Transcription-Free Filler Word Detection with Neural Semi-CRFs

Ge Zhu[1], Yujia Yan[1], Juan-Pablo Caceres[2] and **Zhiyao Duan**[1]

1 Department of Electrical and Computer Engineering
2 Adobe Research

What are filler words?

“In linguistics, a filler is a sound or word that participants in a conversation use to signal that they are pausing to think but are not finished speaking.”

----Wikipedia

In American English, the most common filler sounds are uh and um.

Heather Bortfeld et al. "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender." *Language and speech,* 2001, pp: 123-147.

Motivation

Improving Speech Recognition

Media Editing

Alzheimer's Disease Biomarker

Deception Marker

1.Sharon Goldwater et al. “Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase ASR error rates,” in *Proc. of ACL*, 2008, pp. 380–388.

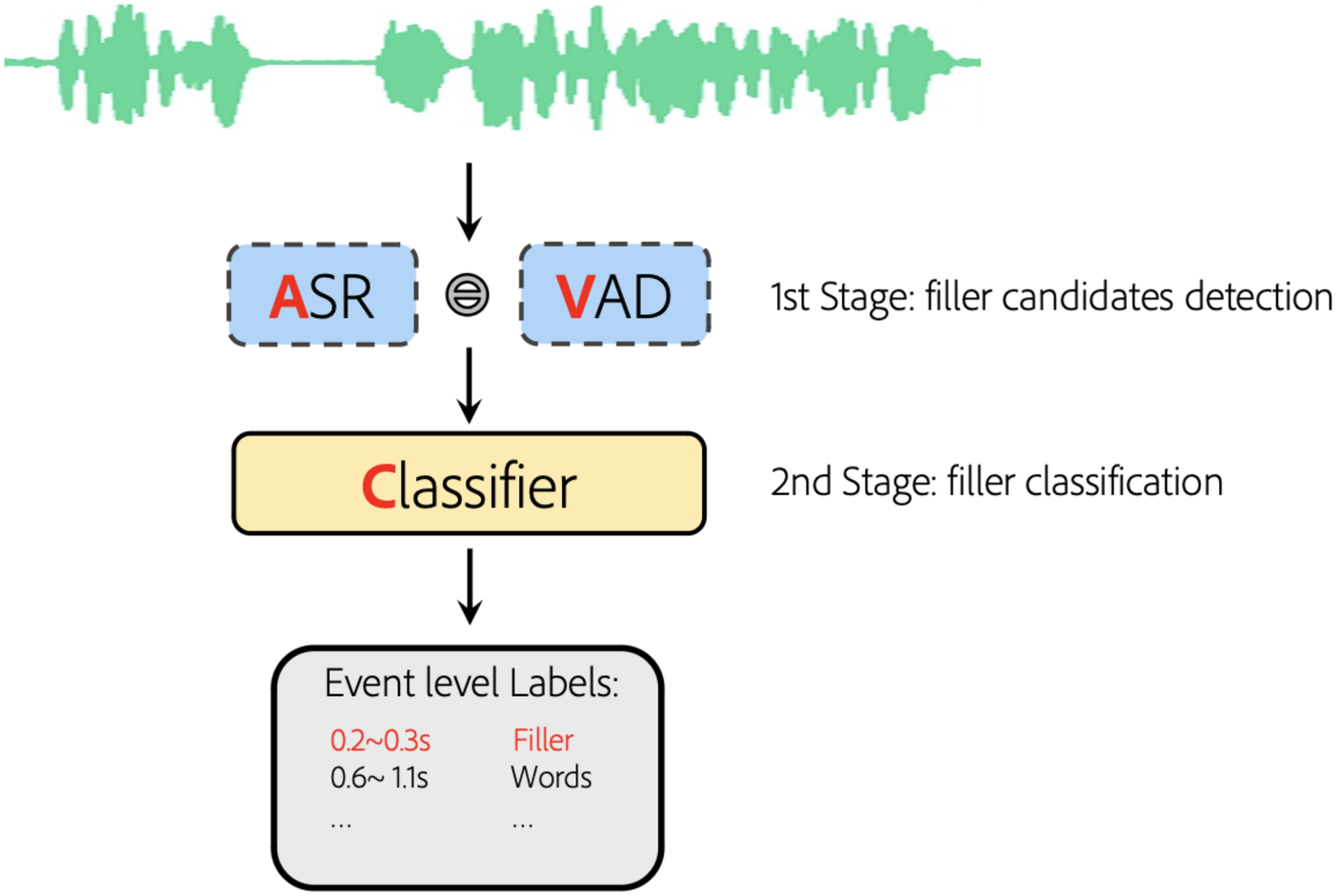
2. https://podcast.adobe.com/, https://www.descript.com/filler-words

3. Jiahong Yuan et al. “Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer’s disease.,” in *Proc. Interspeech*, 2020, pp. 2162–2166.

4. Joanne Arciuli et al. “Um, I can tell you’re lying”: Linguistic markers of deception versus truth- telling in speech,” *Applied Psycholinguistics* , vol. 31, no. 3, pp. 397–411, 2010.

Previous Work

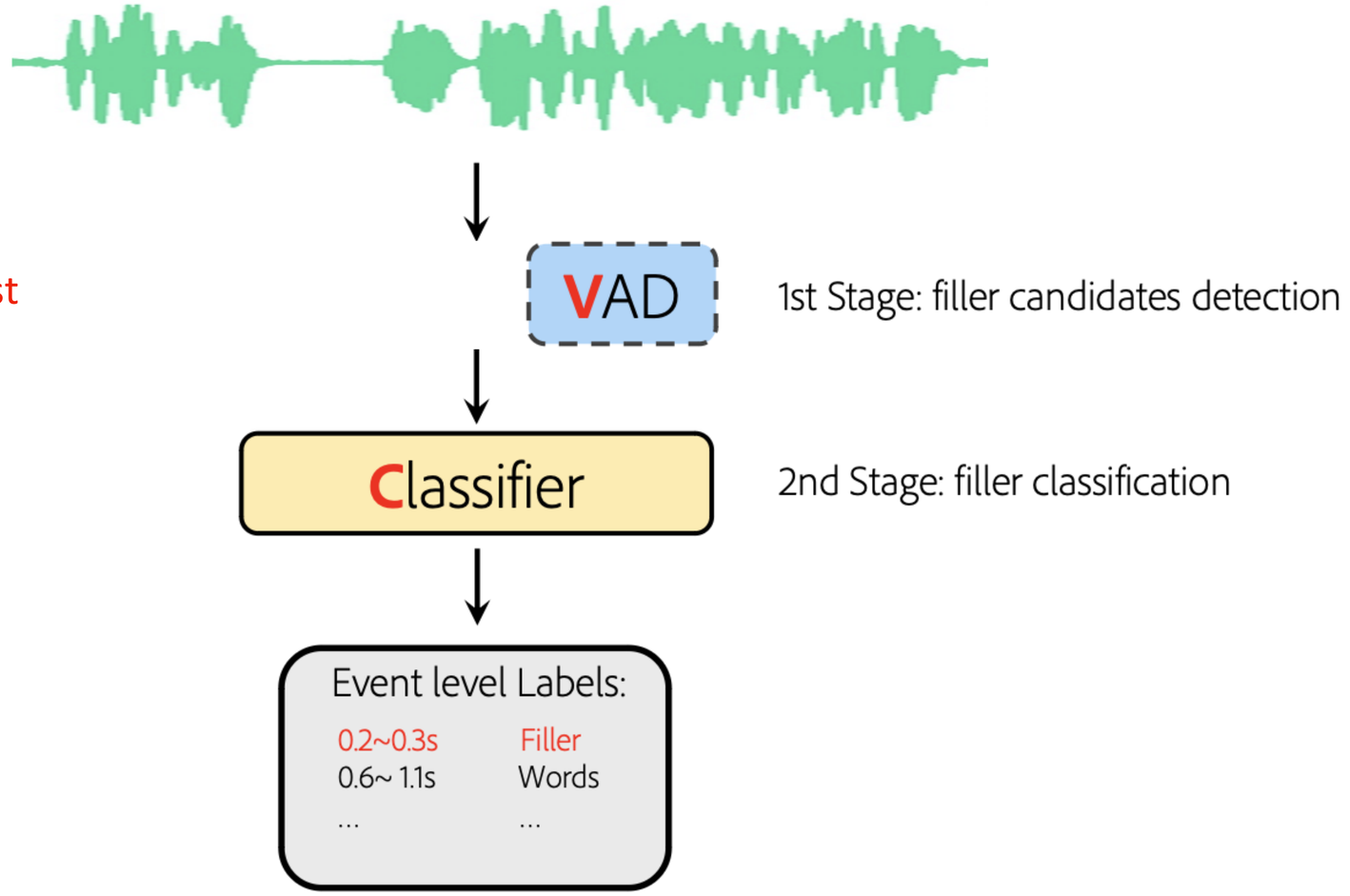
Limitation: verbatim ASR systems can be expensive and unreliable



Ge Zhu et al. “Filler Word Detection and Classification: A Dataset and Benchmark.” in *Proc. Interspeech*, 2022. pp. 3769-3773.

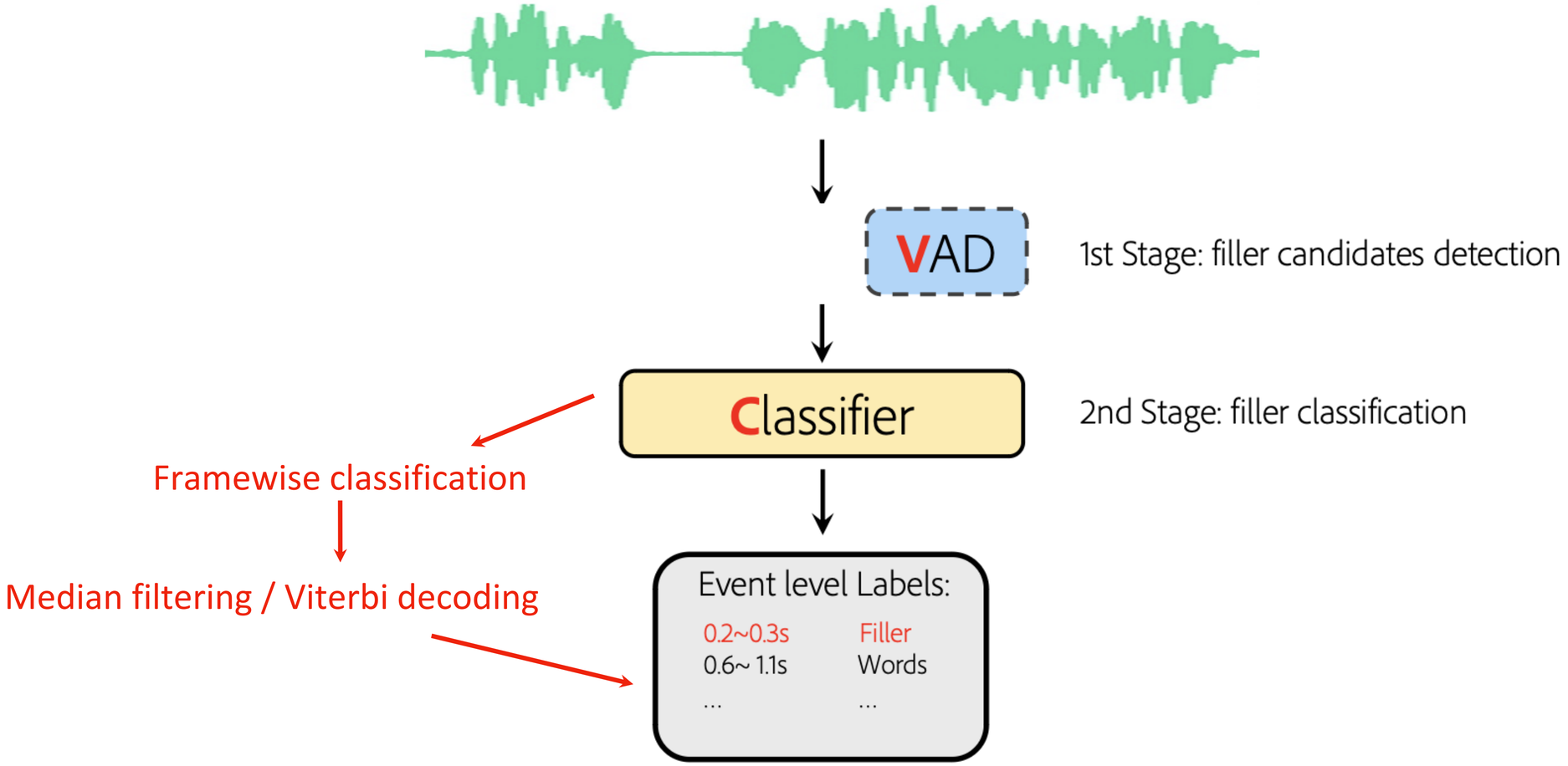
Previous Work

Transcription-free solution:
Remove the ASR module at the cost of a lower classification accuracy



Ge Zhu et al. “Filler Word Detection and Classification: A Dataset and Benchmark.” *Interspeech*, 2022. pp. 3769-3773.

Potential problem in transcription-free systems: post-processor tuning



Improvement 1: directly output event-level labels

Neural Semi-Markov Conditional Random Field (Semi-CRF)

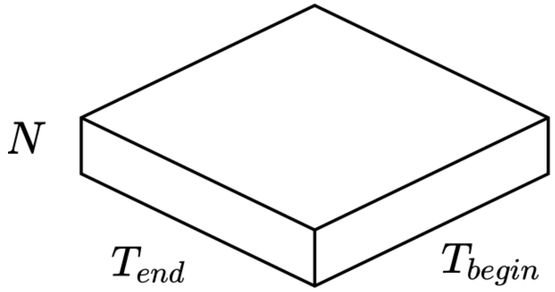
$$\mathbf{x} = \{x_1, ..., x_n\}$$

Neural Networks

Framewise Embedding

$$\mathbf{w} = \{w_1, ..., w_n\}$$

$$s_i = (b_i, e_i, l_i)$$

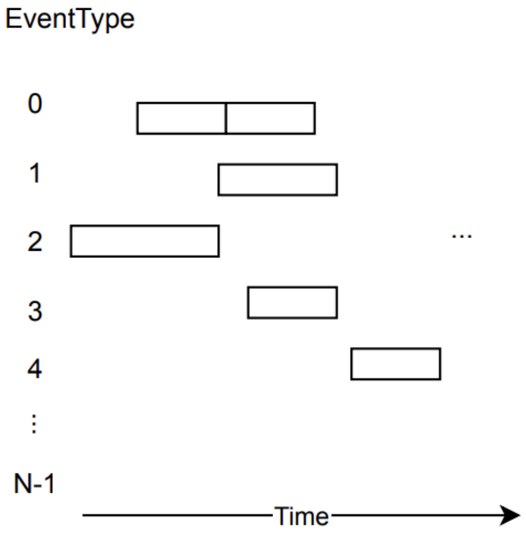


Score

Score Model

$$\mathrm{p}(\hat{\mathbf{s}}|\mathbf{w}) = \frac{\mathrm{score}(\hat{\mathbf{s}}, \mathbf{w})}{\sum_{\mathbf{s}' \in \mathbf{S}} \mathrm{score}(\mathbf{s}', \mathbf{w})}$$

Decoding

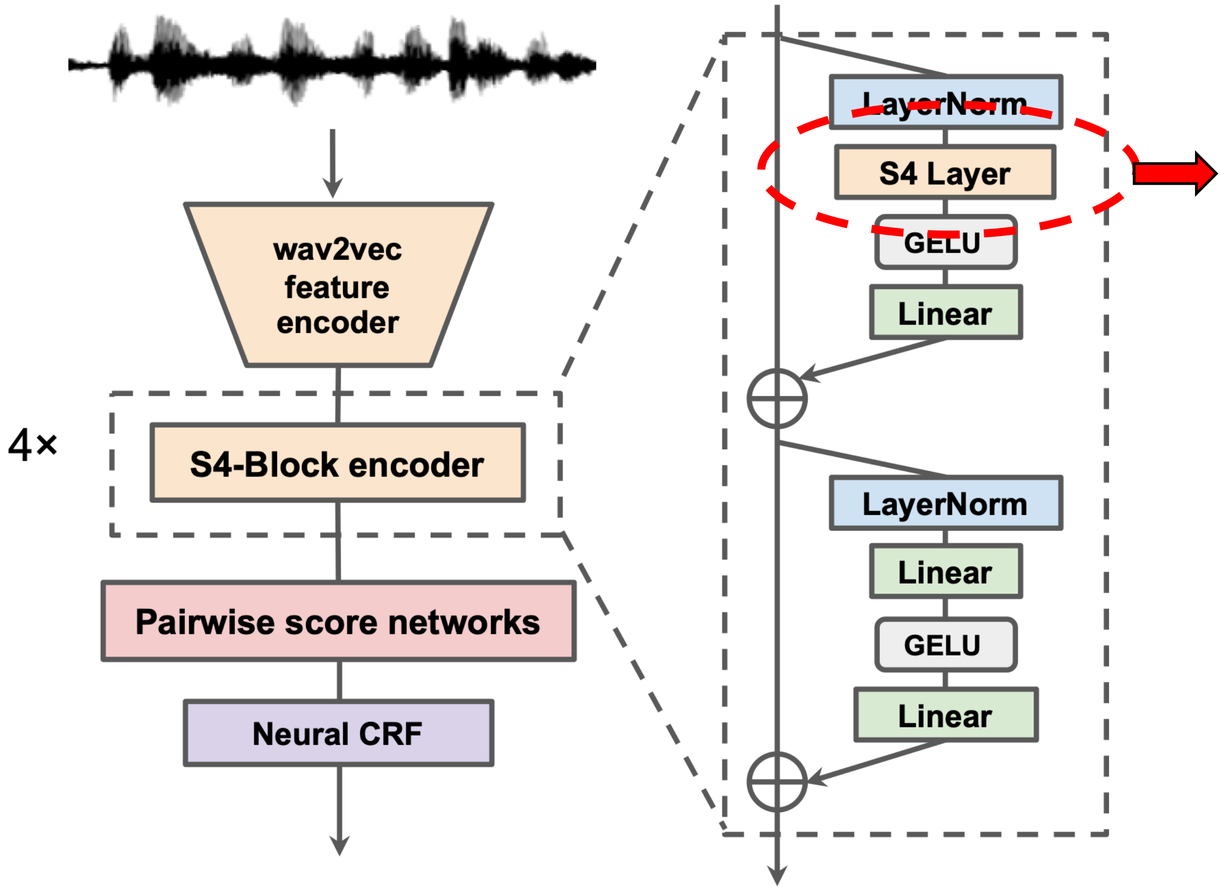


Yujia Yan et al. "Skipping the frame-level: Event-based piano transcription with neural semi-CRFs." in Proc. *NeurIPS* 2021, pp. 20583-20595

Zhixiu Ye et al. ”Hybrid semi-Markov CRF for neural sequence labeling.” in *Proc. ACL,* 2018, vol.2, pp: 235–240.

Improvement 2: better sequence modeling backbone

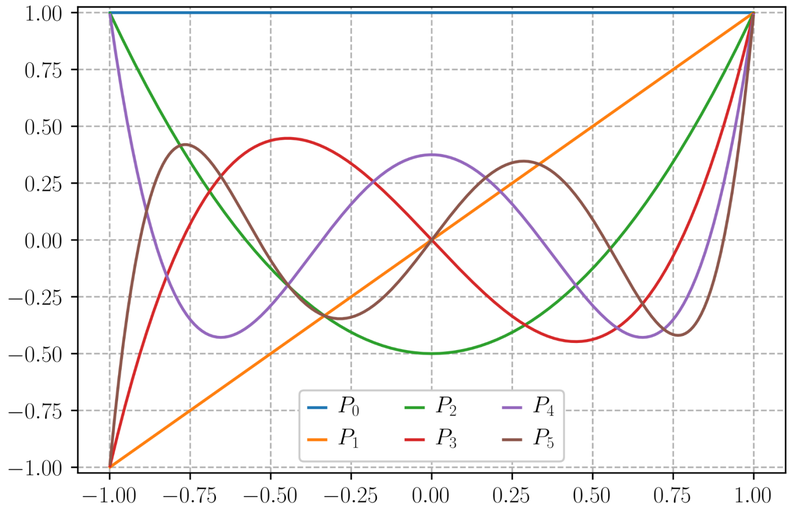
Structured State Space Sequence Models (S4)



1. State Space Sequence Models:

$$x'(t) = \boldsymbol{A}x(t) + \boldsymbol{B}u(t)$$
$$y(t) = \boldsymbol{C}x(t)$$

2. Compress sequence with Legendre series (Structured)



Albert Gu et al. "Efficiently modeling long sequences with structured state spaces." in *Proc. ICLR* 2022.

Experimental Setup

Dataset: PodcastFillers

- 145 hours of podcast episodes in English from SoundCloud
- ~35k filler words (“uh” and “um”)
- ~50k non-filler events (“words”, “repetitions”, ”breaths”, “music”, “laughter”, “agree”, “overlap” and “noise”)

Front-end feature: pre-trained wav2vec frame encoders

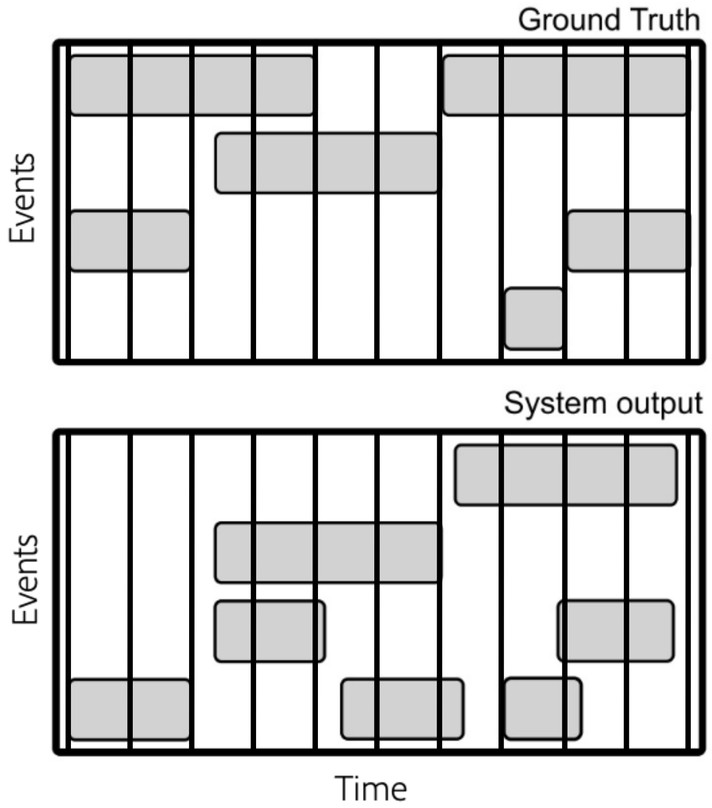
Training:

- Use ground-truth VAD
- Only train the filler classifier (S4 + semi-CRF), but with multiple event labels

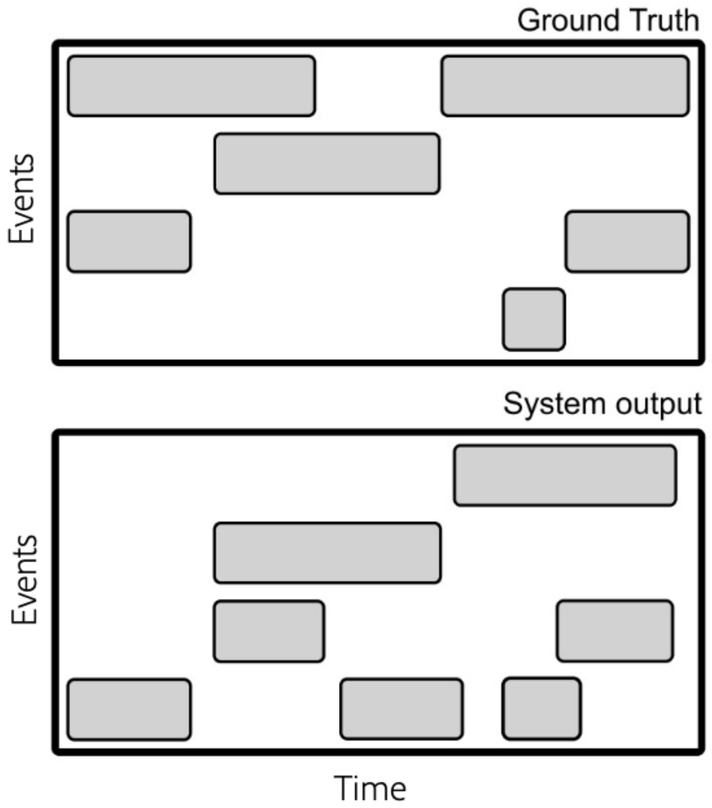
Inference: use a pretrained robust VAD + filler classifier

https://podcastfillers.github.io/

Evaluation metrics: F1 measure



Segment-based



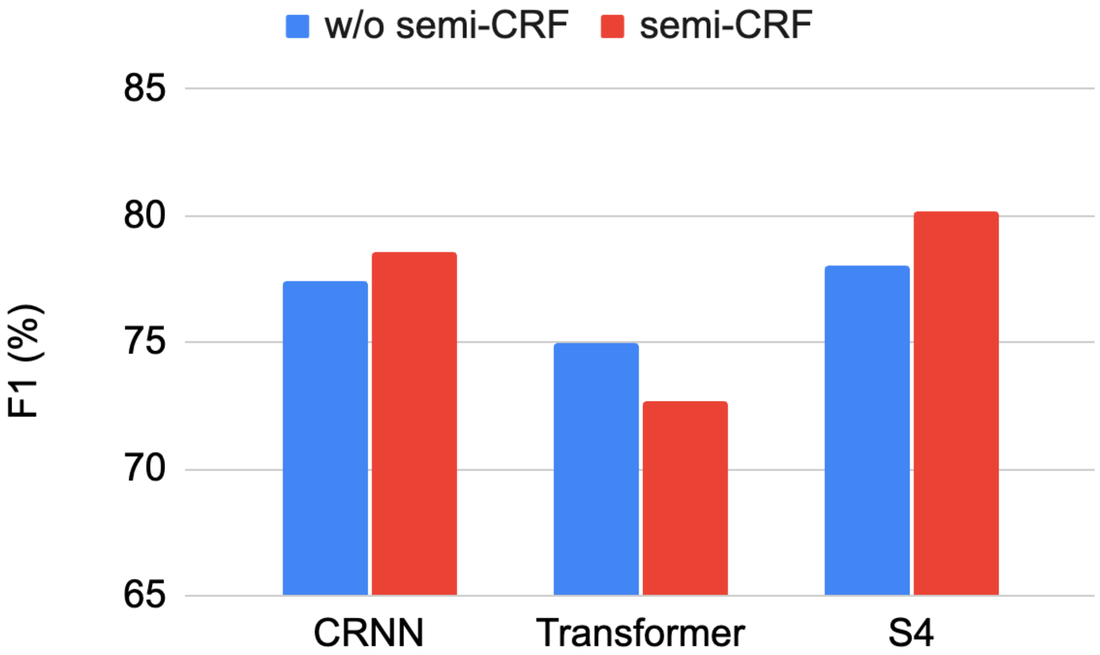
Event-based

Annamaria Mesaros et al, “Metrics for polyphonic sound event detection”, *Applied Sciences*, 2016, 6(6) pp:162

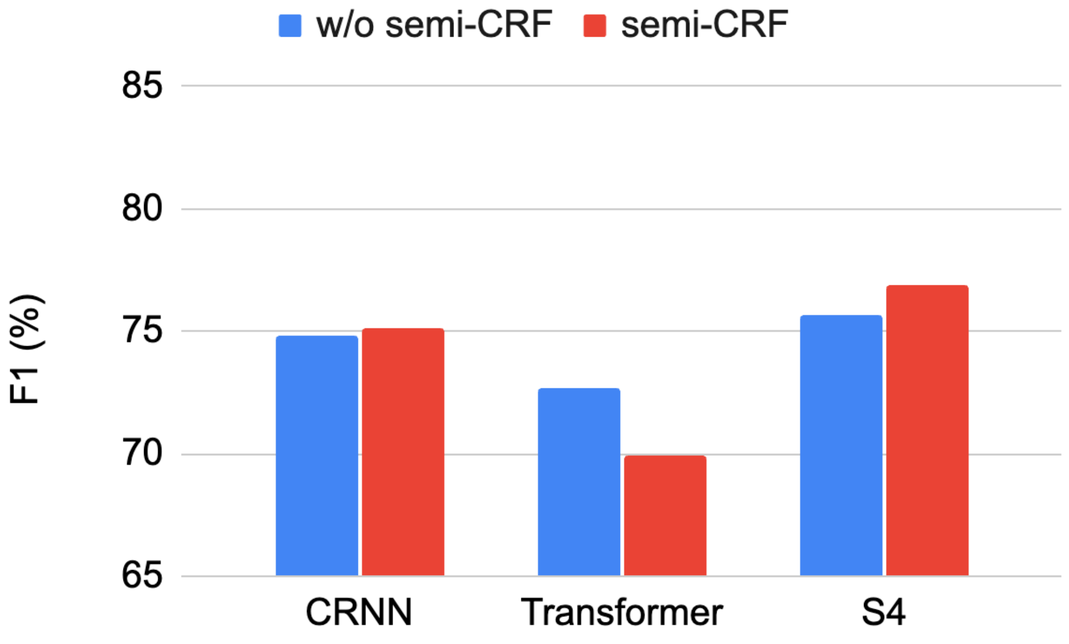
Comparison with State of the Art

System	Transcription based?	Segment			Event		
		Precision	Recall	F1	Precision	Recall	F1
AVC-FillerNet	**Yes**	**93.0**	**95.4**	**94.2**	**91.7**	**94.0**	**92.8**
VC-FillerNet	No	78.4	69.7	73.8	74.8	76.9	73.8
VC-S4CRF (proposed)	No	80.2	80.1	80.2	79.5	74.5	76.9

Ablation 1: On Neural Semi-CRF



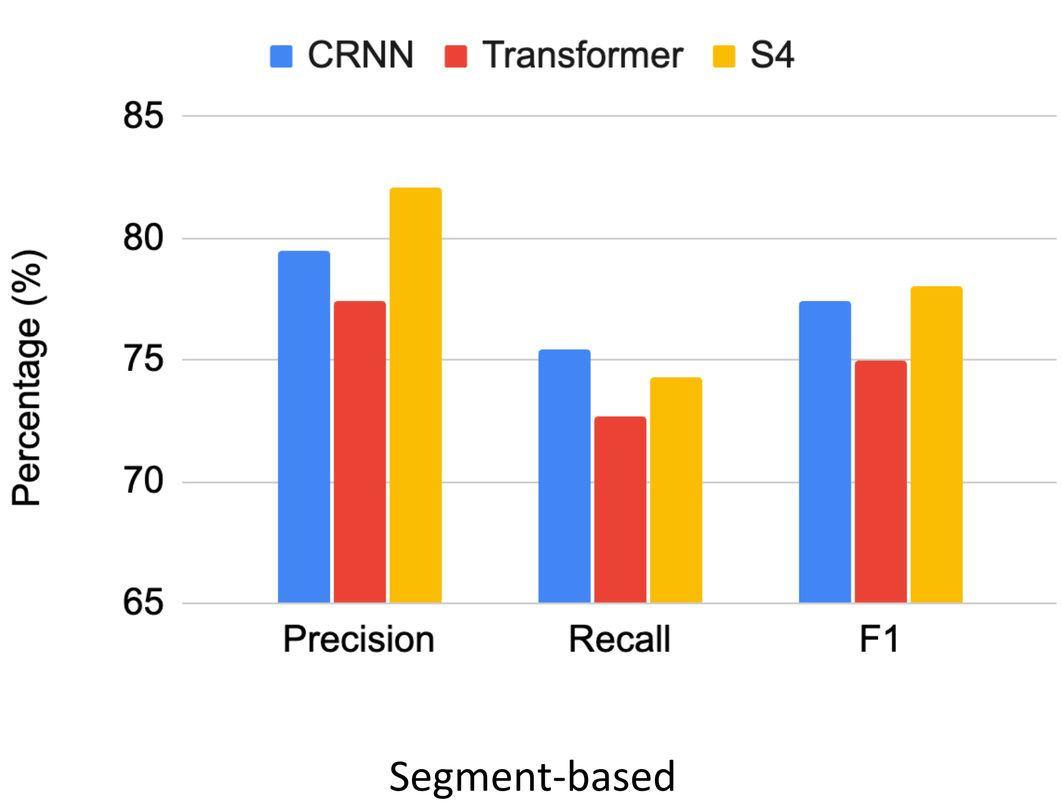
Segment-based



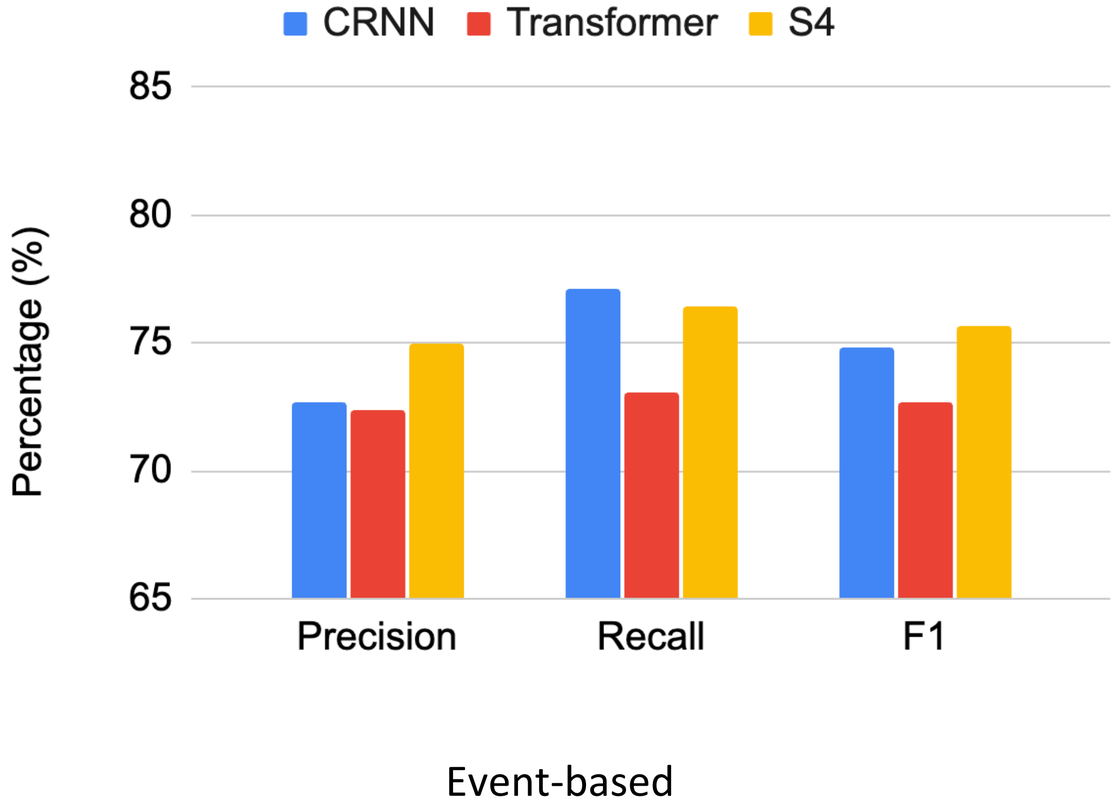
Event-based

*In systems without semi-CRF layers, we applied median filtering on the framewise outputs and tuned the decision thresholds on the validation split of PodcastFillers

Ablation 2: On Backbones (without semi-CRF)



Segment-based



Event-based

Conclusions

In this paper, we improved the transcription-free filler word detection systems by:

(1) introducing neural semi-CRF to directly output event-level labels, which outperforms framewise confidence + post median filtering pipeline

(2) introducing S4 models as embedding backbone, which outperforms the widely used CRNN backbones proposed in SED

Thank you!