# Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass *(MIT CSAIL & MIT-IBM Watson AI Lab)*

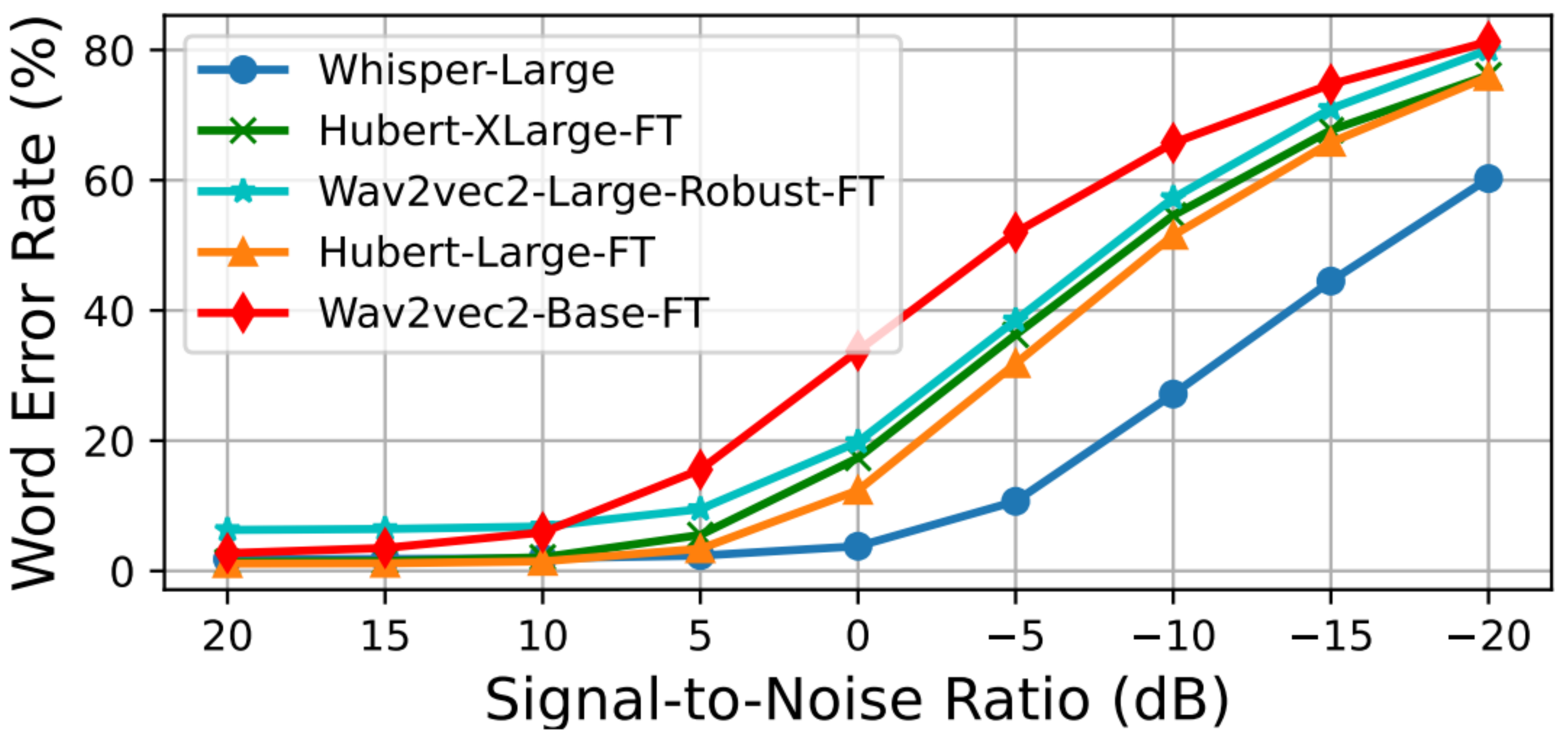
ID:2193

github.com/yuangongnd/whisper-at

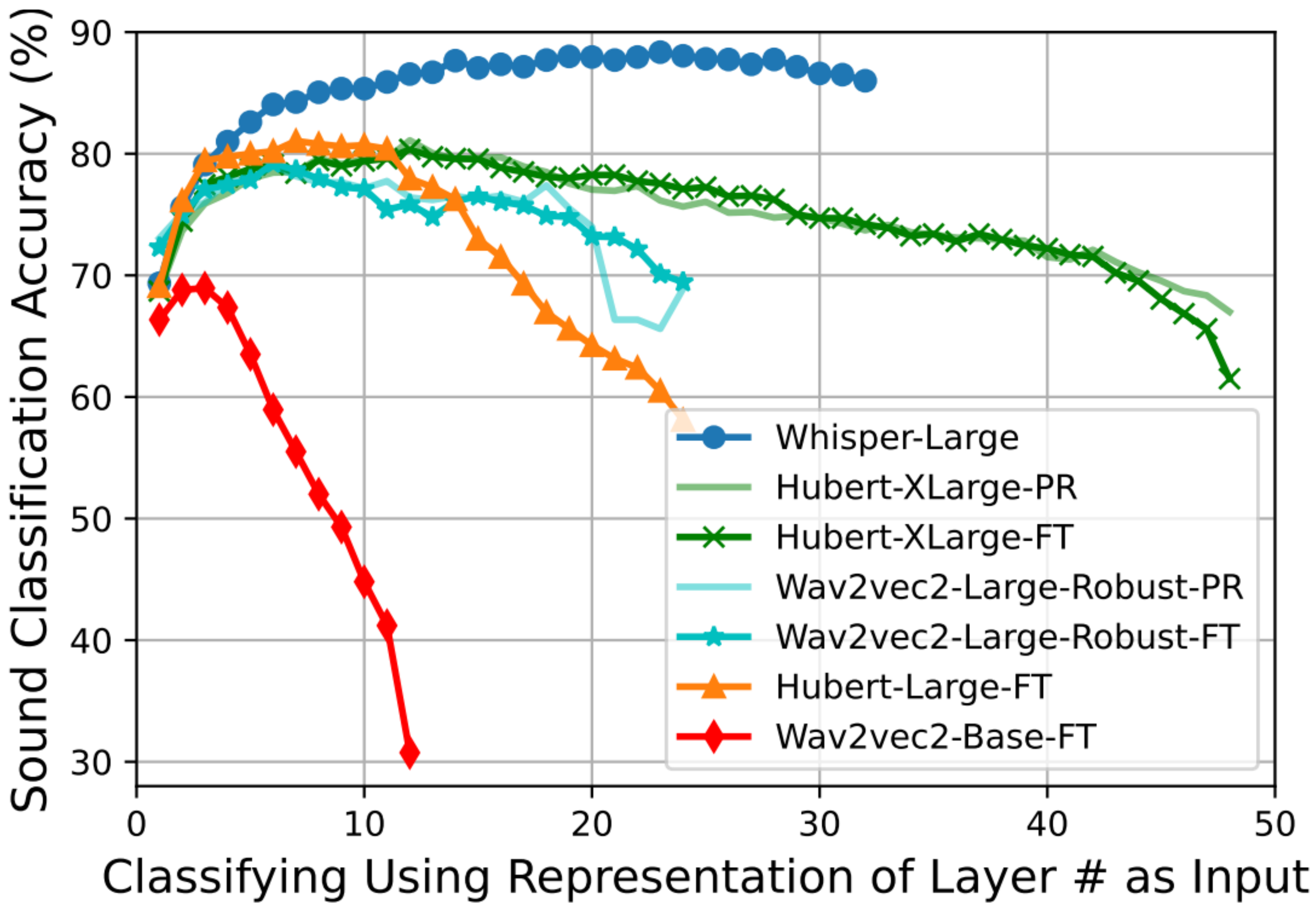
INTERSPEECH 2023

## An Intriguing Finding: Noise-Robust ASR Learns *Noise-Variant* Representations

**We usually believe a noise-robust ASR representation is noise-invariant, but it is NOT true for Whisper.**

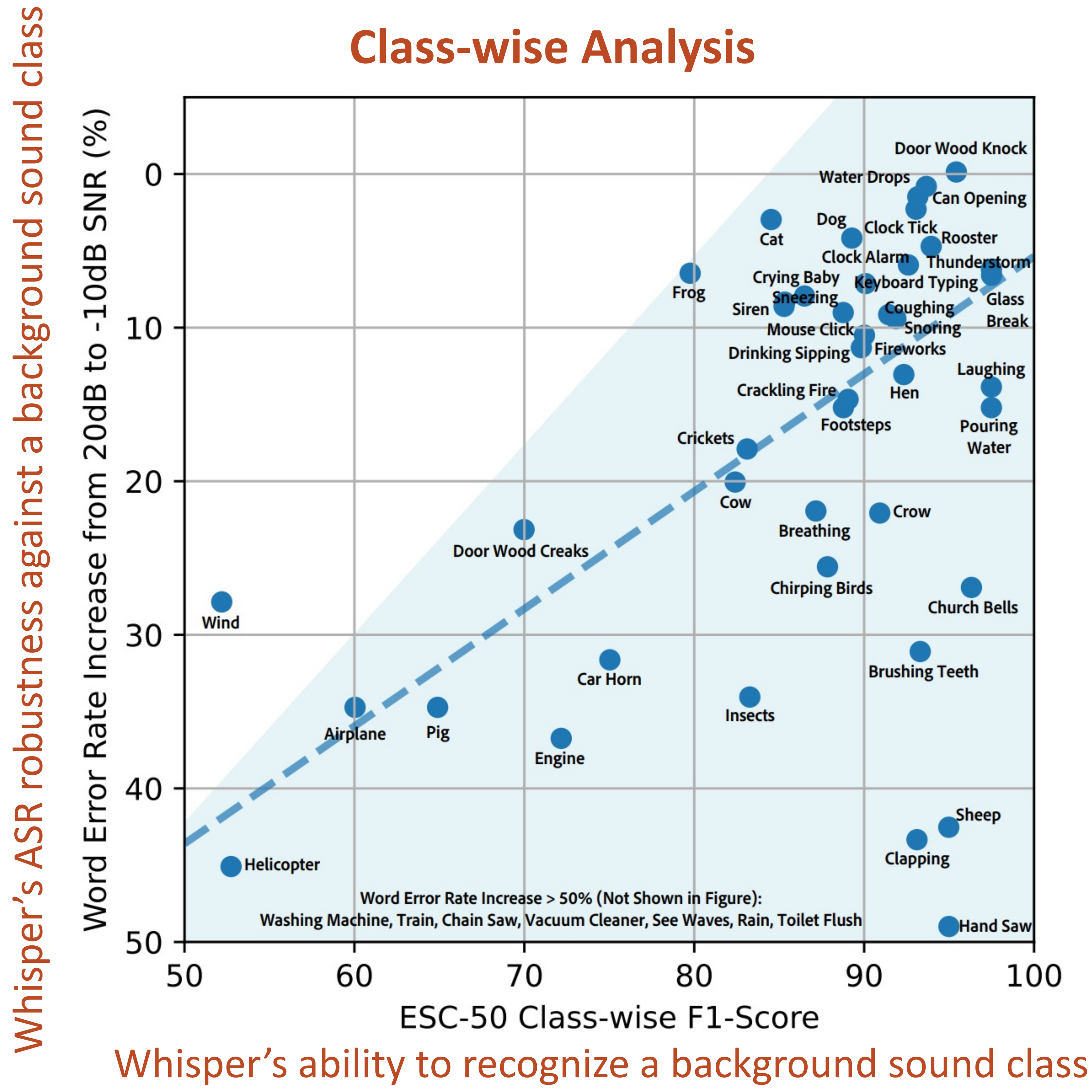



Whisper ASR is most robust against ESC-50 background sounds




Meanwhile, Whisper representations lead to the **best** linear probing background sound classification accuracy on ESC-50, indicating they encode **most background sound information**

### Class-wise Analysis



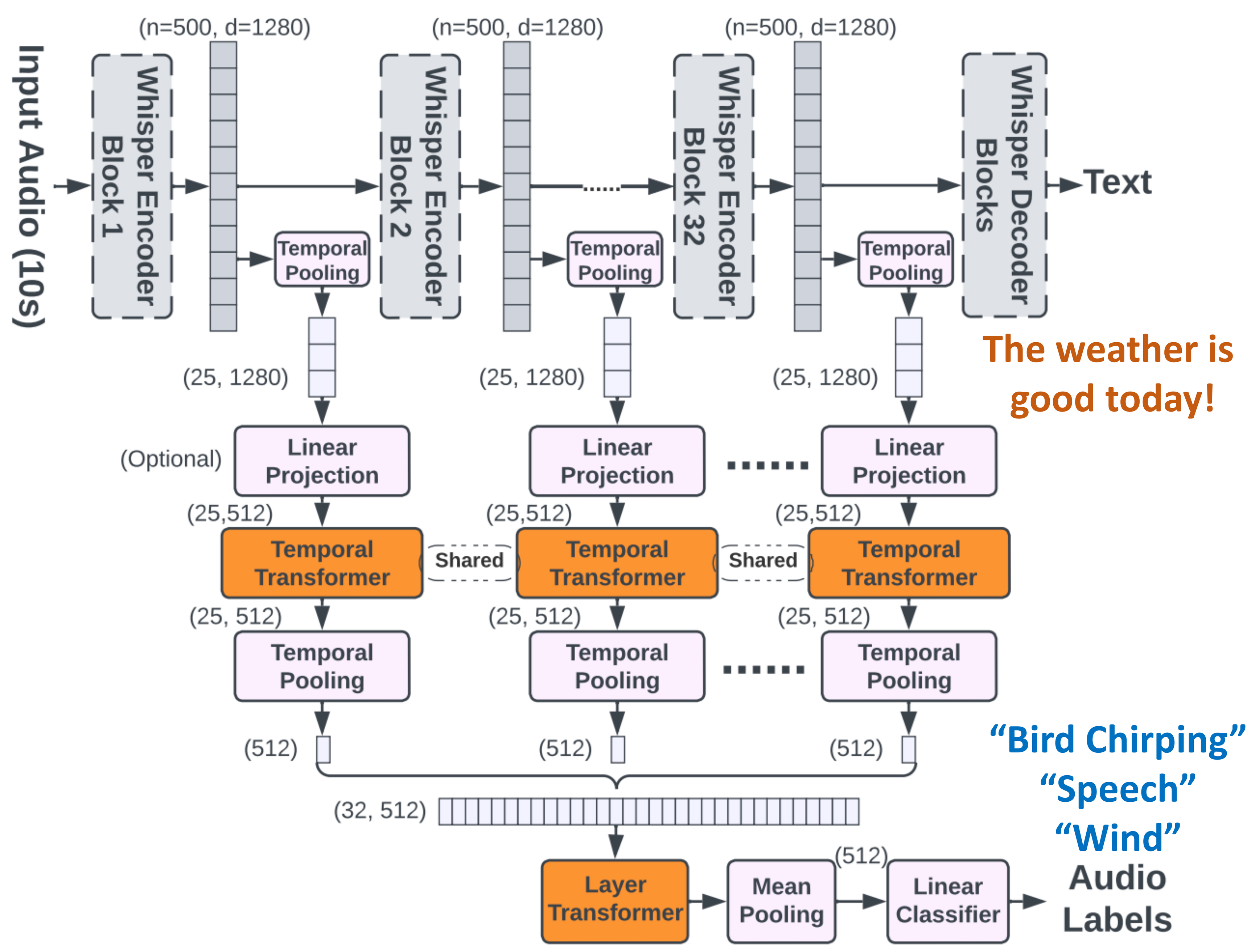

Whisper's ability to recognize a background sound class

The ability to recognize a background sound type is a necessary but not sufficient condition for Whisper to be robust to it.

**Key Insight:** A noise-robust ASR does not have to learn a noise-invariant representation, and there exists other ways to be noise-robust - a noise-conditioned model like Whisper can, and indeed does, work very well.

## Whisper-AT: A *Unified* Audio Tagging and Speech Recognition Model

### Model Architecture




- Whisper model is ***frozen***, so Whisper ASR performance is not impacted.
- Time and layer-wise Transformer (TLTR) to capture information from representations of ***all*** 32 layers.
- TLTR is also a strong model for other audio classification tasks (e.g., speech emotion classification).

### Results

| Model | Audio Tagging | | | | | ASR |
|---|---|---|---|---|---|---|
| | AS-20K | AS-2M | ESC-50 | AT Params | AT Speed-Up | |
| AudioSet Baseline | - | 31.4 | - | - | - | N/A |
| AST | 34.7 | 45.9 | 88.8 | 87M | 1X (133G FLOPs) | N/A |
| Whisper-AT | 32.8 | 41.5 | 91.7 | 7M | 42X | Same as Whisper |

- Whisper-AT has the same ASR performance as Whisper.
- Whisper-AT has comparable Audio Tagging performance to AST, while being 12X smaller and 42X faster for the audio tagging task.
- With <1% extra computational cost to ASR cost, Whisper-AT can recognize audio events, in addition to spoken text, in a single forward pass.
- Same API as Whisper, easy to implement.

```
# Implement in 6 lines of code:
! pip install whisper-at
import whisper_at as whisper
model = whisper.load_model("large")
result = model.transcribe("audio.mp3")
audio_tag_result =
whisper.parse_at_label(result)
print(result["text"], audio_tag_result)
```

Github

HuggingFace

## Acknowledgement

This research is supported by the MIT-IBM Watson AI Lab