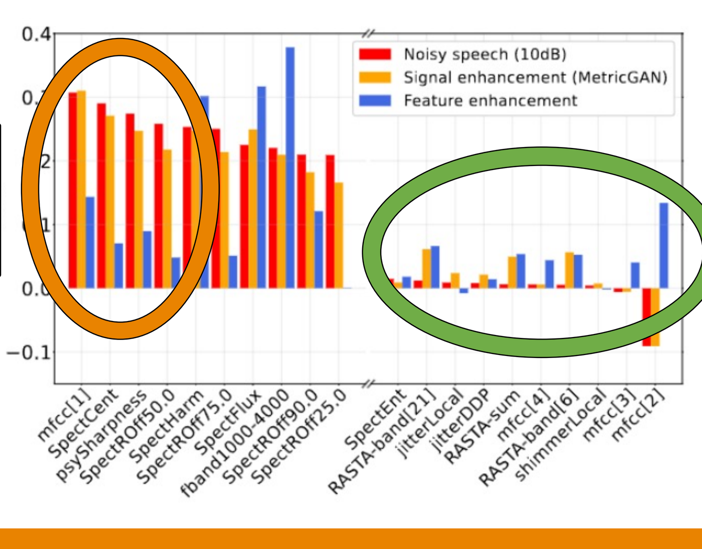
Emotion Recognition Performance (CCC)

	10dB			5dB			0dB		
	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.
Model w/o enhancement	0.278	0.288	0.097	0.228	0.262	0.076	0.194	0.214	0.058
DCCRN	0.151	0.138	0.140	0.111	0.087	0.081	0.083	0.068	0.081
MetricGAN	0.342	0.297	0.110	0.227	0.247	0.111	0.168	0.135	0.073
Only using robust features	0.364	0.385	0.159	0.302	0.370	0.139	0.268	0.321	0.117
Enhancing all features	0.450	0.400	0.179	0.412	0.403	0.177	0.393	0.376	0.147
Selective feature enhancement	**0.530**	**0.485**	**0.185**	**0.467**	**0.457**	**0.178**	**0.397**	**0.392**	**0.151**

1. **Feature enhancement > signal-based enhancement**
2. **Feature enhancement > feature selection**
3. **Enhancing all features > Selective feature enhancement**

Individual feature analysis (Aro. 10dB)

Performance drop after the enhancement



Clear improvements with a feature-based enhancement