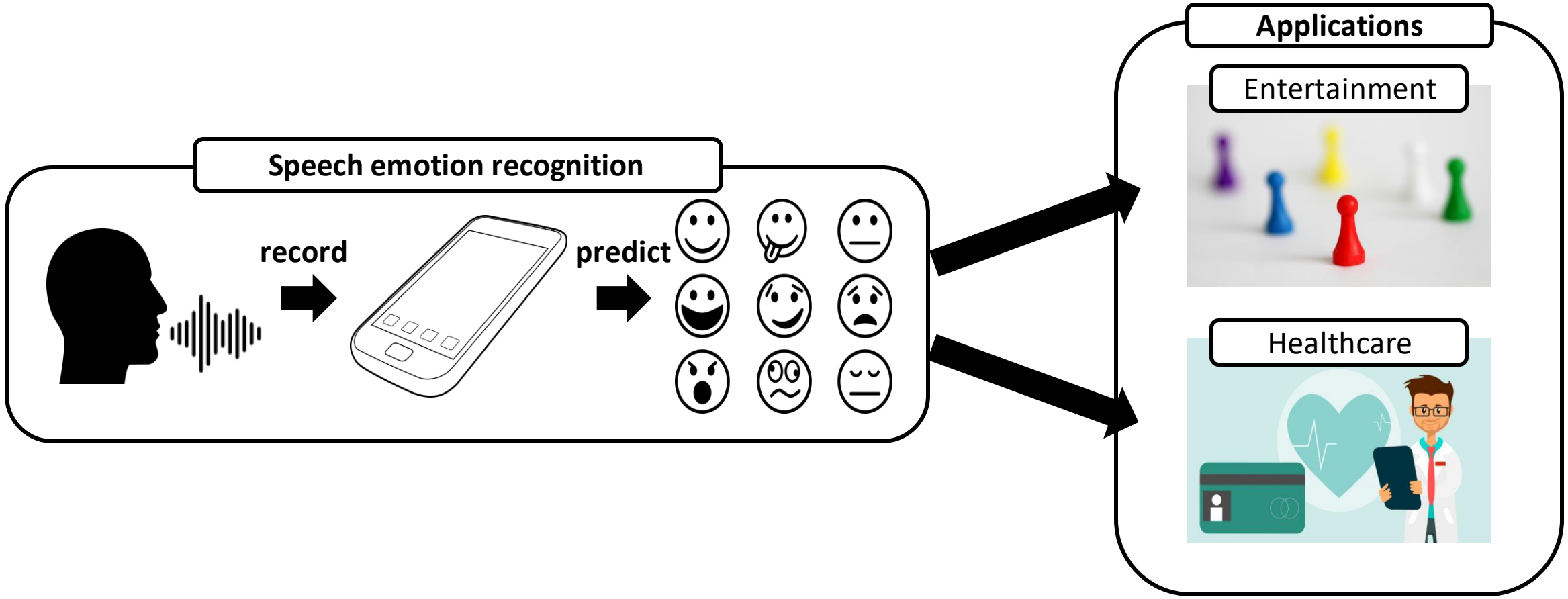
Not All Features Are Equal: Selection of Robust Features for Speech Emotion Recognition in Noisy Environments

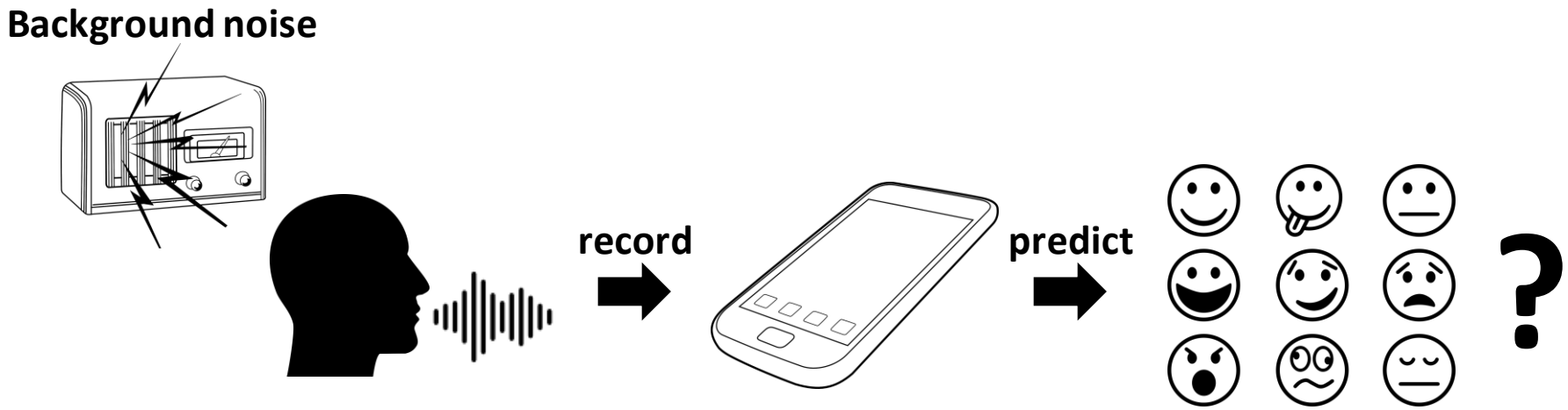
Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso

Speech Emotion Recognition (SER) in Real-World Applications



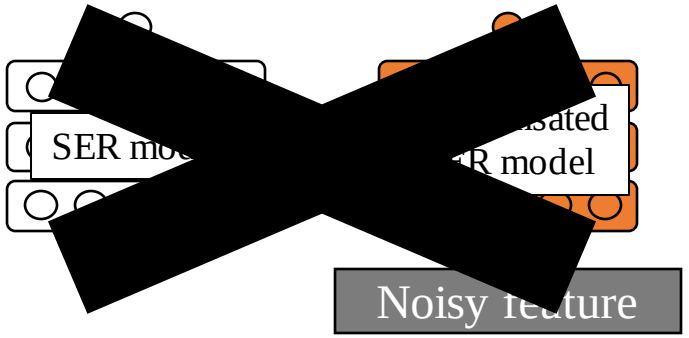
Speech Emotion Recognition (SER) in Real-World Applications

- **Performance degradation caused by the background noise**
  - Speech can be acquired from **unconstrained noisy environment**
  - Background noises **distort the features** used for SER system
    => **disrupts the prediction performance in real-world applications**



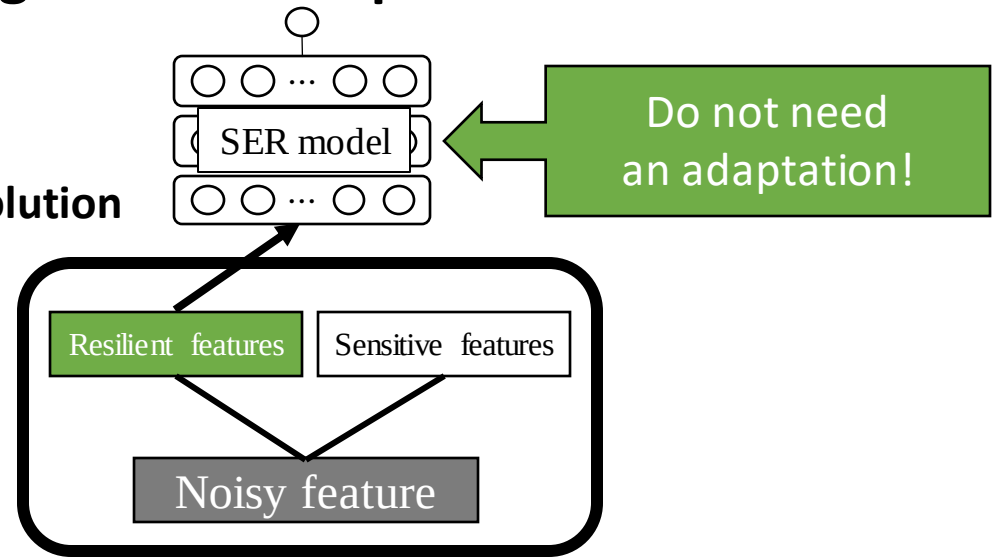
Our work

- **Examine the robustness of individual features**
  - There exist **features resilient to a background noise**
- **Build a robust feature selection method for noisy SER**
  - Improves the performance **without using a model adaptation**



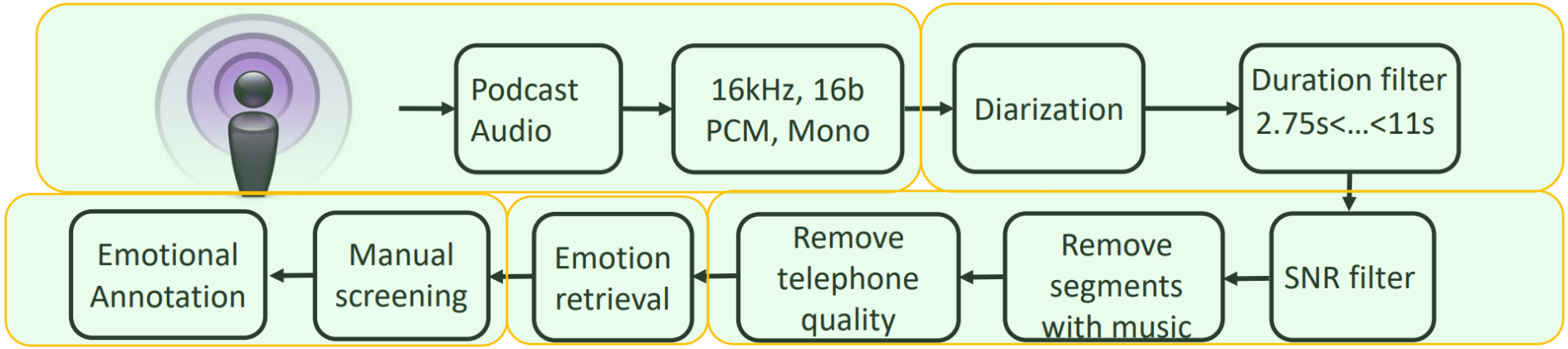
Previous approaches

Our solution



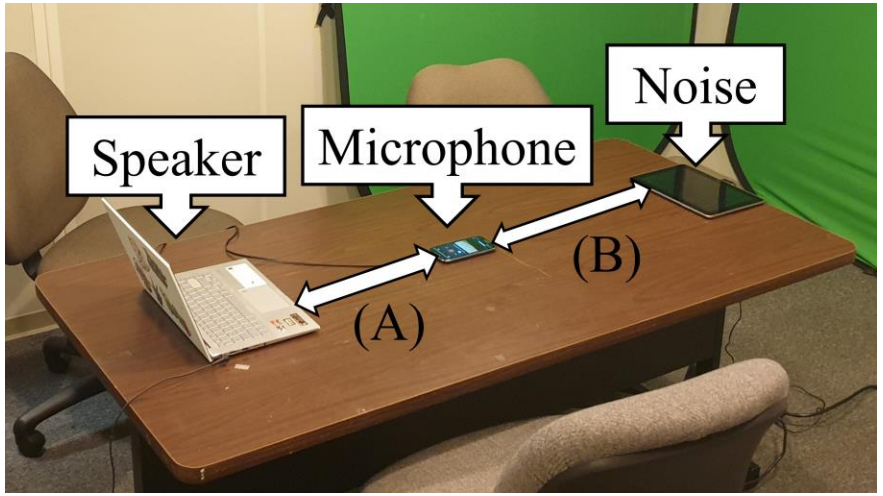
The MSP-Podcast Corpus (v1.8)

- **Spontaneous emotional speech dataset**
  - Podcast recordings are collected **(> 113 hours)**
- **Annotated on Amazon Mechanical Turk**
  - We focus on **emotional attributes (arousal, valence, dominance)**



Noisy Version of the MSP-Podcast Corpus

- **Simulate noisy speech recorded from real-world applications**
  - Use non-copyright radio shows as a noise
  - Directly record the MSP-Podcast and radio noise on smartphone
  - 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**
  - Emotional labels are transferred from the clean MSP-PODCAST corpus

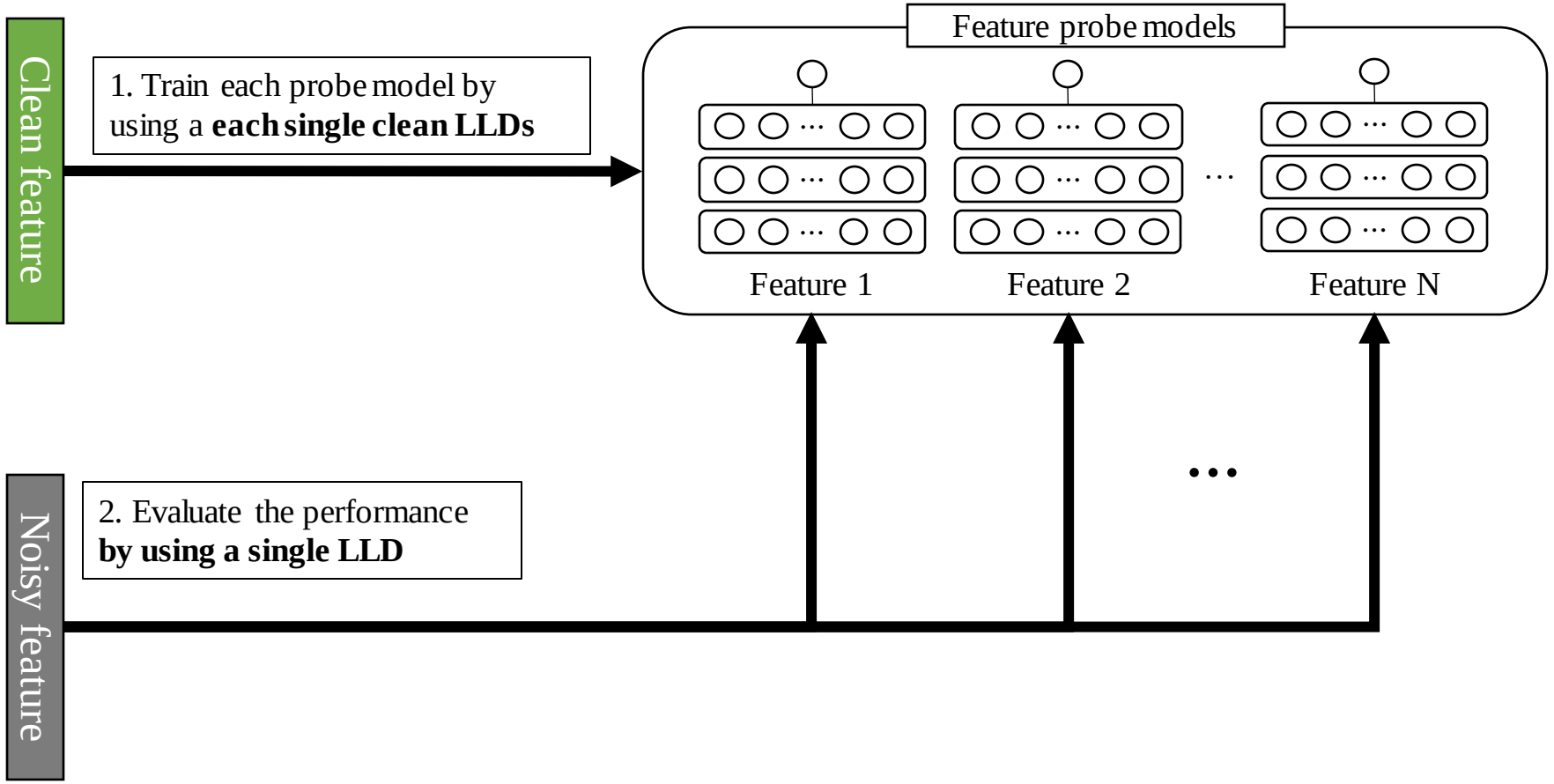
Resources

- **Data preparation**
  - MSP-Podcast v1.8 (clean speech set)
    - Recordings are annotated for emotional attribute labels (arousal, valence, dominance)
  - Noisy version of MSP-Podcast (noisy speech set)

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

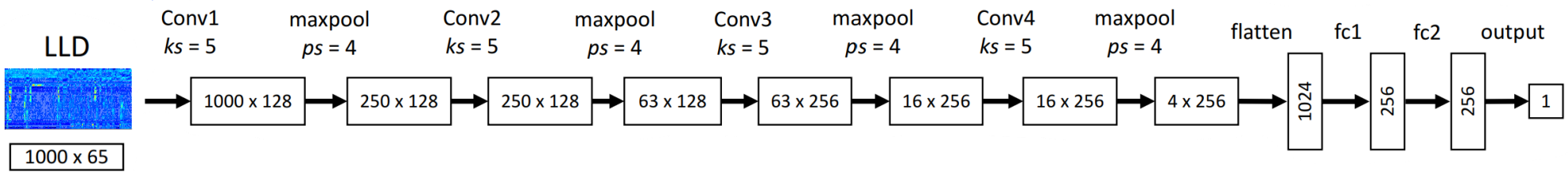
  - **All models are trained with the clean set**
  - **Development sets are used for single feature analysis and feature selection**
- **Acoustic features**
  - 2013 ComParE feature set is used
  - 65 dimensions of low-level descriptors (LLDs)

Single Feature Assessment



Single Feature Assessment

- **Emotion recognition model**
  - Architecture



  - Each model predicts an emotional attribute score
    - Arousal, dominance, valence
  - Multitask learning is used [Parthasarathy & Busso, 2020]
    - $\mathcal{L}_{total} = \alpha \times \mathcal{L}_{aro} + \beta \times \mathcal{L}_{val} + (1 - \alpha - \beta) \times \mathcal{L}_{dom}$
    - Concordance correlation coefficient (CCC) is used
  - 10% dropout is applied to the input

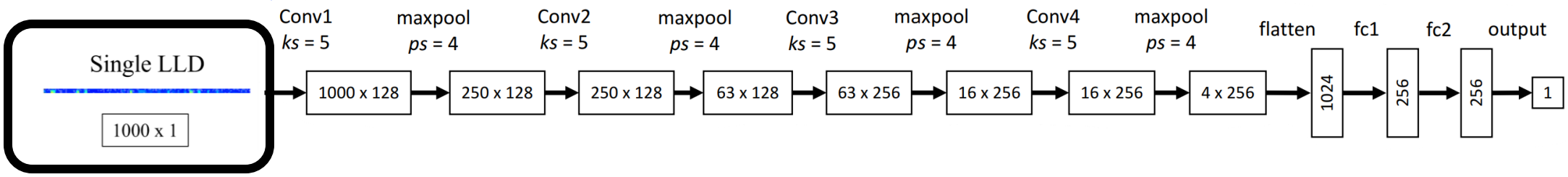
	α	β
Arousal	0.7	0.3
Valence	0.1	0.8
Dominance	0.0	0.2

Coefficients for multitask learning

Srinivas Parthasarathy and Carlos Busso, "Semi-supervised speech emotion recognition with ladder networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2697-2709, September 2020.

Single Feature Assessment

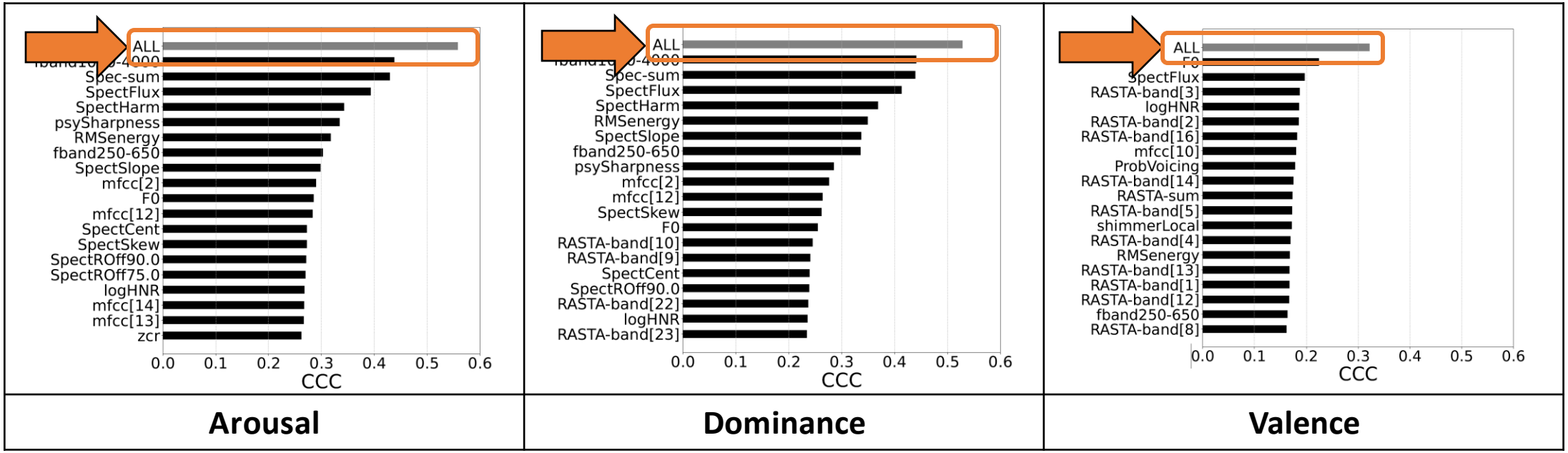
- **Feature probe models**
  - Architecture



  - All the models have the same architecture as the model trained with all the LLDs
  - They also follow the same training strategy as the emotion recognition model
  - **A single feature is used as an input**

Single Feature Assessment

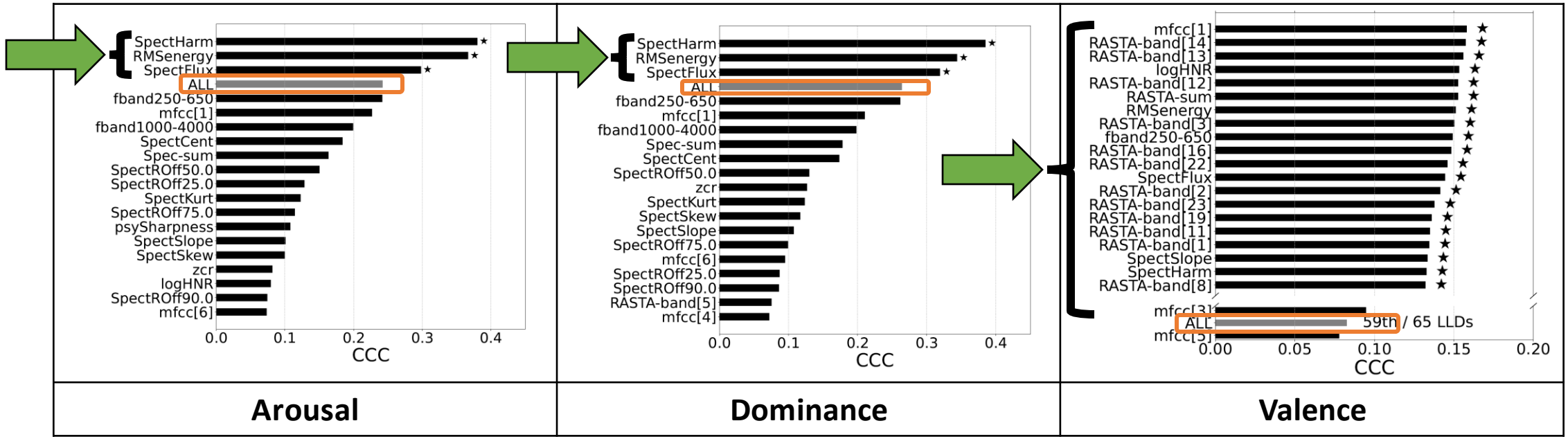
- **Clean Condition**



  - Using all features shows the best performance

Single Feature Assessment

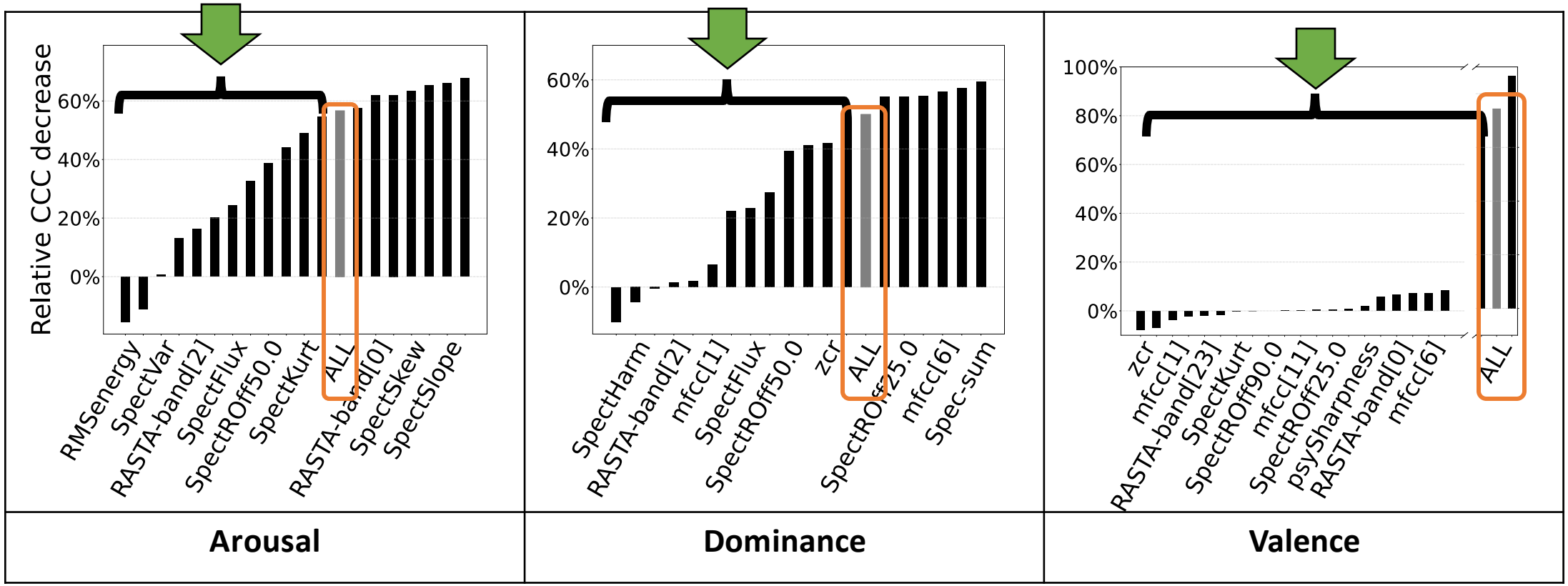
- **Noisy condition (10dB)**



- A model trained with a single LLD **perform better than using all the LLDs**

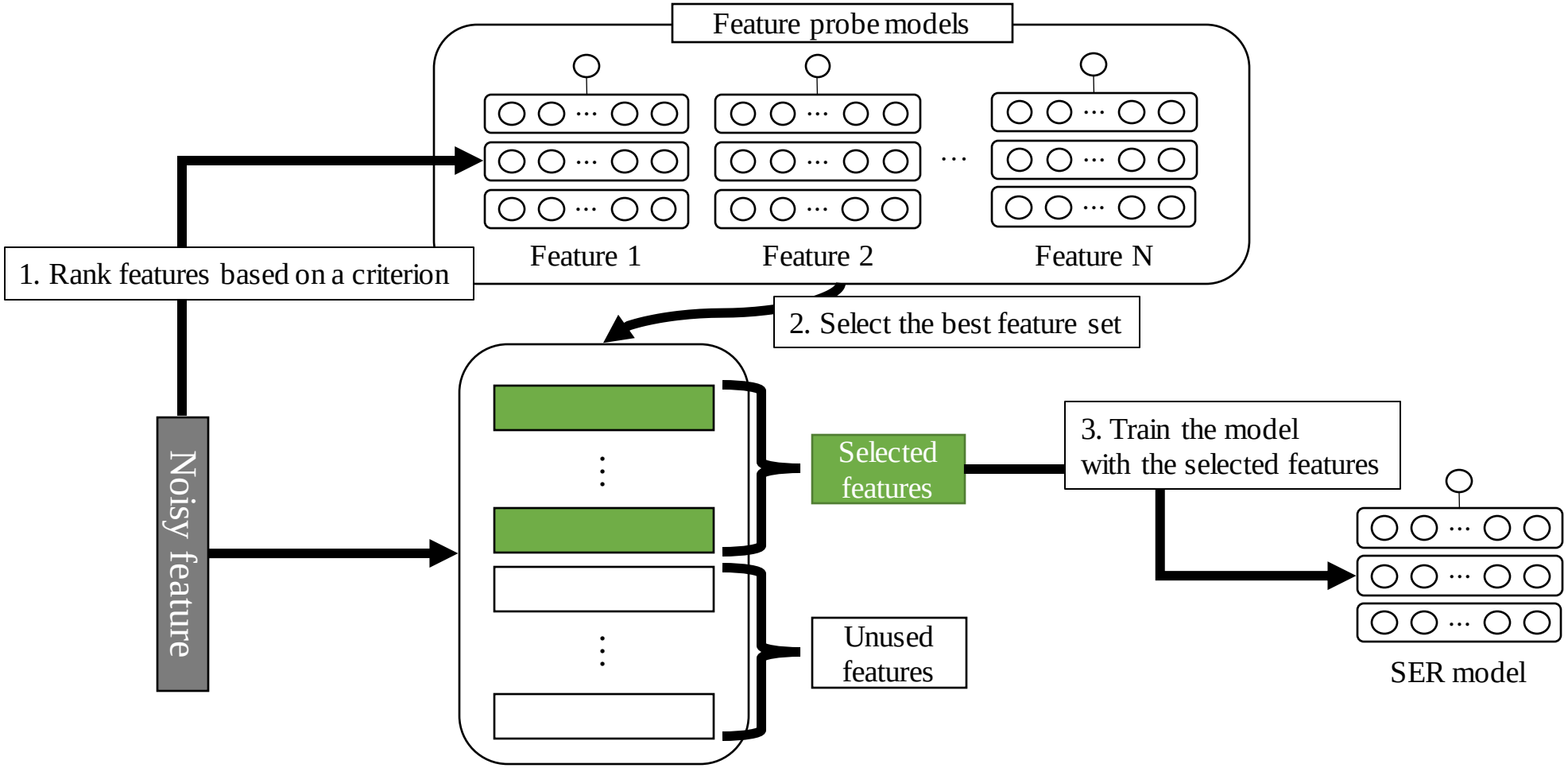
Single Feature Assessment

- **Relative performance decrease**



- Some single LLDs show **the less performance decrease than using all LLDs**

Robust Feature Selection For Noisy SER

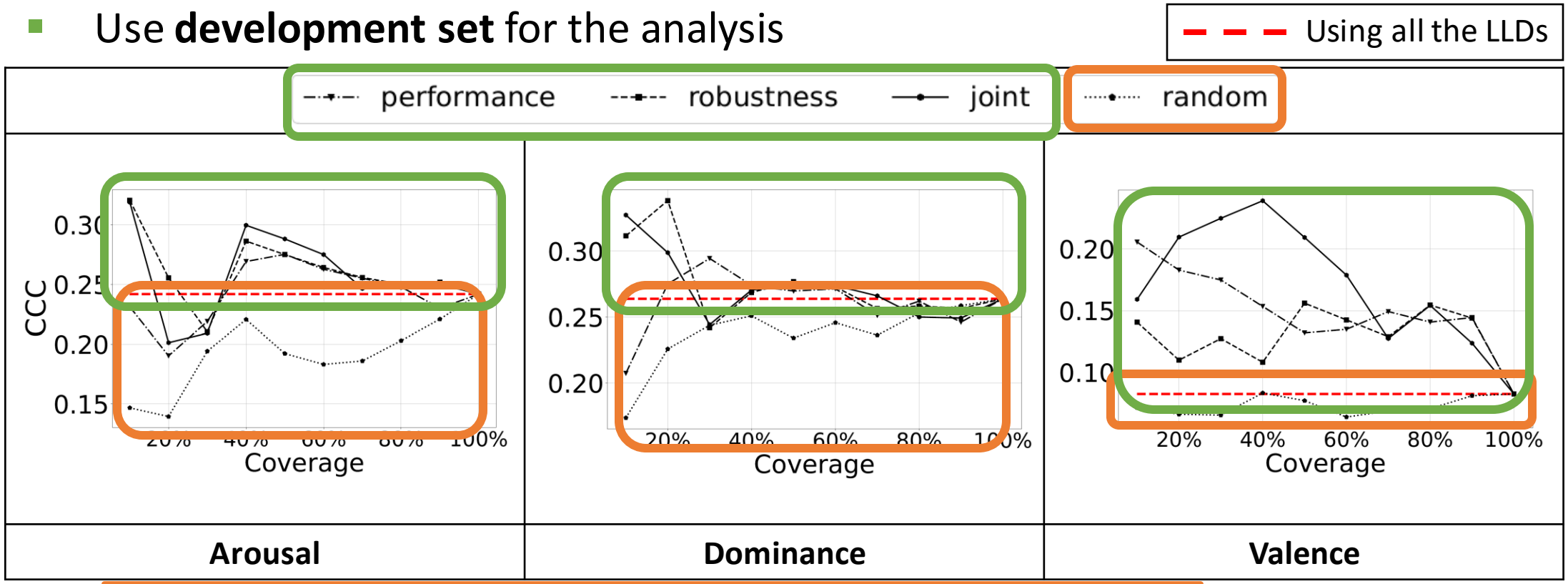


Feature Selection Metrics

- **Performance**
  - **Absolute performance** in the noisy condition
  - $\mathcal{R}_{performance} = CCC_{noisy}$
- **Robustness**
  - **Relative performance decrease** from the clean to the noisy condition
  - $\mathcal{R}_{robustness} = \frac{\{CCC_{noisy} - CCC_{clean}\}}{CCC_{clean}}$
- **Joint**
  - **Summation of the performance and the robustness ranks**
  - $\mathcal{R}_{joint} = 0.5 \times \mathcal{R}_{performance} + 0.5 \times \mathcal{R}_{robustness}$

Cumulative Performance by Adding LLDs

- **Coverage**
  - Use **development set** for the analysis

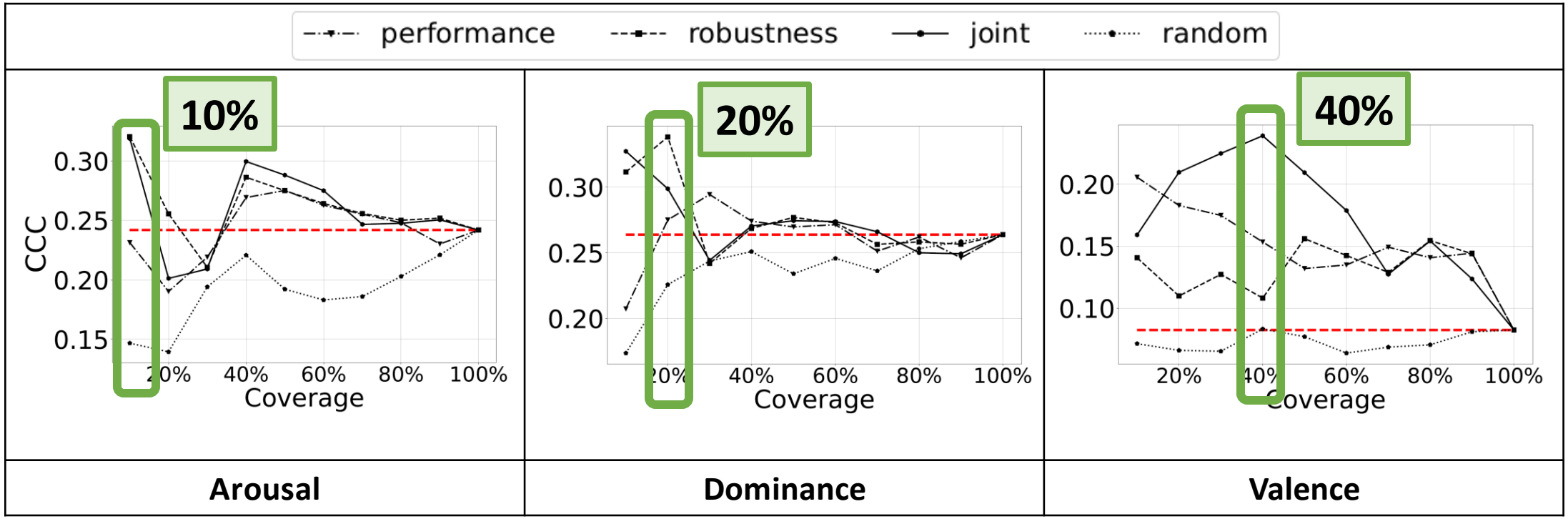


Random selection does not improve the performance

There exist feature sets better than using all LLDs

Coverage Selection

- **Selected Coverage**
  - Select the best feature set **based on the development set analysis**



Results

- **Comparison between clean and noisy condition**
  - Use **Test set** for the evaluation

	Clean			10dB		
	Arousal	Dominance	Valence	Arousal	Dominance	Valence
Performance	0.401	0.399	0.165	0.265	0.298	0.109
Robustness	0.379	0.429	0.151	0.316	**0.357**	**0.139**
Joint	0.414	0.413	0.192	**0.346**	0.319	0.115
Random	0.376	0.405	0.181	0.157	0.239	0.074
All features	**0.572**	**0.505**	**0.212**	0.278	0.288	0.097

Clean condition:
Using all the features is the best

Noisy condition:
Selecting the features is better!

  - Improvements: 24.4% (Arousal) / 23.9% (Dominance) / 43.2% (Valence)
  - Randomly selecting the features does not improve the performance
    - **Using a smaller number of features does not necessarily improve performance**

Results

- **Mismatched noisy condition**
  - Train SER model with the clean speech
  - Use **10dB condition** to select the resilient features

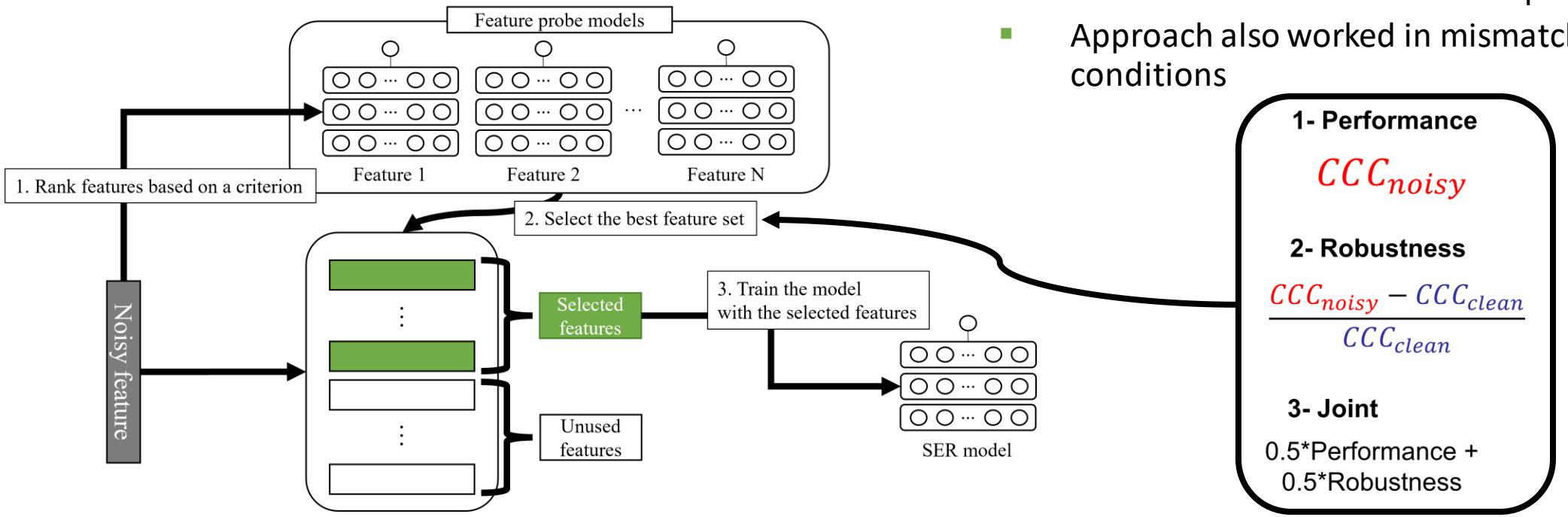
Do not need to match the condition for feature selection

	5dB			0dB		
	Arousal	Dominance	Valence	Arousal	Dominance	Valence
Performance	0.288	0.305	0.096	0.236	0.258	0.083
Robustness	0.252	**0.340**	**0.115**	0.201	**0.290**	**0.084**
Joint	**0.340**	0.302	0.109	**0.292**	0.257	0.076
Random	0.141	0.221	0.063	0.116	0.183	0.048
All features	0.228	0.262	0.076	0.194	0.214	0.058

  - Improvements
    - 5dB: 49.1% (Arousal) / 29.7% (Dominance) / 51.3% (Valence)
    - 0dB: 50.5% (Arousal) / 35.5% (Dominance) / 44.8% (Valence)

Conclusions

- **Not all features are equal**
  - Some features are resilient to background noises for SER task



- **Robust feature set selection**
  - Rank-based feature selection is better than using all features in noisy condition
  - Random selection does not help
  - Approach also worked in mismatched SNR conditions

1- Performance

$$CCC_{noisy}$$

2- Robustness

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

3- Joint

0.5*Performance + 0.5*Robustness

Thank You

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

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