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

Amrutha Nadarajan, Krishna Somandepalli, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA
email: nadaraja,somandep@usc.edu, shri@ee.usc.edu

ABSTRACT

Audio-signal acquisition as part of wearable sensing adds an important dimension for applications such as understanding human behaviors. As part of a large study on work place behaviours, we collected audio data from individual hospital staff using custom wearable recorders. The audio features collected were limited to preserve privacy of the interactions in the hospital. A first step towards audio processing is to identify the foreground speech of the person wearing the audio badge. This task is challenging because of the multi-party nature of possible ambulatory interactions, lack of access to speaker information and varying channel and ambient conditions. In this paper, we present a speaker-agnostic approach to foreground detection. We propose a convolutional neural network model to predict foreground regions using a limited set of audio features. We show that these models generalize across the proxy corpora we collected in house to approximately match the deployment environment. The proxy corpora contained full audio and was used as a test-bed to analyze our models in greater detail. We also evaluated the models in the workplace setting to measure speech activity. Our experimental results show promising direction for analyzing workplace behaviors with privacy protected sensing.

Index Terms— speaking patterns, foreground detection, Speech Activity Detector, wearable sensing, audio

1. INTRODUCTION

Advances in wearable sensing have changed the way we think about audio signal acquisition. The availability of portable miniaturized, self powered systems [1, 2, 3] has enabled acquiring speech data in naturalistic settings. It offers ecologically-valid ways of studying social communication and interaction in rich real-world contexts. As we move towards wearable means to record audio data in unknown environments, we have potentially multi-party interactions with multiple, often unknown ambient sources. This move, away from clean and controlled recording environments presents interesting audio processing challenges.

Privacy concerns also arise with enabling sensing in naturalistic settings, especially when collecting data in sensitive environments such as one's home and workplace. Digital recorders approach this issue by allowing users to control the recorded content. The users can control what to record by deleting the recordings retroactively [1] or collect select features (e.g., energy to determine duration of voice activity). This eliminates the need to save raw audio recordings [3, 2]. In the former case, the cognitive load of the participant increases, but we have access to raw audio for further analysis. In the latter case, we can perform (limited) online feature extraction to extract pre-designed features to perform audio analysis. As a consequence of not having access to raw audio, we cannot obtain human annotations to develop downstream machine learning tasks.

As we move into naturalistic settings and privacy-protected sensing of the participant, we need to rethink the design of fundamental audio signal processing modules. Separating speech from different ambient sources comes naturally to humans. But, this becomes an arduous task from a computational perspective, especially when the number of sources is unknown. This is compounded due to the lack of prior information on the environment characteristics.

In this paper, we focus on the problem of foreground speech detection in data collected from wearable audio recorders. The specific experimental use case is based on recordings (from a first person view) using "audio badges" worn by clinical staff in a large hospital setting as a part of a larger study focused on workplace behavior and performance [3].

1.1. Foreground - definition

A close-talk audio recording, typically in a multiparty conversation can be categorized into four types of segments based on content [4]: 1) speech from the person wearing the audio recorder, 2) cross talk or speech from other people nearby, 3) ambient sounds and noise, and 4) silence. These segments can co-occur in different combinations (except silence). Following [4], we define foreground speech (*foreground*) as speech regions captured by the recording device that belong to the person wearing the recording device. Speech segments other than the foreground belong to cross talk. Non-speech regions either belong to silence or noise.

Depending on the application, *foreground* has been defined in several ways:

- def.1 *any* speech [5]
- def.2 speech from *a* person of interest, usually the one wearing the audio recorder [4]
- def.3 anything that is not ambient audio (which we term background) [6]

According to def.1, a foreground (FG) detector would just be a speech activity detector (SAD), def.3, would be a general auditory scene descriptor – an indirect way to detect foreground regions. As will be described in Section 2, we are ultimately interested in using features from the wearer's speech to predict psychological/affective measures of individuals in workplace. Thus, def.2 would be the most appropriate way to define FG for our experimental setup.

A canonical approach to detect FG would be to use a speaker verification system [7] with a SAD, assuming that the voice characteristics of the wearer of device are known apriori. But, often in large scale behavioral studies, we are not always privy to this information. Hence, in this paper we present a speaker-agnostic approach to FG detection. In Section 5 we show how the deep learning network designed is able to perform reliably well without a priori knowledge of the speaker i.e, speaker agnostic.

1.2. Background on foreground detection

Identifying FG speech is relatively an easier task when we have information on the speaker identity, clean recording conditions or known multi-speaker setting because speaker specific models can be effectively designed. Most prior work has examined audio data where two or more of these conditions have been satisfied. In this section, we primarily focus on prior work with def. 2 of FG.

In known multi-channel conditions, HMMs and GMMs have been used (e.g., [8, 4]) for detecting FG. Short-time correlation of all channel priors were used to reliably extract FG regions in [9]. Deep learning methods work exceptionally well on supervised FG detection when we have raw audio information. The survey paper by Wang et. al., [10] showed different approaches to address this problem, such as estimating an ideal binary mask, target binary mask to localize the speech regions in a spectrogram. Specifically, CNNs have also been shown to work well with musical background [11] to extract FG. [12] used ensemble of DNN models for monoaural source separation. A common theme in all these works is that they use some form of speaker or channel information for modelling FG.

In contrast, there have been fewer approaches that detect FG in a speaker-agnostic manner. Speaker-adapted eigen faces was used in [13, 14] for FG detection without prior information on speaker wearing the mic. However, these methods still need access to the raw audio. Our work examines FG detection when it is not feasible to collect raw audio data.

In the domain of wearable sensing, FG detection largely remains unexplored. A robust VAD using dictionary learning approach was designed for audio collected using smartphones in [15]. A speech detection and localization algorithm for smart headphones was designed in [16]. As described in Section 1, the problem we address however is compounded by unknown multi-party interactions in a wearable recording setup without access to raw audio or apriori speaker information.

2. DATASETS

2.1. Deployment dataset

As described in Section 1, the foreground (FG) Detector was designed to be deployed in audio recorded in a highly sensitive hospital environment [3]. In this section, we briefly describe the study where the audio was collected. As a part of MOSAIC[1] program, in early 2018, we collected a preliminary set of multimodal sensory data for "TILES: Tracking Individual Performance with Sensors" to study how workplace stressors affect the overall health, personality, workplace behavior and affect of hospital employees (belonging to the clinical population at the USC Keck Hospital, Los Angeles, CA). Data was collected for a period of ten weeks from 213 nurses/hospital workers using specially designed audio badges called TAR [3] that employees wore during work-shift hours. From this sample, we chose a subset of N=50 (32 Female, 18 Male, 30 day shift, 20 night shift nurses) for our analysis. The average number of audio files (each of length 20 seconds) per subject was 6138.3 ± 3300.52. For the 50 participants, about 1220 hours of data was collected. The participants answered self reports daily or once in two days pertaining to psychological measures like affect, personality, workplace behavior and health - some of which included daily surveys on affect, stress and anxiety.

Because the experimental environment was a hospital, our study and practices complied with HIPAA regulations [17]. As such, we had no access to raw audio or ground-truth annotations (e.g., FG labels). Hence, we adopted a three step approach to design a robust FG detector, as described below. 1) We train our models on ICSI corpus (a generic, multi-party meetings based corpus) 2) We select the top-performing models in step 1, to test/validate and fine tune on an in-house dataset collected using TAR [3] 3) We deploy the best model on TILES data to get frame level predictions, use the speaking activity estimates in Section 3.3 and assess the validity of these estimates in predicting workplace behaviours.

2.2. Datasets for training/validation

2.2.1. Public audio dataset: ICSI meeting corpus

We chose a publicly released dataset, the ICSI meeting corpus [18] (ICSI), to train an FG detector (See Section 3.1 for model details). We chose ICSI because it has natural conversations with multi-party interactions (close talk and far field). ICSI also provides densely annotated audio and transcriptions from which FG regions can be reliably obtained. [18].

2.2.2. In-house dataset: SAIL Meeting Corpus (SMC)

We also collected in-house data which we call SAIL Meeting Corpus (SMC) where we recorded multi-party conversation during weekly research meetings. The participants in the research meetings wore the same audio badges as that of the participants in the TILES study. We also ensured that the positioning of the recording device on the individual was similar to that of our deployment data. In addition, we also recorded raw audio which we annotated in-house for four labels: FG (def.2 as defined in Section 1.1), cross talk, noise and silence. The data was labelled using Sonic Visualizer [2] at a 100ms time precision. This labelled dataset enabled us to fine-tune and validate models trained on ICSI data for deployment.

Next, we defined two FG activity estimates as described in Section 1.1. Because we had ground truth labels on SMC, we can evaluate the robustness of these estimates. We hypothesized that these activity estimates may be related to some of the ground truth available for the deployment dataset.

Table 1. Datasets and their train/validation/test split information

Dataset		%	# speakers	# sessions
ICSI	train	60	26	247
	test	20	12	100
	val	20	11	81
SMC		100	9	12

3. METHODS

As described before, deep learning methods have been effective in a variety of audio processing tasks including automatic speech recognition and audio event detection (AED); for example, log-Mel features with CNN models have been shown to effectively classify over 500 audio event classes on Audioset data [19]. We can consider modeling log-Mel features with CNN analogous to learning custom filters over hand-designed filter banks like MFCCs to maximize a task-specific objective. The network architecture used in [20], was referred to as *VGG-slim* because it is a simplified version of the image classification architecture, VGG [21]. Based on our initial experiments with VGG-slim for FG task, we further modified the architecture by reducing the number of convolutional layers and tuned the number of nodes in the fully connected layers. This resulted in a simpler version of the model, which we refer to as *VGG-slimmer*.

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

[2] https://www.sonicvisualiser.org/

Note that, as part of TAR feature extraction [3], we only had access to MFCCs which we used as input features instead of log-Mel features.

3.1. *VGG-slimmer* Architecture

For brevity, we describe the CNN architecture using a simplified notation: Let **conv_block[K]** denote sequence of a 1) convolutional layer with *K* number of filters (kernel size=3x3, strides=1x1), 2) Maxpooling (kernel size=2x2, strides=2x2), and 3) Dropout layer; **FC[n]** denotes a fully connected layer of *n* nodes followed by a Dropout layer. We set the dropout rate at 0.2 (increasing it to 0.5 did not improve our performance).

VGG slimmer :: **INP** $\rightarrow$ **conv_block[64]** $\rightarrow$ **conv_block[128]** $\rightarrow$ **conv_block[256]** $\rightarrow$ **conv_block[512]** $\rightarrow$ **FC[1024]** $\rightarrow$ **FC[128]** $\rightarrow$ **Sigmoid[1]**

where **INP** = Input of dimension 51x*M*, where *M* is the time context window for the input with 51 features. The context was tuned over different choices for *M* ranging from 0.2s (*M*=19) to 1s (*M*=95) with 0.2s increments. To satisfy dimensionality constraints, we use a kernel of 1x3 for the final convolutional layer. The output of the final **conv_block** was vectorized before input to the FC layer. The last layer consists of one node with sigmoid activation.

3.2. Features

The audio badges we used for study deployment used OpenSmile [22] to extract audio features referred to as low-level descriptors (LLDs). These contain canonical audio features such as pitch, intensity, spectral descriptors and auditory filter bank coefficients (like MFCC, PLP). A sampling frequency of 16kHz and a frame-length of 60ms with 50ms overlap were used to extract features. These parameters were pre-set in the configuration of the audio badges at the time of deployment for the study [3]. Hence, we used this configuration across all our experiments.

The input to the network was a 51 dimensional feature vector (per frame) where 42 of those are the 14 MFCCs and their first and second differences. The 9 other features were fundamental frequency, intensity, loudness and their deltas, voicing probability, RMS energy and zero crossing rate. The input features were normalized by substracting mean and dividing by standard deviation. We applied a median filter for post processing on the predictions at frame-level. The kernel size of the median filter was chosen to maximize F1 score on the validation set.

3.3. Foreground speech estimates

We compute two speaking activity estimates from the FG speech regions predicted by our model: 1) foreground activation (**FGA**), the percentage of recording time that foreground speech is present and 2) foreground activation frequency (**FGAF**): the frequency of foreground speech activation in a recording.

Formally, let $\mathbf{S} \in \mathbb{R}^{d \times N}$ be the matrix with d-dimensional features as columns for N frames. Let $p[n]$ be the frame level posterior from the network with $p[n] \in [0, 1]\ \forall n$. Then,

$$\rho[n] = \mathbb{1}\big(\mathcal{M}_k(p[n]) > \epsilon\big) \tag{1}$$

where $\mathcal{M}_k$ is the median filter of length k, $\mathbb{1}$ is the indicator function: $\mathbb{1}(z > \epsilon) = 1$ if $z > \epsilon, 0$ else. $\rho(n)$ having a value of 0 denotes BG (background) and 1 denotes FG. In practice, ϵ of 0.5 is used. Now, we define FGA and FGAF as follows:

$$\text{FGA}(\mathbf{S}) = \frac{1}{N} \sum_{k=0}^{N-1} \rho[k] \tag{2}$$

$$\text{FGAF}(\mathbf{S}) = \frac{1}{l} \sum_{k=1}^{N-1} \mathbb{1}((\rho[k] - \rho[k-1]) > 0) \tag{3}$$

where $\mathbb{1}$ is the indicator function, $l = N/t$ is the segment length and t is the number of frames in each segment. Due to the preset configuration of TAR, t = 1900 corresponds to 20s.

Note, that t can be chosen arbitrarily to control the segment length, such that $l \leq N$. Given an audio segment, FGA can be viewed as a proxy for speaking time of the person of interest, and FGAF for the number of times a person starts speaking.

4. EXPERIMENTS

We conducted several experiments to select an architecture across the two datasets described in Section 2. We conducted experiments with the original VGG *slim* architecture, a smaller (fewer number of nodes per layer) version of VGG *slim* and a smaller version of VGG *slim* without double convolutions. We also varied the number of dense layers (decreasing number of nodes).

We also experimented with varying input using just 14 MFCC features but the task performance significantly improved with the 51-dimensional feature set described in Section 3.2. We tested different normalization methods: min-max, max-absolute scaling and z-standardization of which z-standardization worked best. For the input, we also vary context: ranging from 0.2 to 1 second. We found that increasing context improves performance and generalizability across datasets. The final model we chose had a context of 1 second.

The models were all trained in Keras with TensorFlow backend. We used RMSprop for optimizing the network with an initial learning rate of 1e-4, ρ = 0.9, $\epsilon = 0$ and *decay* = 0.

4.1. Baseline Models and adapting for SMC

We compare performance of the proposed architecture with two baselines. We use a fully connected (FC) DNN consisting of 5 layers with decreasing number of nodes as one of our baselines. Each FC layer was followed by a dropout layer (0.2 dropout). Because our model was a modified version of VGG *slim*, we use this network as the CNN baseline. We used class balanced ICSI data for training as we had adequate number of samples (N=784000) for each class.

Because we cannot collect labels for TILES audio data, we used the SMC corpus (See Section 2.2.2) as a test-bed to evaluate and understand our models to a greater detail. For fine-tuning, we used a k-fold validation on SMC because we only had 9 speakers, and we wanted to test our models for speaker variability. We re-partition the data into 4 splits to fine-tune the models trained on close talk speech from ICSI, each split with data from 3 speakers for training and 6 speakers for validation. We ensure the splits have no common speakers. We used a smaller proportion of the data for fine-tuning to ensure generalizability. We chose the model that performed consistently (least variance of F1-score) across the different splits. We used an early stopping criteria on the validation loss (stop training if the loss did not improve by 1e-4 for at least five consecutive epochs).

4.2. Statistical testing for FGA and FGAF

We then compared the estimates from the predicted FG mask and the GT measures with a Welch two sample t-test to test for statistical differences. The ground truth FGA and FGAF were computed on SMC dataset. For all tests, we chose $\alpha = 0.05$ to decide statistical significance. In the context of our deployment environment, different subjects may start wearing the TAR devices at different times during the day. To simulate this, we recreated the recording set up of [3] where the recorder is active for 20 seconds and inactive for

Table 2. Performance evaluation of different models; Precision (P), Recall(R), EER(Equal error rate), F1(F1 score)

	ICSI accuracy (%)			SMC accuracy (%)			
Model	train	val	test	P	R	EER	F1
FC-DNN	88.4	78.5	75.1	87.0	3.6	48.5	11
VGG *slim*	94.1	89.3	87.1	24.6	94.5	50.6	57
VGG *slimmer*	**93.3**	**90.1**	**90.4**	46.0	85.1	**27.0**	**78**
fine-tuning results							
VGG *slimmer*	-	-	-	81.2	76.9	**18.6**	**84**

40 seconds. Since we do not know when the participant chooses to turn on the recorder, we simulated this with random offsets to begin turning on the recorder. The random offset was chosen uniformly over the range (0,60) seconds in 10s intervals. We repeat this 3700 times for each recording file in the SMC corpus. For each simulation, we estimated the FGA and FGAF measures from the predicted FG labels, as well the ground-truth measures.

We refer to this simulation as *random-offset-simulation*. For each of the 3700 simulations, we computed t-statistic from the Welch-corrected t-test and examined the p-values corresponding to the 5th and 95th percentile of the distribution. This gives us the confidence intervals (CI) of differences between the predicted and true activity measures for the random-offset-simulation.

5. RESULTS

The performance of different networks trained on ICSI and tested on SMC corpus is shown in Table 2. The VGG-*slimmer* model outperforms all the other models we tested suggesting that the modifications we made to the VGG-*slim* architecture are effective for the FG detection task. It is important to note that the design choices we made to VGG-*slim* were solely guided by the ICSI validation accuracy, and not the SMC data. This indicates that our models generalize across-datasets.

Notice that although the recall and EER are significantly lower for a simple FC-DNN model, than the VGG-*slimmer*, the precision is high (row 1 vs 3 in Table 5). In other words, FC-DNN can classify very few frames as FG (low recall) accurately (high precision). However, fine-tuning the VGG-*slimmer* models on SMC data improves the precision from 46% to 81.2% with a minimal loss in recall suggesting that fine-tuning is an important step in adapting models across datasets. This is consistent with prior literature which recommends fine-tuning for models to generalize better in the test domain [23].

In order to assess the stability of the proposed system, we performed 4-fold validation as described in Section 4, and fine-tuned to adapt the model. To adapt the models for SMC data, we initialized the weights of VGG-slimmer from the ICSI model, and retrained with SMC data. We then experimented with freezing the weights of the conv. layers with the ICSI model and trained only the FC layers with SMC data. The models trained end-to-end perform consistently better across different splits with an average F-measure of 77.1 and a standard deviation of 0.1. Of the 4 different models, we picked the model that showed the best performance on train and test splits.

5.1. Reliability of speaking activity estimates

Overall, we found no statistical differences between the GT and estimates of FGA and FGAF (p-value = 0.2299 and 0.1461) computed across the entire length of the audio file.

Results from the random-offset simulation across 3700 simulations: For FGA, the average t-statistic for $\mu_{true} - \mu_{pred}, DF = 7$ was $t = -1.29 \pm 0.18$ with $(0.05, 0.95)$ CI were $(-1.62, -1.00)$ with corresponding p-values of 0.1498 and 0.3492 respectively. This indicates that our predicted model does not significantly overestimate FGA. For FGAF: $t = 1.94 \pm 0.27$ with $(0.05, 0.95)$ CI of $(2.39, 1.51)$ with p-values of 0.1737 and 0.0478 respectively. At an $\alpha = 0.05$ for statistical significance, we note that FGAF is slightly underestimated by our FG models. These results suggest that the FGA and FGAF measures can be reliably estimated by our FG models.

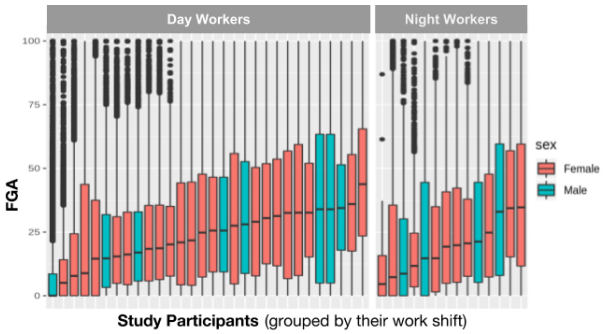

Fig. 1. Distribution of foreground activity (FGA) for each subject with respect to gender and work shift (day/night shift worker)

5.2. FGA and FGAF estimates to predict MOSAIC constructs

We performed preliminary analysis of the speaking activity estimates on the TILES (deployment) dataset. Figure 1 shows the distribution of FGA for different work shifts for male and female subjects. Welch's t-test on FGA by gender showed that Female subjects have significantly higher foreground activity ($t = -6.9, p << 0.01$). We also tested if the FGA estimates explain positive and negative affect from self-reports (two of the MOSAIC constructs) using linear mixed effects (LME) models. We chose LME because we have repeated measurements of subjects' self-report evaluations. A null model was built with positive/negative affect as outcome and subject as a fixed effect (repeating measurement), and controlling for gender. The alternate model included FGA as an additional variable. We used a chi-squared test to test if the LME model with FGA was performing significantly better than the null model. For positive affect: LME with FGA performed better than the null model ($\chi^2 \approx 7.5, p = 0.006$). For negative affect, the LME model with FGA did not perform significantly better than the null model ($\chi^2 \approx 1.4, p = 0.237$). These observations were consistent with LME including FGAF estimates as well. These initial results suggest that the speech activity estimates can explain some of the variance in predicting positive affect measure. A detailed analysis of FGA/FGAF estimates to predict other behavioral measures is a focus of our future work.

6. ACKNOWLEDGEMENT

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. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

7. REFERENCES

[1] Matthias R. Mehl, James W. Pennebaker, D. Michael Crow, James Dabbs, and John H. Price, "The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations," *Behavior Research Methods, Instruments, Computers*, vol. 33, no. 4, pp. 517–523, Nov 2001.

[2] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell, "Soundsense: Scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2009, MobiSys '09, pp. 165–178, ACM.

[3] 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.

[4] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 84–91, 2005.

[5] Neville Ryant, Mark Liberman, and Jiahong Yuan, "Speech activity detection on youtube using deep neural networks.," in *INTERSPEECH*, 2013, pp. 728–731.

[6] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo, "A semi-supervised learning approach to online audio background detection," 2009.

[7] Pejman Mowlaee, "New stategies for single-channel speech separation," *Institute for Electronic system, Aalborg University, Aalborg, Denmark Ph. D. thesis*, 2010.

[8] Thilo Pfau, Daniel PW Ellis, and Andreas Stolcke, "Multispeaker speech activity detection for the icsi meeting recorder," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 107–110.

[9] Kornel Laskowski, Qin Jin, and Tanja Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Eighth International Conference on Spoken Language Processing*, 2004.

[10] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[11] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 429–436.

[12] Xiao-Lei Zhang and DeLiang Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 967–977, 2016.

[13] Ron J Weiss and Daniel PW Ellis, "A variational em algorithm for learning eigenvoice parameters in mixed signals," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 113–116.

[14] Ron J Weiss and Daniel PW Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, vol. 24, no. 1, pp. 16–29, 2010.

[15] Sebastian Feese and Gerhard Tröster, "Robust voice activity detection for social sensing," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 2013, pp. 931–938.

[16] Sumit Basu, Brian Clarkson, and Alex Pentland, "Smart headphones: enhancing auditory awareness through robust speech detection and source localization," in *icassp*. IEEE, 2001, pp. 3361–3364.

[17] Denise L. Anthony, Ajit Appari, and M. Eric Johnson, "Institutionalizing hipaa compliance: Organizations and competing logics in u.s. health care," *Journal of Health and Social Behavior*, vol. 55, no. 1, pp. 108–124, 2014.

[18] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The icsi meeting corpus," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 1, pp. I–I.

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[20] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[23] Nicholas Becherer, John Pecarina, Scott Nykl, and Kenneth Hopkinson, "Improving optimization of convolutional neural networks through parameter fine-tuning," *Neural Computing and Applications*, pp. 1–11, 2017.