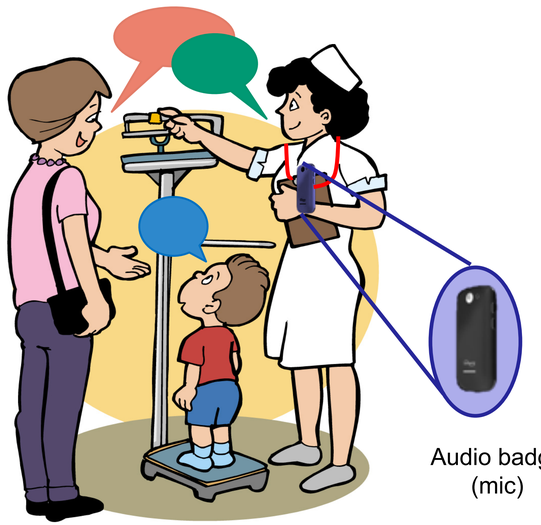
SPEAKER AGNOSTIC FOREGROUND SPEECH DETECTION FROM AUDIO RECORDINGS IN WORKPLACE SETTINGS FROM WEARABLE RECORDERS

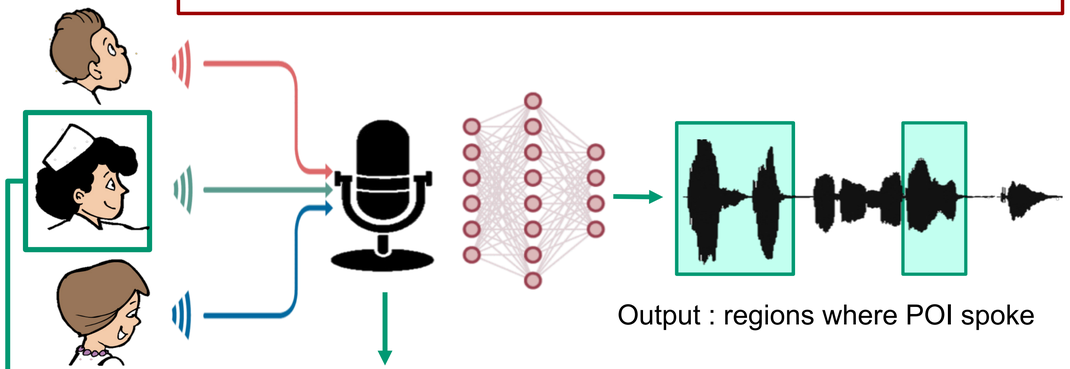
Amrutha Nadarajan *(nadaraja@usc.edu)*, Krishna Somandepalli, Shrikanth Narayanan
Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles

Experimental Setup

- Speech foreground (FG) detector designed for a wearable audio recording
- Data collected using "privacy-preserving" audio-badges [1]
- Designed for a sensitive environment
  - no raw audio, limited audio features
  - no a priori information on speaker characteristics



Audio badge (mic)



Foreground Definition

- Foreground Speech: Speech from person of interest (POI) here POI is the person wearing audio badge
- *vs* Speech Activity Detector : Any speech (including cross talk)

Training/fine-tuning

- Trained on class balanced ICSI corpus
- Validate/test/fine tune on in-house dataset (unseen speaker k-fold)
- Features: 14 MFCCs (+first, second deltas), pitch, intensity, loudness, voicing probability, RMS energy, zero crossing rate

Performance on public/in-house dataset

Table : Performance evaluation of different models: Test accuracy (%), Precision, Recall, EER(Equal error rate), F1(F1 score)

	ICSI	SMC			
Model	Test acc. (%)	Precision	Recall	EER	F1
FC-DNN	75.1	87.0	3.6	48.5	11
VGG *slim* [2]	87.1	24.6	94.5	50.6	57
VGG *slimmer*	**90.4**	46.0	85.1	**27.0**	**78**
fine-tuning results					
VGG *slimmer*	-	81.2	76.9	**18.6**	**84**

Use case for foreground activity

Do speaking estimates explain positive and negative affect?

Linear Mixed Effects model with positive/negative affect as outcome

- Null model :
  - subject as a fixed effect
  - controlling for gender
- Alternate model : Foreground Activity (FGA) as an additional variable
  - For positive affect: LME with FGA performed better than the null model ($\chi^2 \approx 7.5$, $p < 0.05$)
  - For negative affect: LME with FGA did not perform better than the null model ($\chi^2 \approx 1.4$, $p > 0.05$)

Datasets

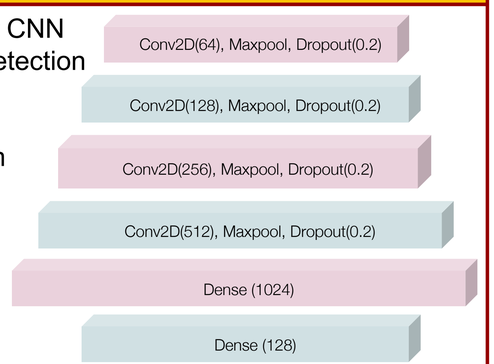
- ICSI - Public audio dataset [3] a generic, multi-party meetings based corpus
- SMC – In house data collected using [1]
- TILES (IARPA-MOSAIC [4]):
  - multimodal sensory data
  - to study overall health, personality, affect
  - clinical population at the USC Keck Hospital
  - self reports on positive, negative affect, stress, anxiety
  - Longitudinal study (10 weeks), N ~ 200

Network Architecture

- *VGG-slim* [2], state of the art CNN architecture for audio event detection
- Modified *VGG-slim* arch: "reduced" version of VGG slim

~50% reduction in number of parameters wrt *VGG-slim*

Fig : ***VGG-slimmer***



Summary

- A foreground speech detector with no a priori knowledge of speaker characteristics was designed using a limited set of audio features
- One use case of speaking activity estimates derived from foreground activity elaborated

REFERENCES

[1] Tiantian Feng, Amrutha Nadarajan, Colin Vaz, Brandon Booth, and Shrikanth Narayanan, "Tiles audio recorder: an unobtrusive wearable solution to track audio activity," in Proceedings of the 4th ACM Workshop on Wearable Systems and Applications. ACM, 2018, pp. 33–38

[2] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in ICASSP, 2017, pp. 131–135

[3 Adam Janin et al., "The icsi meeting corpus," In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on IEEE, 2003

[4] https://www.iarpa.gov/index.php/research-programs/mosaic

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No 2017 – 17042800005.