CASTING TO CORPUS: SEGMENTING AND SELECTING SPONTANEOUS DIALOGUE FOR TTS WITH A CNN-LSTM SPEAKER-DEPENDENT BREATH DETECTOR

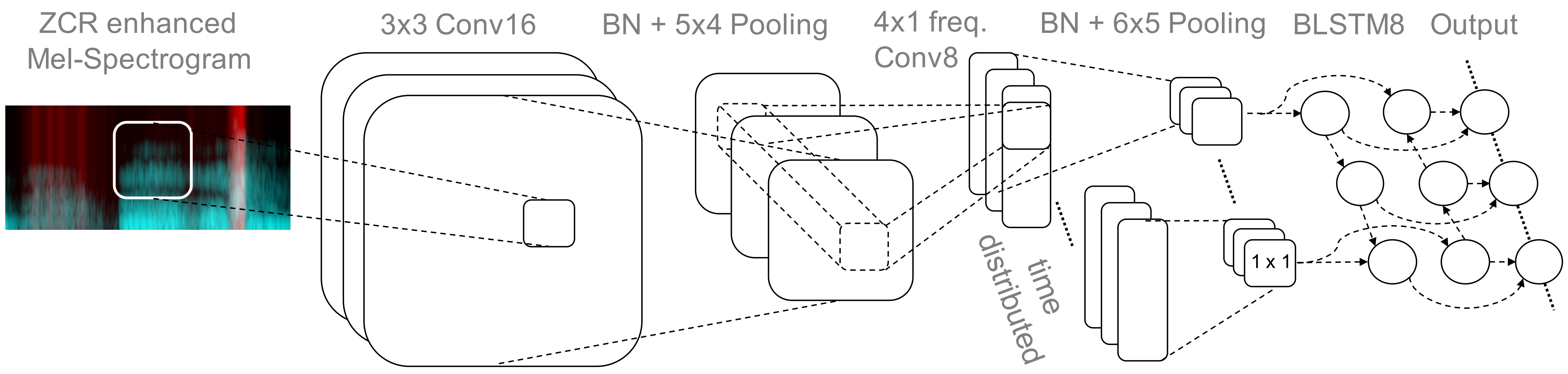
Éva Székely, Gustav Eje Henter, Joakim Gustafson

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

Aim: utilising breath events to create corpora for spontaneous TTS

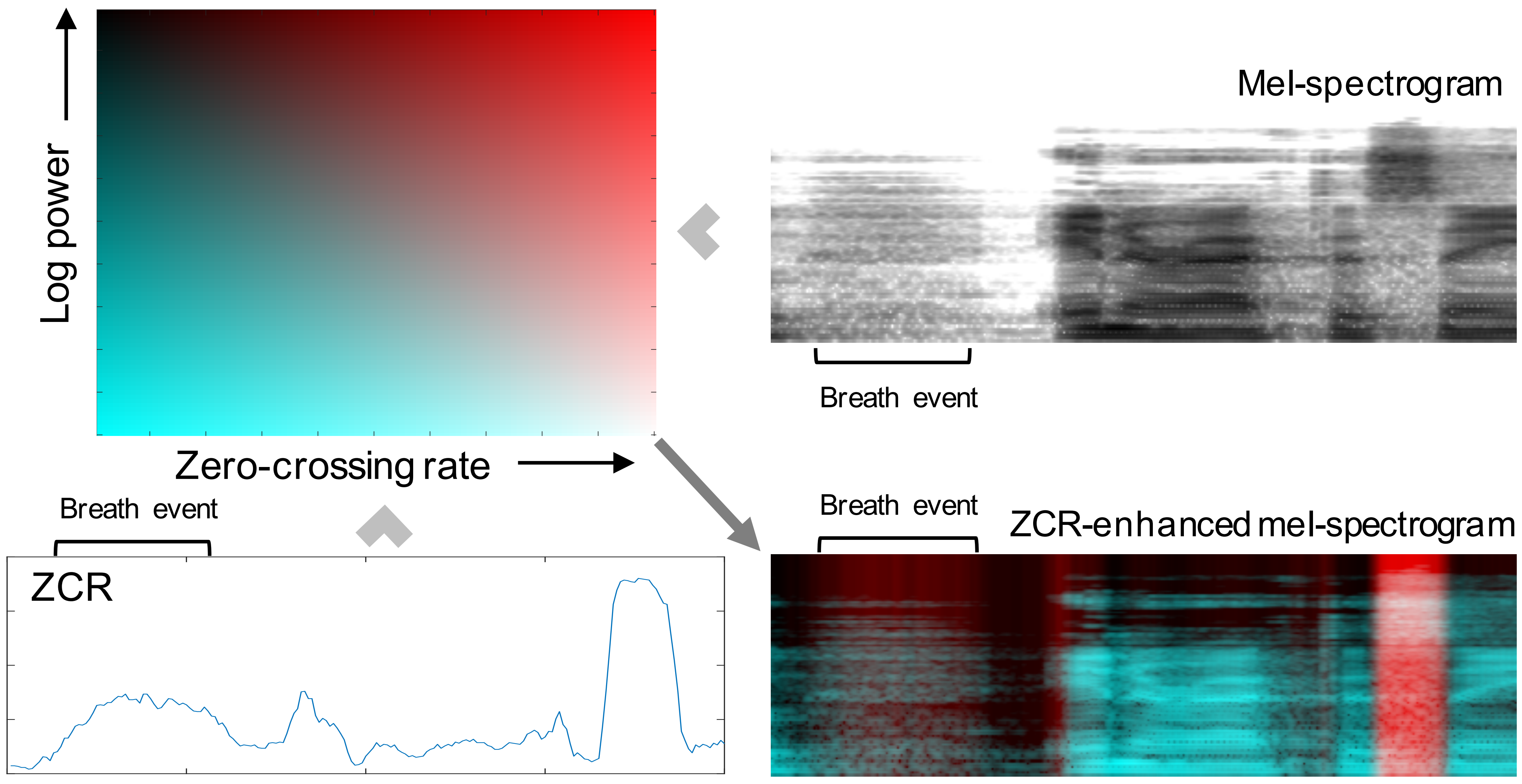
Data: public domain conversational podcast, 2 speakers

Method: semi-supervised approach with CNN-LSTM detecting breaths and overlapping speech on ZCR enhanced spectrograms



Architecture of the classifier

Why CNN-LSTM on spectrograms?
Long context sensitivity. Good performance on other paralinguistic tasks.



ZCR information makes breaths and fricatives more visually distinguishable

Why spontaneous speech data?
More appropriate for conversational settings.

Why found data?
Transcribed conversational speech databases are rare, but dialogue is common in found audio.
In large datasets we can pick and choose the best bits.

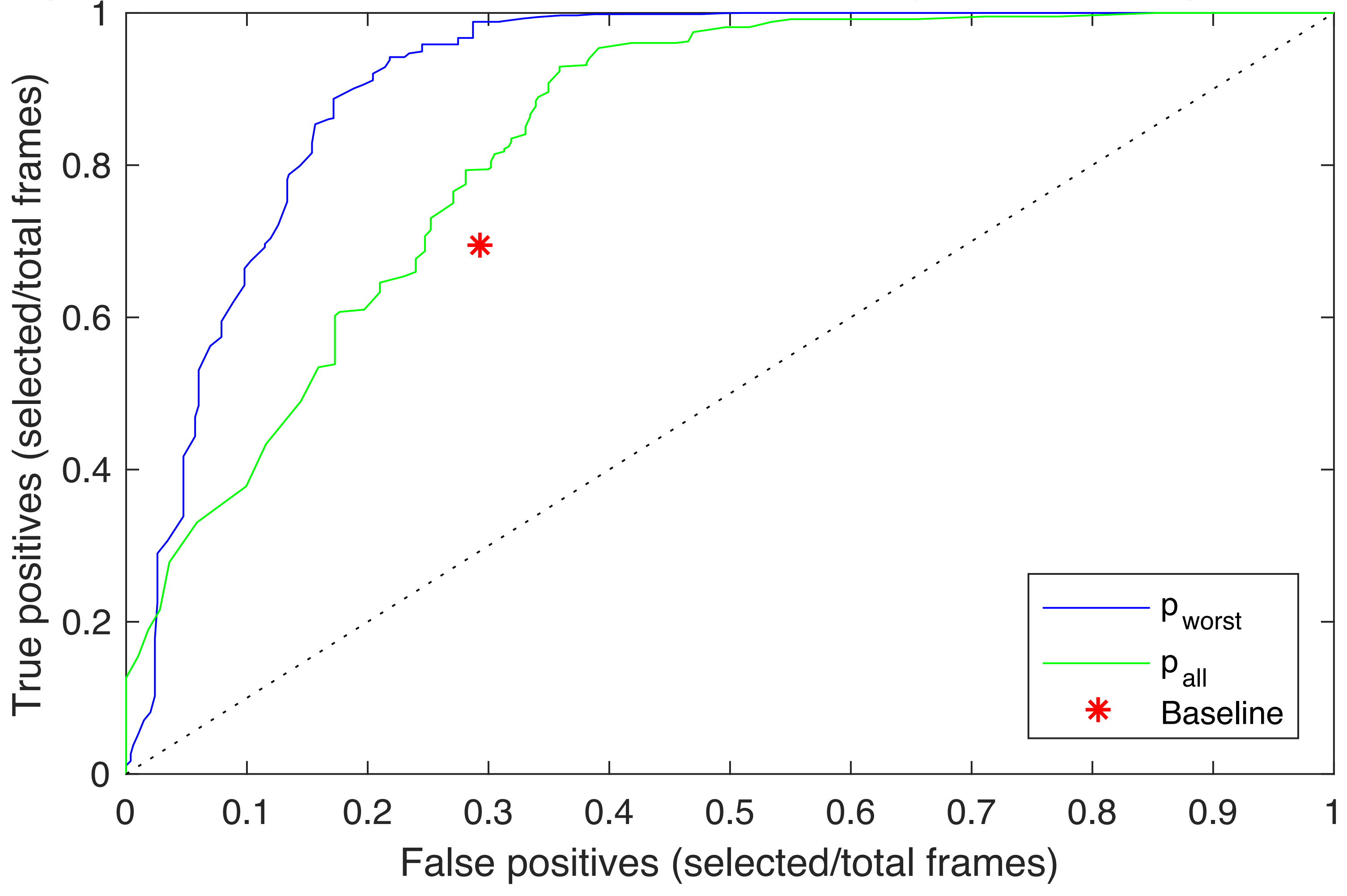
Why breaths?
Spontaneous speech does not neatly divide in sentences. Breath plays an important role in speech planning.

Two possible segment selection criteria

$$p_{\mathrm{worst}}(\mathrm{seg}) = \min_{t \in \mathrm{seg}} p_t$$

$$p_{\mathrm{all}}(\mathrm{seg}) = \exp\left(\sum_{t \in \mathrm{seg}} \ln p_t\right)$$

p_t is the estimated probability that frame t is acceptable (silence or breath or speech from target speaker)



ROC curves for the two segment-selection criteria and the baseline. p_{worst} was chosen as the proposed method for discarding bad segments

Input feature set	All classes	Target speaker breaths	
	Accuracy	Precision	Recall
Monochrome	67.5%	90.5%	81.7%
Viridis	69.9%	82.8%	93.9%
Mono. + ZCR	77.6%	96.3%	95.1%

Classifier performance with different input features

Issue	Baseline	Proposed	*p*-value
None (problem-free)	70	217	$<10^{-44}$
No breath at the beginning	111	4	$<10^{-30}$
Overlap with backchannel	37	17	$4.1 \cdot 10^{-3}$
Contains other speaker	26	7	$6.4 \cdot 10^{-4}$
Noise	6	5	0.84

Baseline vs. proposed on a sample of 250 test-set segments

Conclusions & future work

- ✓ Proposed method outperforms baseline selection method that treats breaths as silences
- ✓ Adding ZCR to the spectrogram improves breath detection
- ✓ Next step: conversational TTS

This research was supported by the Swedish Research Council Project Incremental Text-To-Speech Conversion VR (2013-4935) and by the Swedish Foundation for Strategic Research project EACare (RIT15-0107).
The authors would like to thank the creators of the ThinkComputers podcast for making the recordings available in the public domain.