Not All Features Are Equal: Selection of Robust Features for Speech Emotion Recognition in Noisy Environments

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

icassp 2022 Singapore

Motivation

Background:

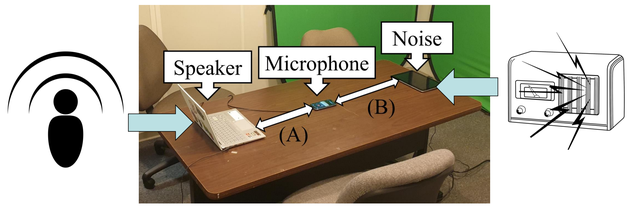
- Background noises distort the features used for speech emotion recognition (SER) systems
  - Disrupts the emotion prediction performance in real-world applications
- Do all features extracted from noisy speech equally degrade the prediction performance?
- Can we select a feature set that is most resilient to background noise?

Our Work:

- Examine the robustness of individual features
- Build a robust feature selection method by ranking low-level descriptors (LLDs) to improve the noise robustness

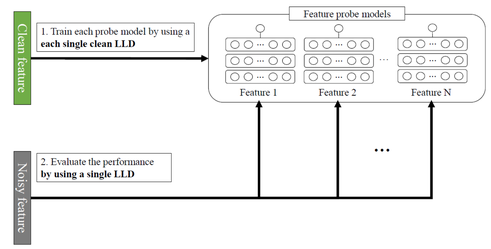
The MSP-Podcast Corpus

- Emotional corpus collected at UT-Dallas (v 1.8)
  - Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
  - Annotated on Amazon Mechanical Turk
    - Emotional attributes (valence, arousal, dominance)
    - Primary and secondary emotions, but not used here
- Noisy version of the corpus by directly recording the emotional speech with non-stational radio noise
  - Simulate noisy speech recorded from real-world applications
  - We collect 10dB, 5dB, and 0dB conditions

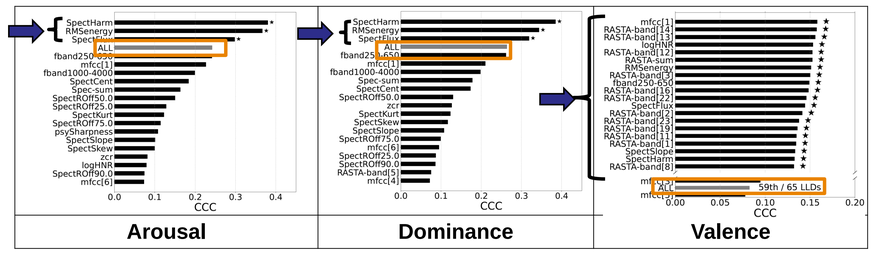


Single Feature Assessment

- We train 65 SER models, each trained with a single LLD
- Trained with clean speech, test with noisy speech
  - Environmental mismatch
- We evaluate the *concordance correlation coefficient* (CCC)



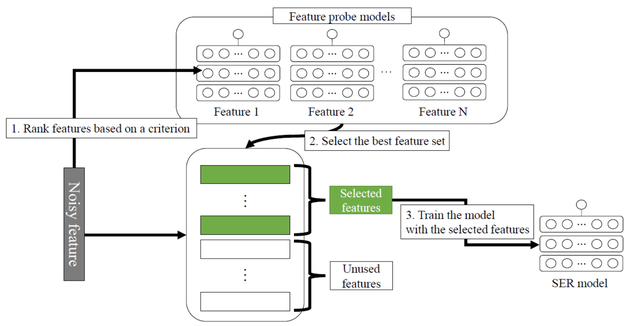
Performance of each LLD in 10dB condition



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

Feature Set Selection for Noisy Speech Emotion Recognition

Feature Set Selection



Selection Metrics

1- Performance

$$CCC_{noisy}$$

2- Robustness

$$\frac{CCC_{noisy} - CCC_{clean}}{CCC_{clean}}$$

3- Joint

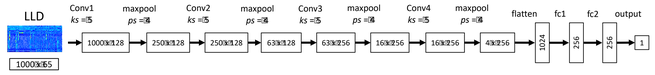
0.5*Performance + 0.5*Robustness

Acoustic Features

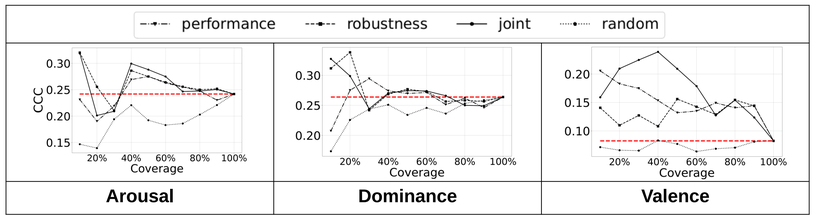
- Interspeech 2013 Computational Paralinguistic Challenge feature set
- 65 LLDs in the set

Emotion Recognition Framework

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



Cumulative Performance by Adding LLDs



Selected Coverage Arousal: 10% / Dominance: 20% / Valence: 40%

Test Set Result

	10dB			5dB			0dB		
	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence
Performance	0.265	0.298	0.109	0.288	0.305	0.096	0.236	0.258	0.083
Robustness	0.316	**0.357**	**0.139**	0.252	**0.340**	**0.115**	0.201	**0.290**	**0.084**
Joint	**0.346**	0.319	0.115	**0.340**	0.302	0.109	**0.292**	0.257	0.076
Random	0.157	0.239	0.074	0.141	0.221	0.063	0.116	0.183	0.048
All features	0.278	0.288	0.097	0.228	0.262	0.076	0.194	0.214	0.058

Matched condition | **Mismatched condition**

Conclusions

- Ranking features based on:
  - Performance
  - Robustness
  - Performance and robustness
- Rank-based feature selection is better than using all features in noisy condition
- Random selection does not help
- Approach also worked in mismatched SNR condition

Future Work

- We will investigate robustness of the feature set depending on type of noise
- Enhance weak features instead of enhancing all features

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

NIH National Institutes of Health