BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025

# Endpointing for spoken dialogue

Author: Sathvik Udupa

Supervisor: Petr Schwarz

**Endpointing: when did a speaker stop speaking?**

## Mel Spectrogram ~ 80 kbps 🙁🙁🙁

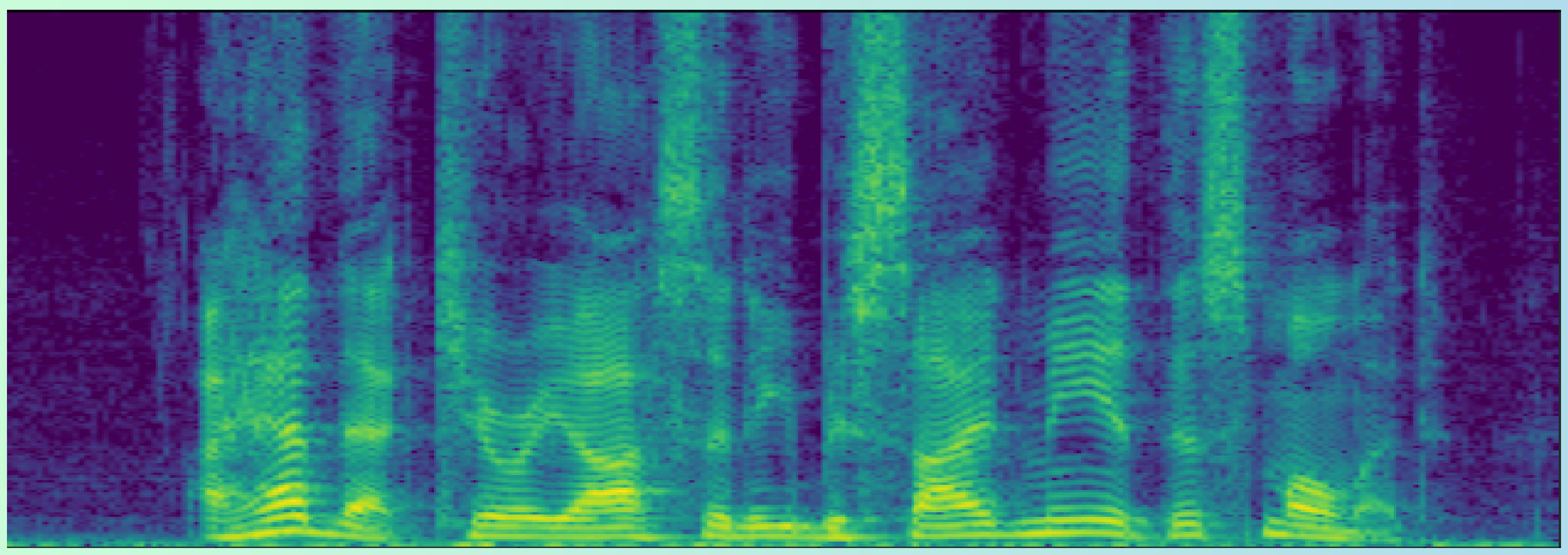

Fig 1: Mel Spectrogram

## Mimi [1] - neural audio codec ~ 1.1kbps 🙂🙂🙂

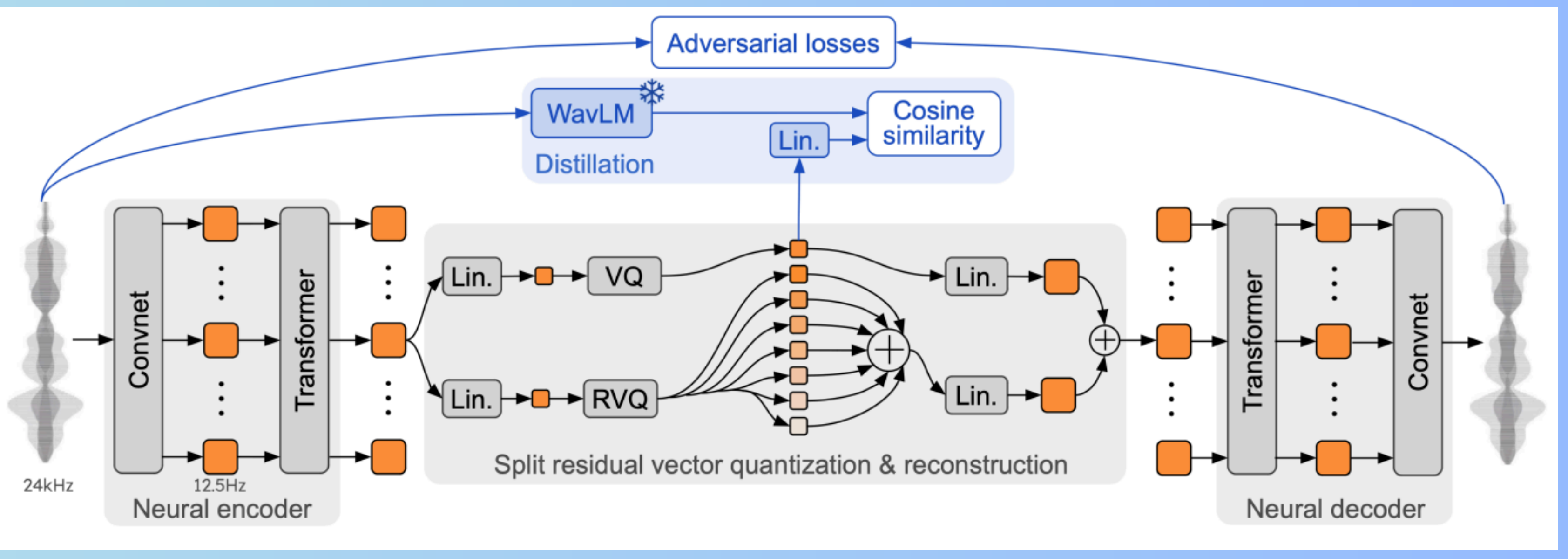

Fig 2: Mimi Codec

## Label delay training - latency / error tradeoff

$$\mathbf{Y}_\tau = \begin{cases} k & \text{if } t \le \tau \\ y_{t-\tau} & \text{if } t > \tau \text{ and } t \le T \end{cases}$$

Fig 3: Label delay

## Endpointer - latency vs error

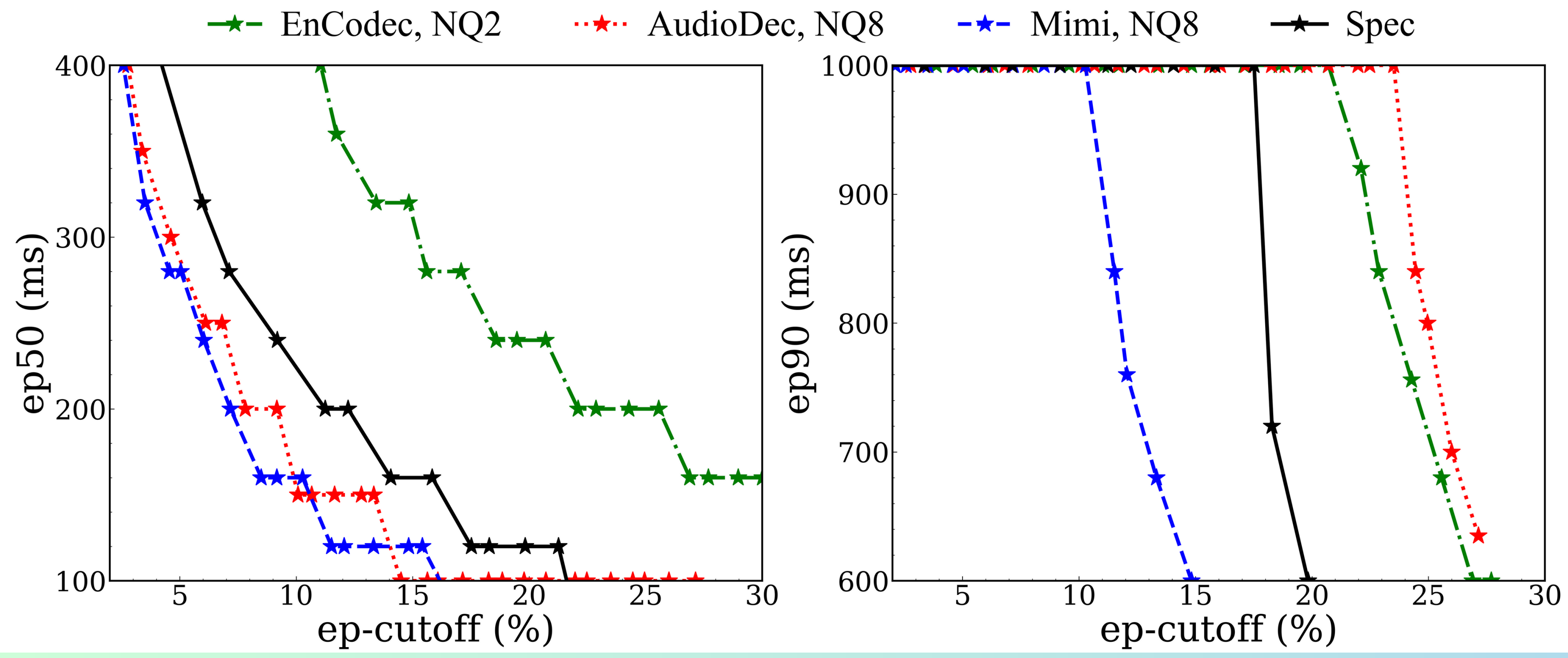

Fig 5: Baseline vs Mimi features

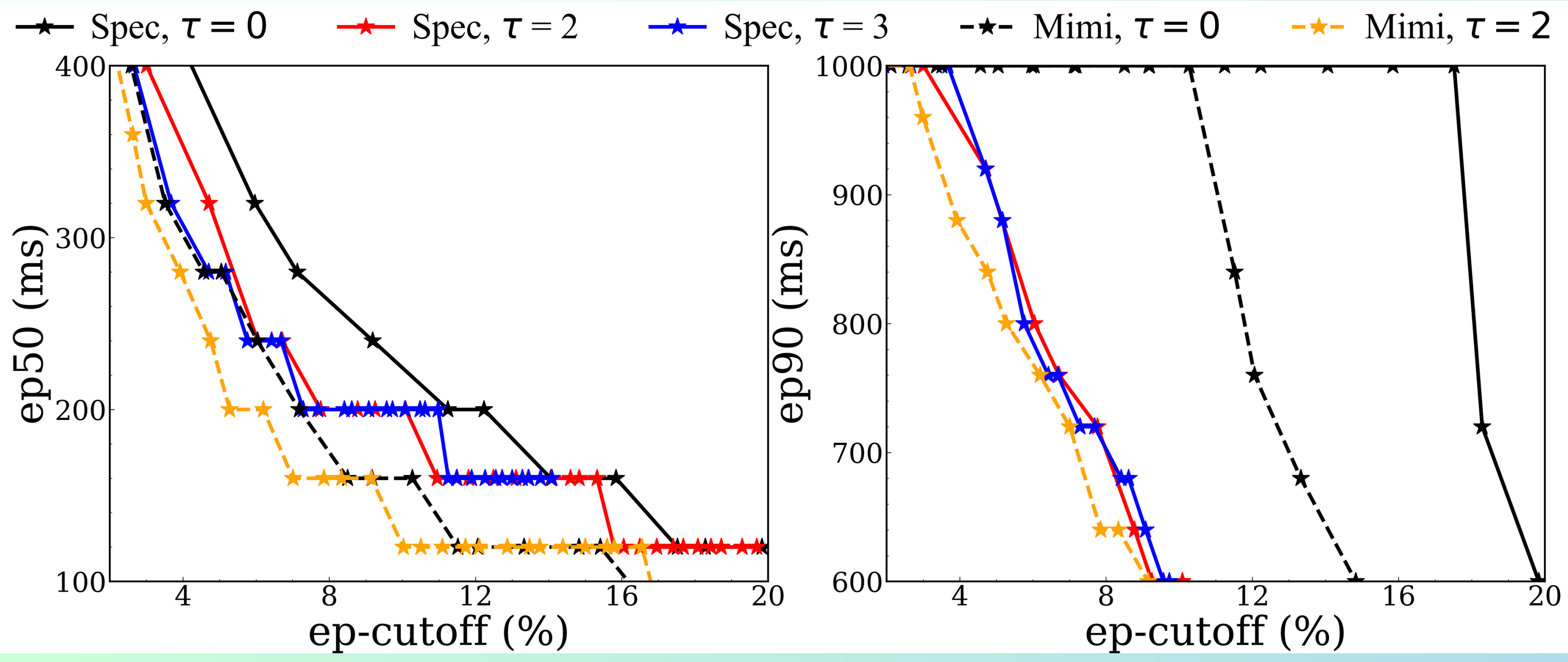

Fig 6: Results with label delay training

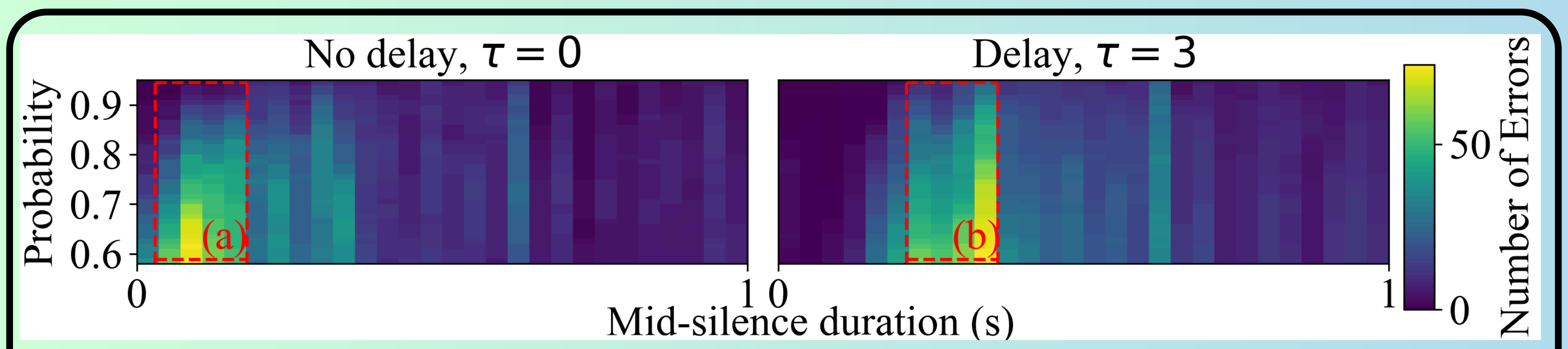

Fig 8: Error locations with and without label delay

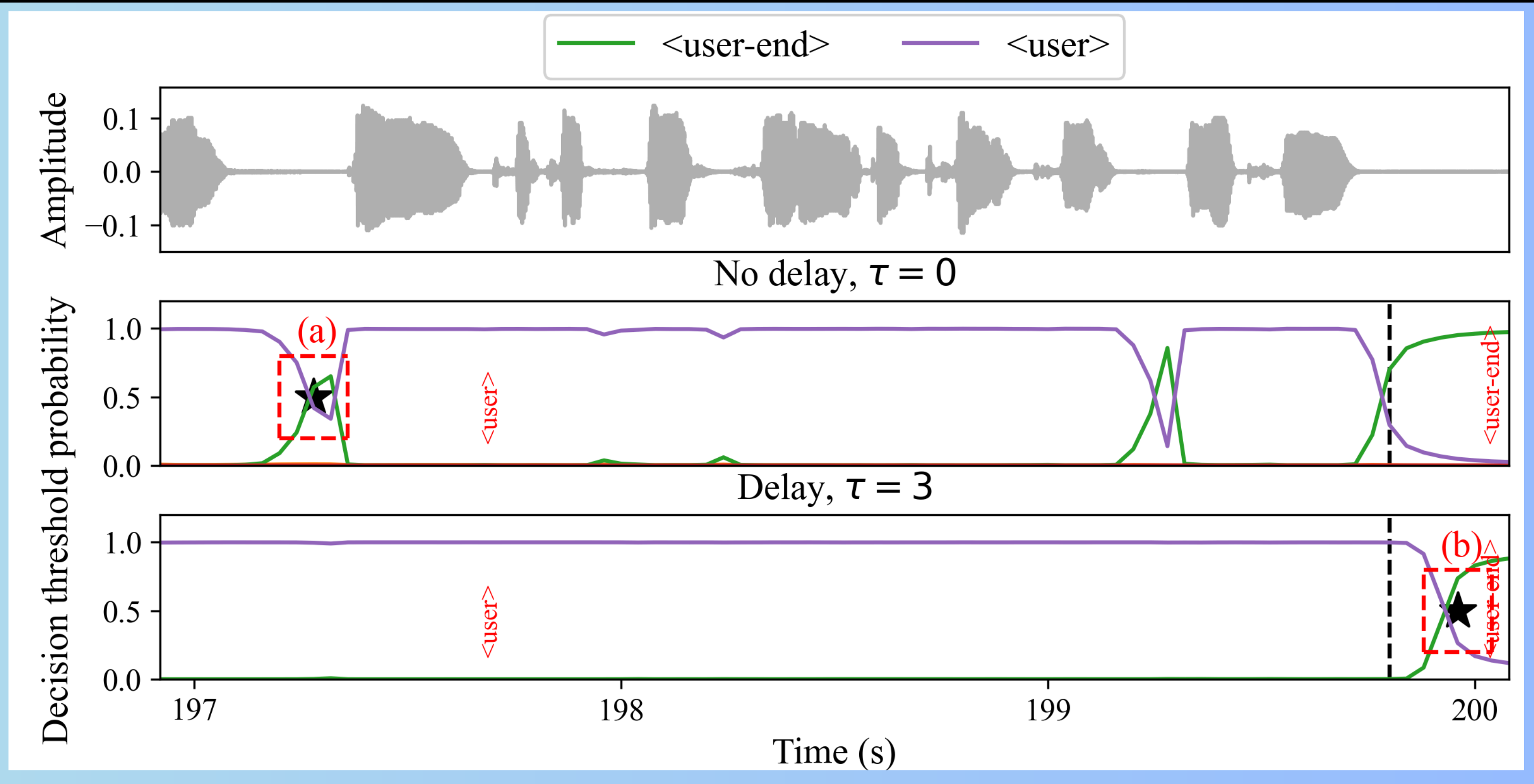

Fig 4: Prediction visualisation

## Integration with Speech LLM - Moshi [1]

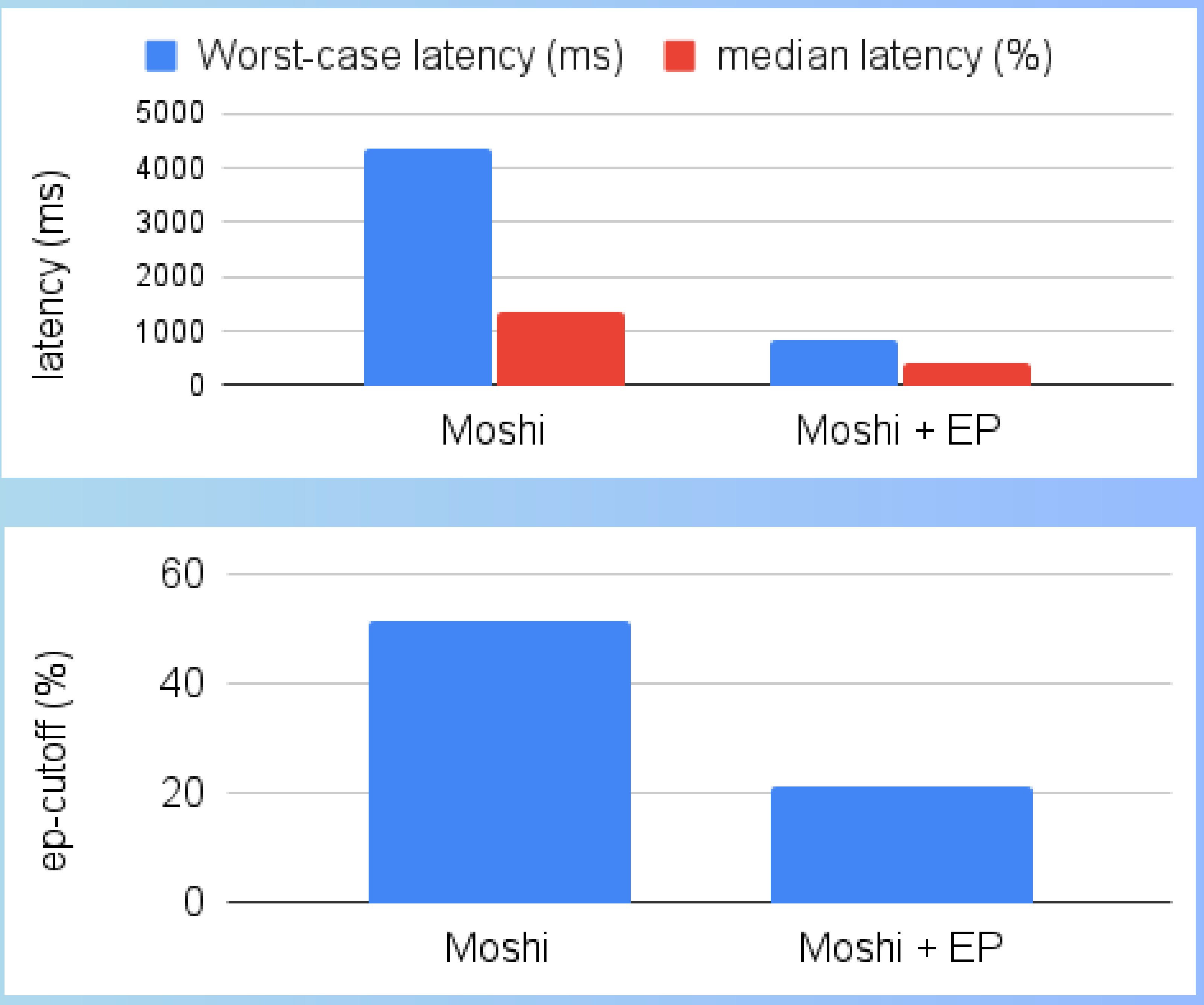

Fig 7: Results with speech LLM

## Summary (at fixed 200ms latency):

- Mimi - 46% relative error reduction
- label delay - 32% relative reduction
- Efficient integration for speechLLM

[1] Défossez, Alexandre, et al. "Moshi: a speech-text foundation model for real-time dialogue." arXiv preprint arXiv:2410.00037 (2024).