Politecnico di Torino

DBMG

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features

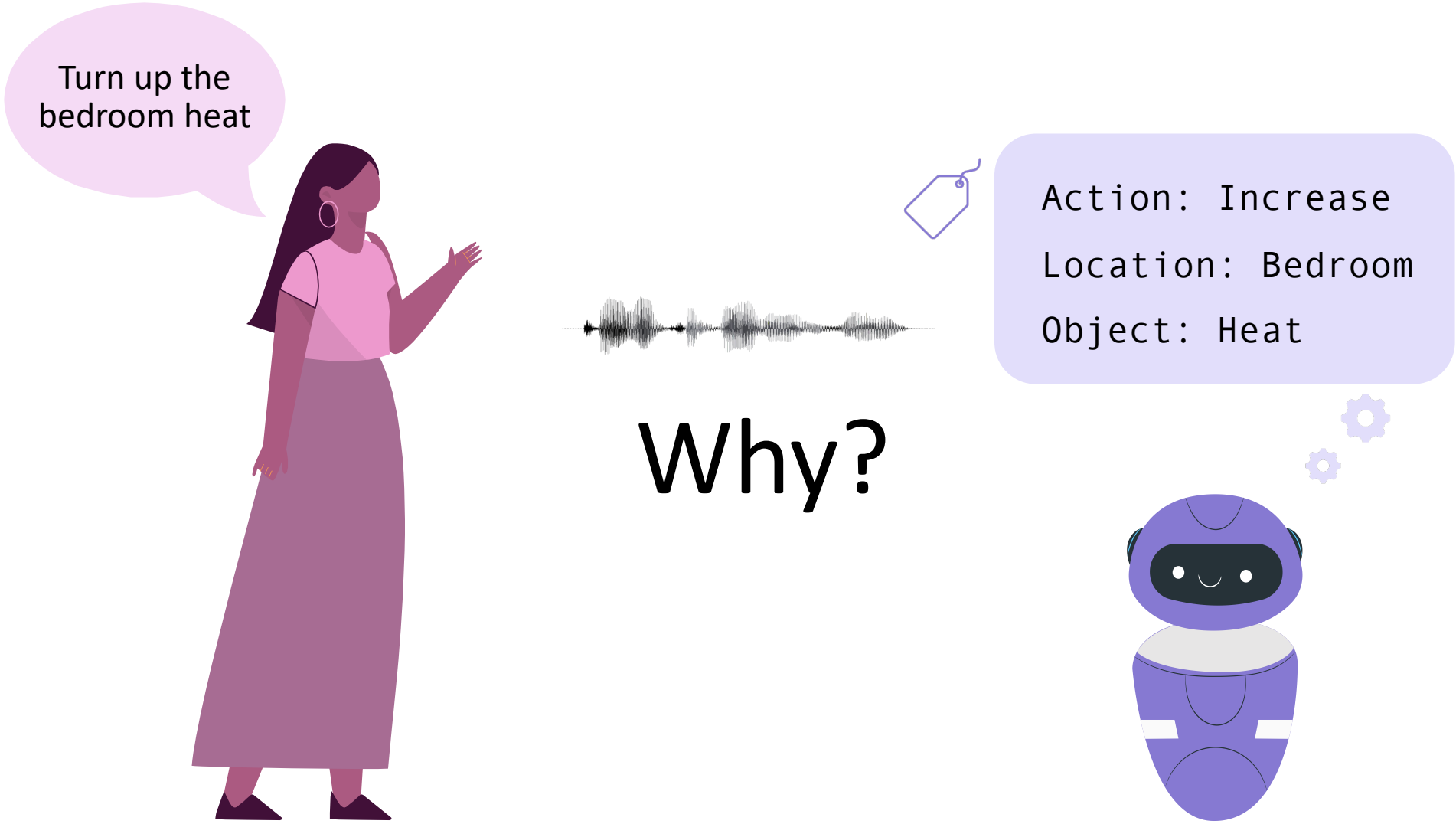
Eliana Pastor, Alkis Koudounas, Giuseppe Attanasio, Dirk Hovy, Elena Baralis

Politecnico di Torino, Bocconi University - Italy

Bocconi

MilaNLP

INTEGRATOR European Research Council

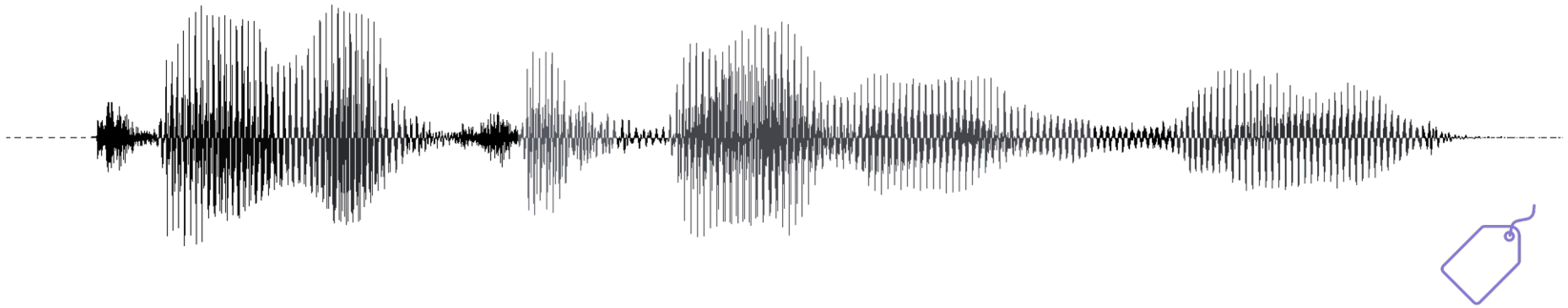

Turn up the bedroom heat
Action: Increase
Location: Bedroom
Object: Heat
Why?

On the need of explaining speech models

Right for the right reasons?

Why is it incorrect?

Encode bias?



Explain the **interaction** between

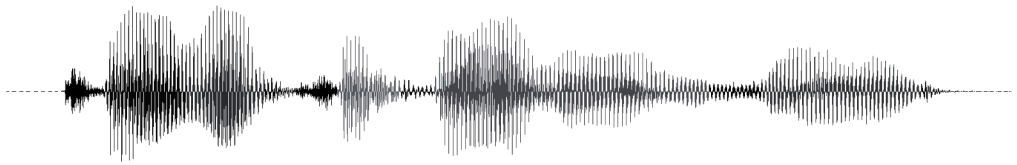
utterance components and **predictions**

in a **human-understandable** manner

RQ1. How do we define interpretable representations describing utterances?

Semantic

Spoken words



Turn up the bedroom heat

Paralinguistic

Prosody & external conditions

Pitch

Noise level

Speaking rate

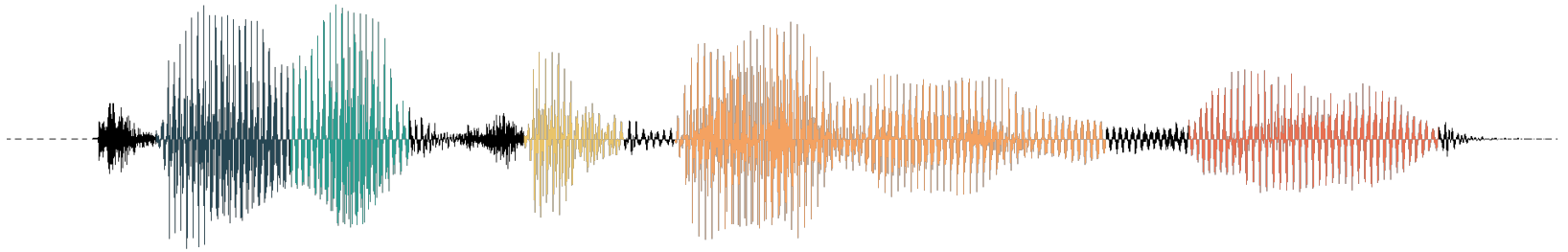
RQ2. How do we explain predictions at the semantic and paralinguistic levels?

Perturbation-based approach

- Perturb the utterance based on an interpretable feature
- Measure the impact on predictions
- The greater the change, the more the model relies on this feature!

Semantic

Use a word-level time alignment model

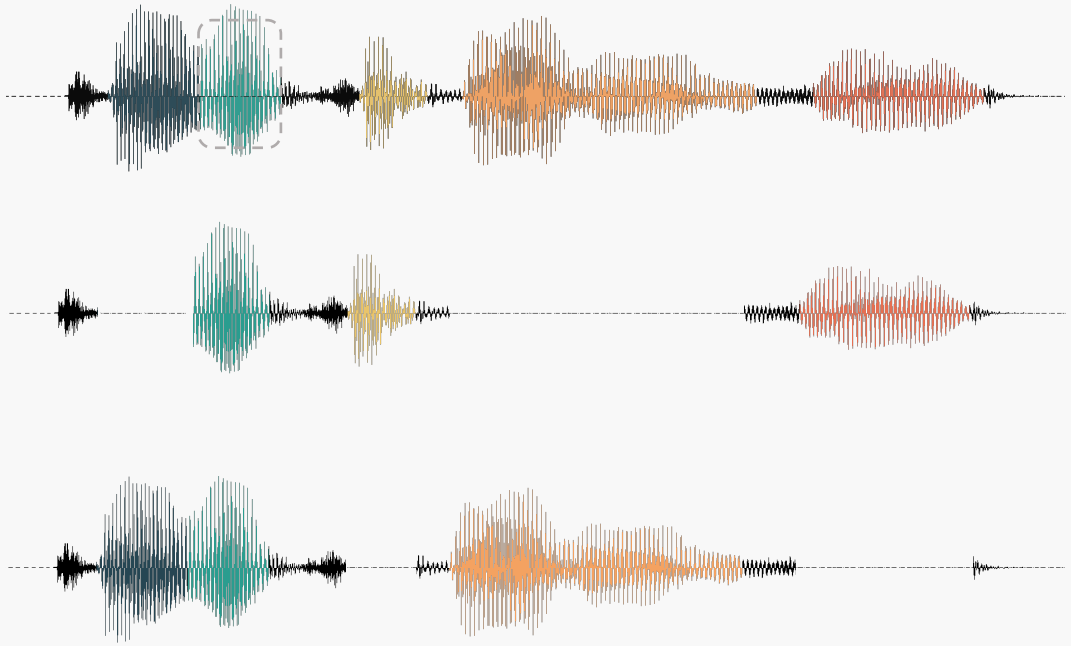


increase 98%

bedroom 85%

Turn up the bedroom heat

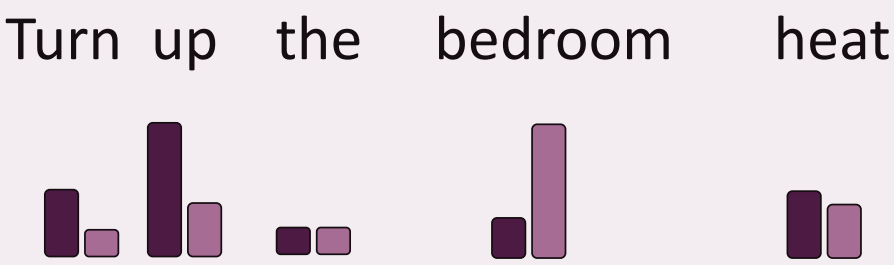
Mask audio segments



increase	bedroom
25%	85%
36%	0%
98%	80%

Aggregate feature impact

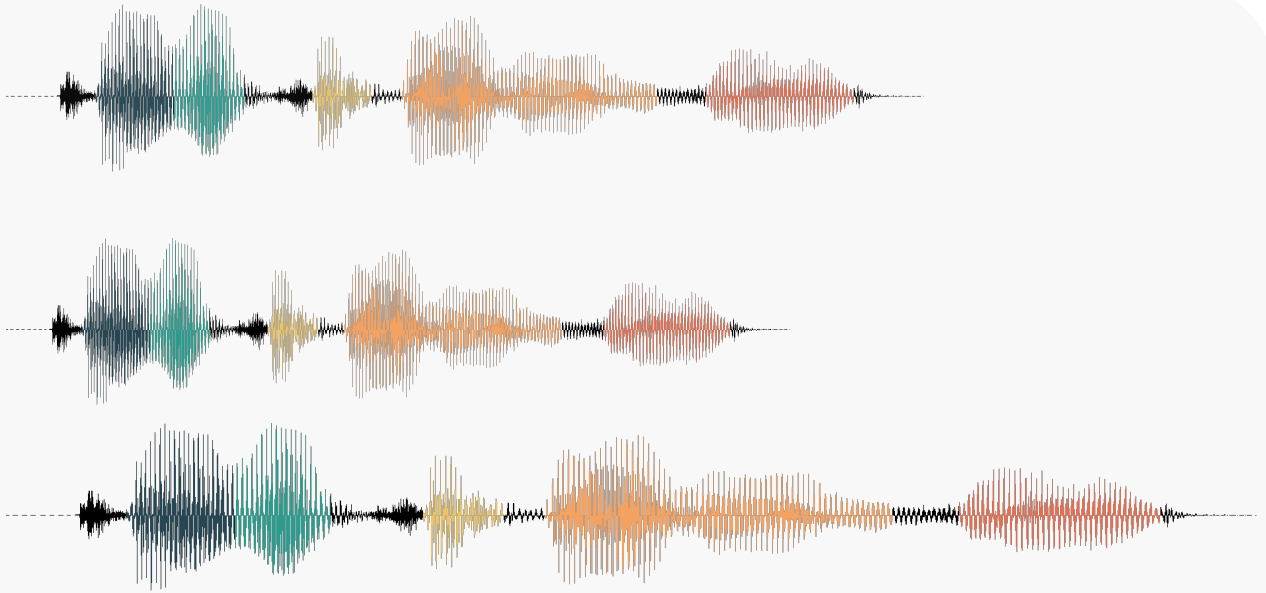
- Leave-one-out
- LIME



Word-level attributions

Paralinguistic

Perturb signal on paralinguistic



increase bedroom

30% 85%

20% 70%

98% 90%

Aggregate feature impact

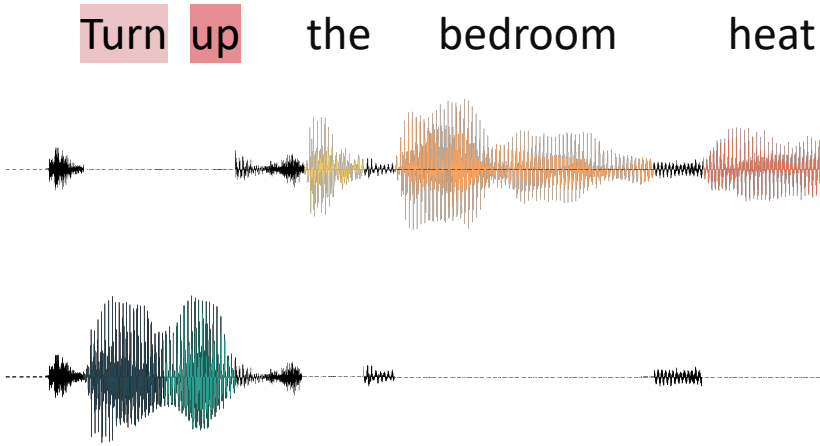
Time stretching

Speaking rate

Explanation evaluation

Faithfulness

Adhere to the model's inner working



- **Comprehensiveness** — Is it all?
- **Sufficiency** — Is it sufficient?

Our word-level explanations are **faithful**!

Explanation evaluation

Plausibility

Reasonable, believable, align with human reasoning

User study

- Plausibility of explanations
- Visualization preference

	Turn	up	the	bedroom	heat.
act=increase	0.250	0.545	0.260	0.139	0.021
obj=heat	0	0	0	0.014	0.550
loc=bedroom	0.002	0.006	0.087	0.997	0.323

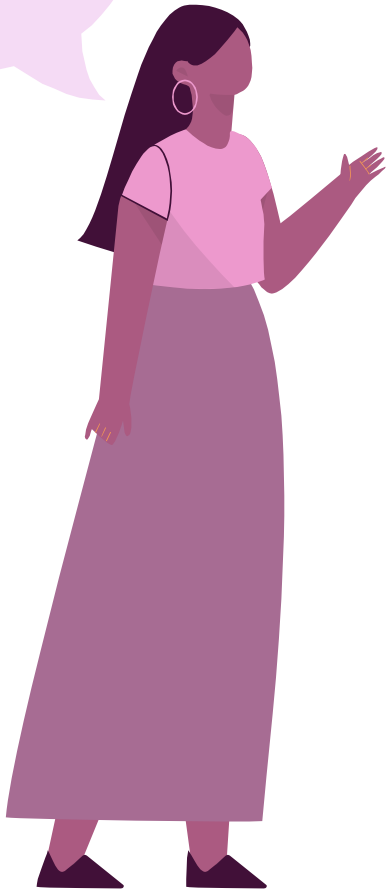
Users found explanations **plausible** + visualization **intuitive** & **straightforward**!

Try it!

elianap/SpeechXAI

```python
from speechxai import Benchmark
from transformers import Wav2Vec2ForSequenceClassification, Wav2Vec2FeatureExtractor


model = Wav2Vec2ForSequenceClassification.from_pretrained("superb/wav2vec2-base-superb-ic")

feature_extractor = Wav2Vec2FeatureExtractor.from_pretrained("superb/wav2vec2-base-superb-ic")



benchmark = Benchmark(model, feature_extractor)

explanation = benchmark.explain(audio_path=audio_path, methodology="LIME")


benchmark.show_table(explanation)
```

Thanks!



elianap/SpeechXAI

eliana.pastor@polito.it

eliana__pastor