RICE UNIVERSITY

By

Wenwan Chen

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE

Ashutosh Sabharwal (Dec 1, 2020 19:08 PST)

Ashutosh Sabharwal

Ankit B Patel (Dec 1, 2020 21:25 CST)

Ankit Patel

Santiago Segarra (Dec 1, 2020 22:01 CST)

Santiago Segarra

HOUSTON, TEXAS

December 2020

ABSTRACT

AmbianceCount: An Objective Social Ambiance Measure from Unconstrained Day-long Audio Recordings

by

Wenwan Chen

Measuring social ambiance in unconstrained environments is of significant importance in mental health due to the association between sociability and psychological outcome. However, it has been challenging to quantify social ambiance since existing objective methods fail to capture the transient ambiance patterns in unconstrained environments. In this thesis, I present AmbianceCount, an automatic and objective method that extracts social ambiance from unconstrained audio recordings by estimating the number of concurrent speakers. AmbianceCount consists of a supervised deep neural network (DNN) embedding extractor to differentiate speech mixtures, and a scoring system for estimation and improving generalization. The performance of AmbianceCount is compared with baseline and evaluated on several synthesized datasets. Lastly, I utilize AmbianceCount to evaluate data from a sociability pilot, with audio data from depression and psychosis patients as well as age-matched healthy controls. Our analysis shows that extracted social ambiance patterns are significantly different across three groups. Besides, it is observed that captured social ambiance patterns are associated with psychometric and personality scores, which is consistent with clinical diagnosis.

Acknowledgements

First of all, I want to thank my advisor Dr. Ashutosh Sabharwal for his guidance, encourage and consistent support. His thorough knowledge and patience helped me through all stages of work both technically as well as mentally. He would always encourage me to take independent decisions about the research and at the same time provide insightful feedback. He is the nicest advisor anyone could have.

Besides my advisor, I would like to thank Dr. Ankit Patel and Dr. Santiago Segarra for valuable suggestions for my project and taking time to review my thesis. I also want thank my collaborators, Dr. Nidal Moudakkam. Her passion in the project and constructive feedback are one of my biggest motivations.

Further, I want to thank all my amazing lab-mates and friends at Rice. I really appreciate your support and company.

Finally, I want to express my appreciation to my parents and Mr. Yan He. You are always there for me and bring so much joy to my life.

Contents

Illustrations

Tables

Chapter 1

Introduction

Ambiance, defined as "the character and atmosphere of a place" [1], is a construct that describes our perception of an environment. Social ambiance, in this context, is our perception of the social environment, characterized by social interactions happening around us. Social ambiance is believed to influence mental health and well-being of individuals [2, 3].

Despite its importance, the measurement of social ambiance has been a challenging for two reasons. First, to my best knowledge, there are no self-report measures that can capture social ambiance throughout the day, especially since social ambiance is constantly changing and it is not practical to ask participants report often in a day. Second, existing objective methods fail to capture transient social ambiance patterns since they only provide coarse-scale and aggregate information e.g., ambient volume [4], chatter and noise levels [5]. Therefore, an objective and fine-scale measure to quantifying social ambiance is required.

In this thesis, I propose AmbianceCount, which measures social ambiance by estimating the number of concurrent speakers from unconstrained audio recordings. Different from speaker diarization [6] that counts speakers after identification, AmbianceCount directly estimates the number of concurrent speakers *without* identification. This also mimics the process of how humans perceive an overlapped speech. Note that I use audio recordings as data source because they capture an audio environment continuously and participants are generally more comfortable with audio

recording compared to videos/images recording. Besides, conversations, as important signs for nearby social interactions, create a type of sound texture that represents the social ambiance of a place.

The AmbianceCount consists of a front-end embedding extractor to differentiate speech mixtures, followed by a back-end scoring system that output the estimation result and to improve generalization. The deep neural network (DNN) embedding extractor is based on x-vectors [7] architecture, which retains the relevant sequential information of input speech mixtures, and has great performance for similar tasks like age estimation [8] and spoken language recognition [9]. To explore the best architecture for the embedding extractor, several candidate neural networks are investigated and compared. Besides, to mimic humans perception of speech mixtures, I define the task to as a classification-regression that jointly optimize jointly minimizing cross-entropy and Mean Square Percentage Error (MSPE) loss [10]. Evaluation results show that AmbianceCount performs well in noisy and reverberant environments, and is able to generalizes well to unseen data.

To demonstrate the utility of AmbianceCount in mental health, I apply AmbianceCount to data from a Sociability pilot, an IRB-approved dataset that captured multi-modal social interactions from three cohorts: depression patients, psychosis patients and age-matched control subjects. Results show that social ambiance patterns captured by our algorithm are significantly different across three groups, which is consistent with clinical observations. The social ambiance patterns are also associated with psychometric and personality scores.

The contributions of this thesis can be listed as follows:

1. I propose AmbianceCount, an objective method that measure social ambiance from unconstrained audio recordings by estimating the number of concurrent

speakers.

2. I evaluate the performance on different social scenarios and different synthesized datasets. Results show that AmbianceCount is robust against noise and reverberation, and is able to generalize well on unseen data.

3. I apply AmbianceCount to Sociability trial and observed that extracted social ambiance patterns are significantly different across three groups. Besides, it is observed that captured social ambiance patterns are associated with psychometric and personality scores.

Chapter 2

Related Work

2.1 Objective Ambiance Measurement

A majority of existing work extract acoustic or visual cues to infer ambiance. To improve local search experience, [5] investigates the recognition of physical ambiance categories (occupancy, human chatter, noise and music levels) using audio features collected in-situ by users. To predict human perceptions of outdoor places, [11, 12, 13] predict human perceptions of wealth, uniqueness and safety using geo-tagged images. Despite the influence of ambiance on mental health and well-being, [2, 3], to the best of our knowledge, there is no objective measure for social ambiance that is characterized by social interactions happening nearby.

Intuitively, real-world audio recordings can be good data sources for social ambiance study, since they capture nearby social interactions continuously and people are generally more comfortable with recording audio compared to recording videos/images. One possible avenue is to use speaker diarization [14], that can be used to reconstruct the social ambiance from audio recordings by detecting nearby speakers. However, current diarization systems work under an assumption that only one speaker is active at a time. Existing methods thus fail when applied to real-world scenarios where multiple overlapping events active at the same time.

The fact is, speaker overlaps, prevalent in most social scenarios, create a type of sound texture that represent social ambiance. Instead of discarding all the in-

formation in the overlaps, AmbianceCount aims to utilize the number of concurrent speakers estimated from a speech mixture as proxy for social ambiance.

2.2 Concurrent Speaker Count Estimation

It is challenging to estimate the number of sources in a speech mixture. An intuitive idea would be to use several microphones and to fully utilize the space information [15]. However, this is not applicable for social ambiance measurement, where long-term continuous data is collected from participants with unrestricted mobility. Consequently, counting from the single-channel has been considered in many studies [16, 17, 18, 19]. In [18], single-channel overlapping speech is exploited to estimate the indoor occupancy level. A novel entropy-based method is proposed for room occupancy in both small and large crowd configurations. The influence of room setting, speaker position are also investigated.

Efforts have been made to understand the underlying challenges to pave an operational strategy. In [16], the authors explore several neural network architectures to find the best strategy for speaker count estimation. They generalize the problem by fusing classification and regression and later prove that classification outperforms regression for all network architectures. Furthermore, factors that influence the performance are assessed, including reverberation, segment duration and volume differences between utterances. These preliminary results provided key insights into the making of a robust count estimation system. Work in [19] addresses the underlying model and establishes a working foundation for further analysis. The authors describe that with an increasing number of speakers overlapped, the speech mixture is less time varying in the long term, but more chaotic in the short term.

All the prior methods, however, define the task as a classification problem where

the relations between different possible values are lost. The network outputs the estimates by picking the most likely class, without considering the deviations of wrong estimates from ground truth. There are two observations in terms of how humans estimates the number of concurrent speakers: (i) most errors we make fall into adjacent classes, and (ii) humans are in general more discriminative when there are fewer sources in the mixture. Inspired by above observations, our proposed AmbianceCount aims to mimic how humans process the information by penalizing the network based on their percentage deviation from ground truth. Specifically, the ordering of different classes are kept by defining the task as a classification-regression problem. Besides, for the same absolute error, the algorithm is stricter when the ground truth number of speakers is smaller.

Chapter 3

Mixture Dataset Preparation

To develop an algorithm for the speaker count estimation and to ensure the training convergence, the first step is to tag the training and evaluation datasets correctly with the speaker count. Given that a realistic dataset of fully overlapped speakers is not available, I synthesized a dataset with speech mixtures.

The main idea is to create speech mixtures from monologue speech, then tune it with sound effect libraries to simulate ambient sounds and reverberation. Thousands of realistic variations can be created by adjusting speech audio properties such as amplitude and rate, as well mixing sounds with different sound effects libraries, *e.g.*, MUSAN [20] and RIR-NOISE [21].

Following sections describe our process of preparing the mixture dataset.

3.1 Mixture Creation

Since our model should be speaker independent, monologue speech corpus with a higher number of different speakers is preferred. Therefore, we utilize LibriSpeech corpus, a speech data set based on LibriVox's audio books. By combining three LibriSpeech subsets, clean-360, clean-100 and other-500, I get 960 hours of speech in utterances from 2338 speakers (1228 female speakers and 1210 male speakers), sampled at 16 kHz. Utterances were segmented when the silence intervals were longer than 0.3 seconds or coincided with sentence breaks in the reference text, so it is

assumed that a speaker is continuously active within each monologue utterance.

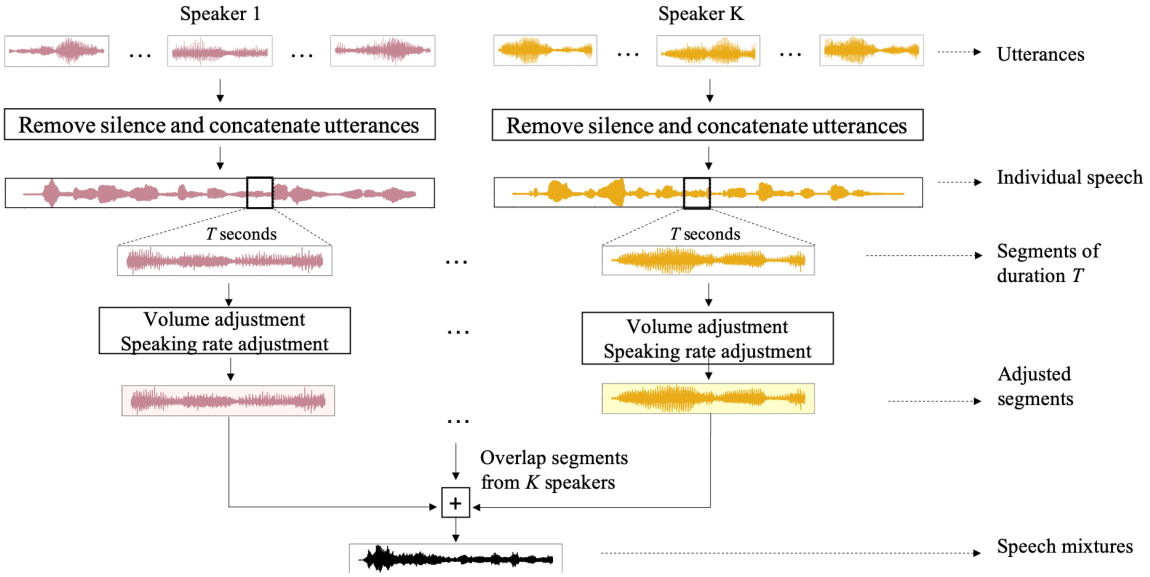


Figure 3.1 : An example of creating K-speakers mixtures.

Figure 3.1 describes, for example, how to create a K-speakers mixture from LibriSpeech utterances.

3.1.1 Utterances to recordings

For each speaker, a 15-30 min recording is generated by concatenating LibriSpeech utterances of that speaker. By doing so, I will later be able to select a random excerpt from each speaker and investigate the influence of duration T regardless of the various lengths of the original utterances. To this end, silence is removed at the beginning and the end of the utterances, concatenate them into a recording of that specific speaker.

3.1.2 Selecting segments

Next, an audio segment is randomly selected from each recording. It is necessary for developing a content-independent algorithm since in most speech corpus, people are required to read the same material. By randomly selecting segments from the concatenated recording, it is more likely that everyone in the mixture is reading different content.

3.1.3 Adjusting segment

The mean amplitude of selected segments is normalized, which is analogue to an ideal scenario where speakers are equally distanced from a single microphone and are talking in the same volume. Next, segments are randomly adjusted in volume and speed to simulate how people speak in real-world scenarios. Consider the influence of microphone locations as well as people speaking in different volumes and speeds, I randomly apply to the mixture a volume factor between -3 dB to +3 dB and a speed factor ranging from -0.9 dB to +0.8 dB.

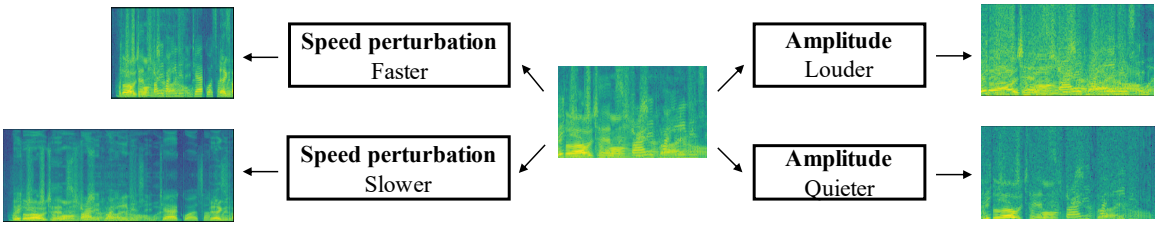


Figure 3.2 : Audio Properties adjustment

3.1.4 Clean speech mixture

Finally, adjusted segments from K speakers are trimmed to T seconds and overlapped with each other to generate a speech mixture, labelled with speaker count of two. At

this point, we have what we call a clean speech mixture, in which at any given time, K speakers are simultaneously active, without noise and reverberation.

The same mixing approach is also applied for TIMIT [22] and THCHS [23] speech corpus respectively, which are later used for classifier training and testing. I standardize all sounds into a single format, as monologue speech and sound effects come in different file formats (*e.g.*, WAVs, FLACs) and I use 16 kHz as our sample rate.

3.2 Creating scenarios

To cover different acoustic scenarios, I add three categories of sound effects: background noises, foreground noises and reverberation.

In any sound, some elements seem more prominent while others will seem to recede. In real-world recordings, non-speech sounds can be either background sound, foreground sound or a combination of both [24]. Usually, background noises either sound weaker or are continuous enough to make up the background texture of a soundscape. So in our process, background noises are added to the entire recording, and repeated as necessary to cover the full length. Multiple overlapping background noises are sometimes added. On the contrary, foreground noises, standing out against the background, are added sequentially, according to a specified interval, and do not overlap. To build a diverse background or foreground noise, I combine MUSAN-NOISE and MUSAN-MUSIC [20], and get a sound effect dataset that covers sound of things (*e.g.*, dialtones, fax machine noises), natural sounds (*e.g.*, thunder, wind), and music without vocal(*e.g.*, Western art music and popular genres). Finally, the speech mixtures are reverberated via convolution with simulated room impulse responses (RIRs) described in [21], which is to simulate how the mixtures sound like in different room settings. As is summarized in Table 3.1, thousands of realistic scenarios are

simulated by combining all the sound effects above.

Table 3.1 : Sound effects for creating scenarios

Sound effects	Parameters	Sound Effect Library
Background noise	SNR:0-18 dB	MUSAN-NOISE, MUSAN-MUSIC
Foreground noise	SNR:0-18 dB	MUSAN-NOISE, MUSAN-MUSIC
Reverberation	Small/Median/Large room	Simulated room impulse responses (RIRs)

Chapter 4

Speaker Count Estimation

In this chapter, I present AmbianceCount, an objective method to measure social ambiance by estimating the number of concurrent speakers.

Consider a scenario in which K people are co-located and are talking in pairs or in sub-groups. AmbianceCount is designed to estimate the number of concurrent speakers for each segment as belonging to one of 11 classes $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \text{other}\}$.

Given that our system is designed to be text-independent and there is no prior information about speakers, the performance highly depends on the ability of the neural networks to capture speech mixture characteristics.

While existing work [16, 17] developed end-to-end neural networks, I split the end-to-end approach into two parts: a DNN to produce embeddings and a separately trained classifier to compare them, which facilitates the use of backend methods to improve generalization. Different from the task-oriented style of end-to-end method, our method can work well in out-of-domain data, and has outperformed end-to-end methods in tasks like speaker verification [25].

As shown in Figure 4.1, the system consists of a front-end utterance embedding extractor and a back-end classifier. $s(t)$ denotes the input speech mixtures of duration T. In the following sections, I will explain the basic steps of the algorithm: Acoustic feature extraction (Section 4.1), utterance-level embedding extraction (Section 4.2), back-end scoring and domain adaptation (Section 4.3).

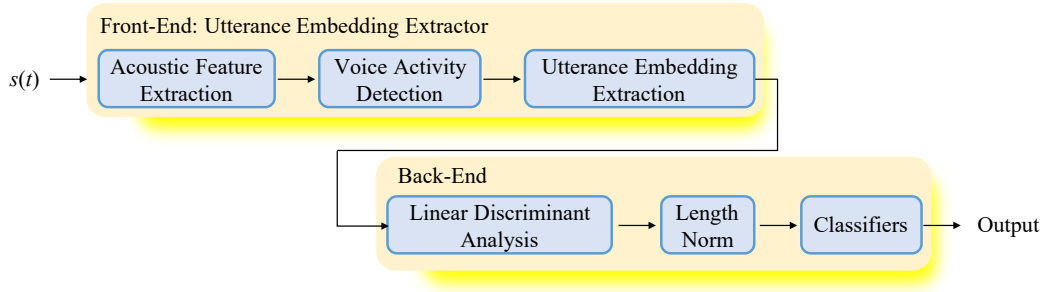


Figure 4.1 : The overview of AmbianceCount

4.1 Acoustic Features Extraction

Acoustic features, as input representations to neural networks, is a critical step to capture the characteristics of overlapped speech. Such representations can reduce redundancy of signals and capture significant sound characteristics, like phase information or periodicity.

Two types of features are explored and combined for experiments: Time-frequency features and low level pitch features.

4.1.1 Time-Frequency Representation

Time-frequency features are commonly used to represent speech signals due to their periodicity. The most commonly used time-frequency acoustic features in speech processing are Short-Time Fourier Transform (STFT), Filter Banks and Mel Frequency Cepstral Coefficients (MFCC). Computing these three involve the same pre-processing and framing procedures.

- **STFT** Speech signals change over time, but it is safe to assume that they are *quasi-stationary* within a short period of time. As a result, signals are split into 25 ms frames and a STFT is conducted over the short-time frame

using Hamming windows. After that, the power spectrum of the transform is computed, which is squared magnitude of the STFT of the signal

- **Filter Banks** Filter Banks are generated by applying 40 triangular filters to STFT on a Mel-scale to extract frequency bands. The Mel-scale aims to mimic the non-linear human perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies.

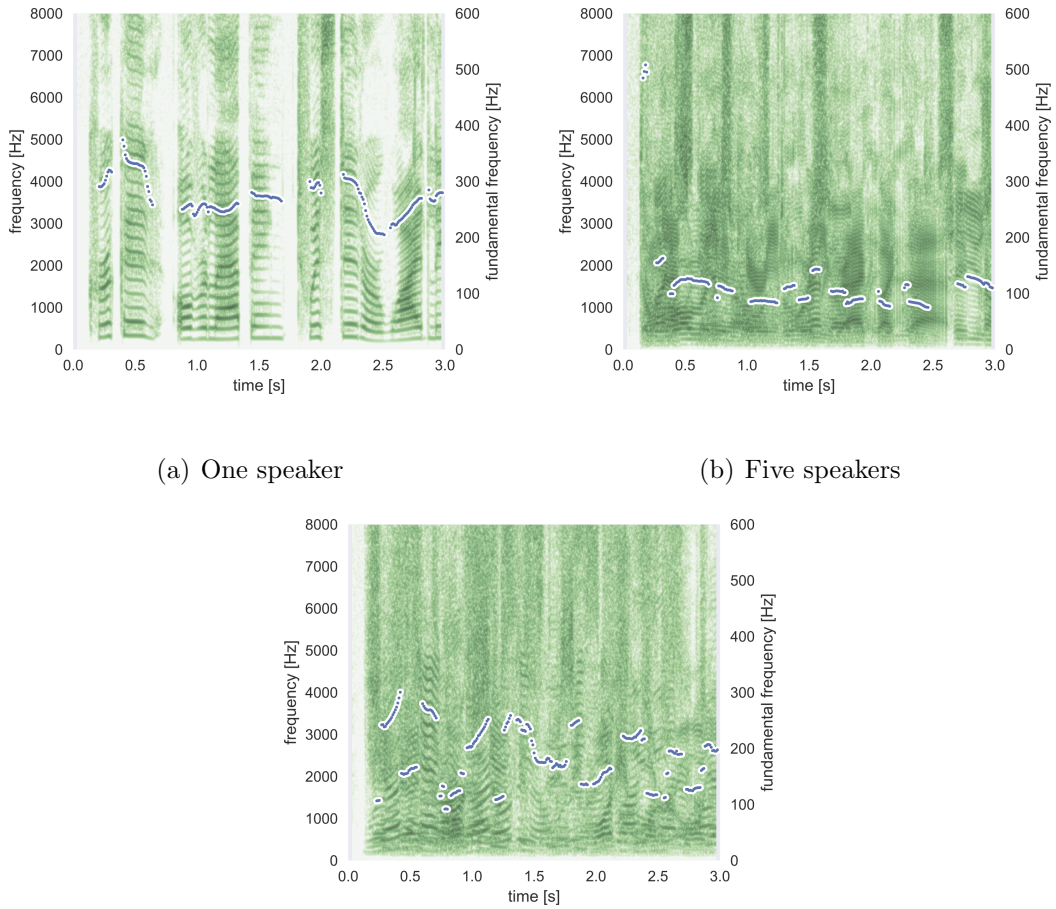
- **MFCC** MFCC is a compressed representation of Filter Banks by applying Discrete Cosine Transform (DCT) to Filter Banks and keeping only 23 cepstral coefficients. The reason for compression is to decorrelate the filter bank coefficients since some algorithms are not susceptible to highly correlated input.

Note that both Filter Banks and MFCC involve computing STFT. The extra step for Filter Banks is to mimic human perception of sound, and the extra steps for MFCC are motivated by the susceptibility of some machine learning algorithms to highly correlated input. It is worth exploring all above time-frequency representations to get more domain information about overlapped speech and to see how different neural networks react to correlated input.

4.1.2 Pitch

Pitch, also referred as fundamental frequency, is another good measure of speech periodicity. It is defined as the lowest frequency of a periodic waveform, and the ear identifies it as the specific pitch of the sound.

Figure 4.2(a), 4.2(b) and 4.2(c) show the spectrogram and pitch of different speech mixtures. It is observed that pitches become more discontinuous with an increasing number of speakers talking simultaneously. It seems that pitch is able to capture some



(a) One speaker

(b) Five speakers

(c) Nine speakers

Figure 4.2 : Spectrogram and Pitch of different speech mixtures

short-term characteristics from speech mixtures. Above observation is consistent with [19] where they describe that with an increasing number of speakers overlapped, the speech mixture is less time varying in the long term, but more chaotic in the short term.

4.2 Utterance Embedding Extraction

The Utterance Embedding Extraction can be considered as a neural network based front-end that maps sequences of acoustic features into utterance embedding vectors. As is shown in Figure 4.3, the framework architecture consists of three blocks. The first block is to extract frame-level features using neural networks, e.g. Time-Delay Neural Network (TDNN), Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN). The second block is a pooling layer that converts frame-level features to utterance vectors. The third block is to extract utterance-level embeddings before output layers. Different from basic x-vector extractor, an attention mechanism [26] is added to give different frame-level weights in order to capture long-term variations of speech mixture. To generate additional training data, SpecAugment [27] is utilized which applies directly to the acoustic frame vectors. Moreover, I propose a combined classification-regression objective function that help the computer estimate speaker count in a human's perspective, which will be covered in the following section.

4.2.1 Frame-level Feature Extractor

Usually data augmentation plays an important role in developing a robust algorithm by add noise to raw audio and enlarge the training data. Given that real-life scenarios are noisy and unpredictable, I need to explore other methods of generating additional noisy samples. SpecAugment [27] is gaining increasing attention recently because it is cheap to apply and does not require any additional data. Different from traditional data augmentation methods that apply on raw audio, SpecAugment operates directly on the spectrogram features of the input audio, as if it were an image. Specifically, it consists of three types of deformations of the spectrogram: time warping, frequency masking and time masking.

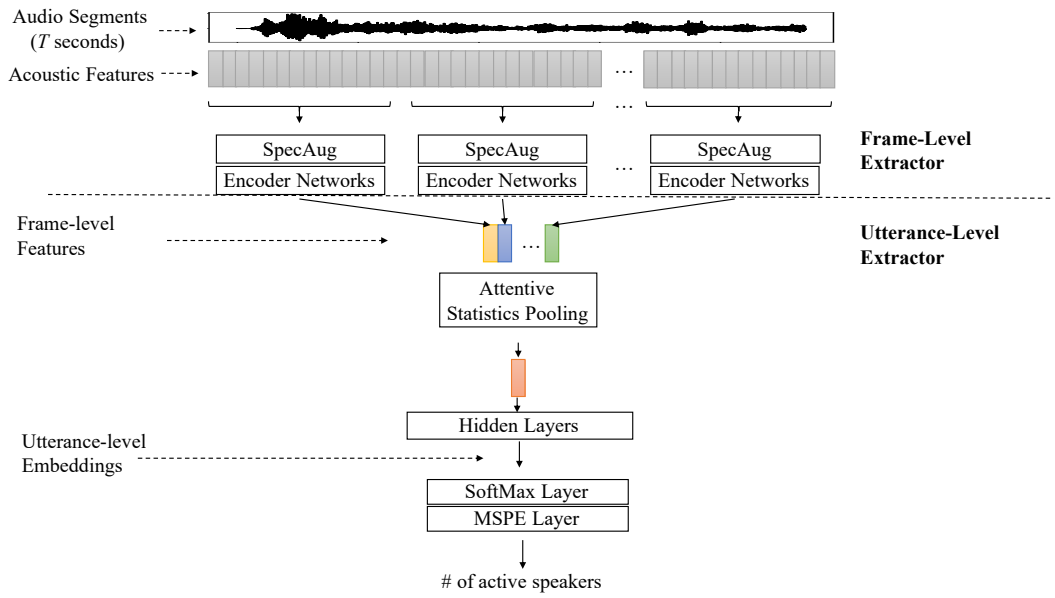


Figure 4.3 : Architecture of extracting utterance-level embeddings

Augmented acoustic features are then fed into an deep encoder network to extract frame-level embeddings. The DNN is leveraged to find a representation that can best discriminate different speech mixtures. I am going to explore five different network architectures: TDNN, Extended TDNN (E-TDNN), Factorized-TDNN (F-TDNN) and ResNet, which achieve great performance either in speaker count task or some other related field like speaker diariazation [6], speaker recognition [28, 29], age estimation [8] and spoken language recognition [9]. All architectures under investigation are summarized in Figure 4.4. Note that in this section I give basic topology of DNN architectures. Dimensions, normalization and dropout information will be finalized in the experiment section.

- *TDNN*: Time Delay Neural Networks (TDNNs) [7], also known as one dimensional Convolutional Neural Networks (1-d CNNs), are an efficient and well-performing neural network architecture for speech recognition. It is composed

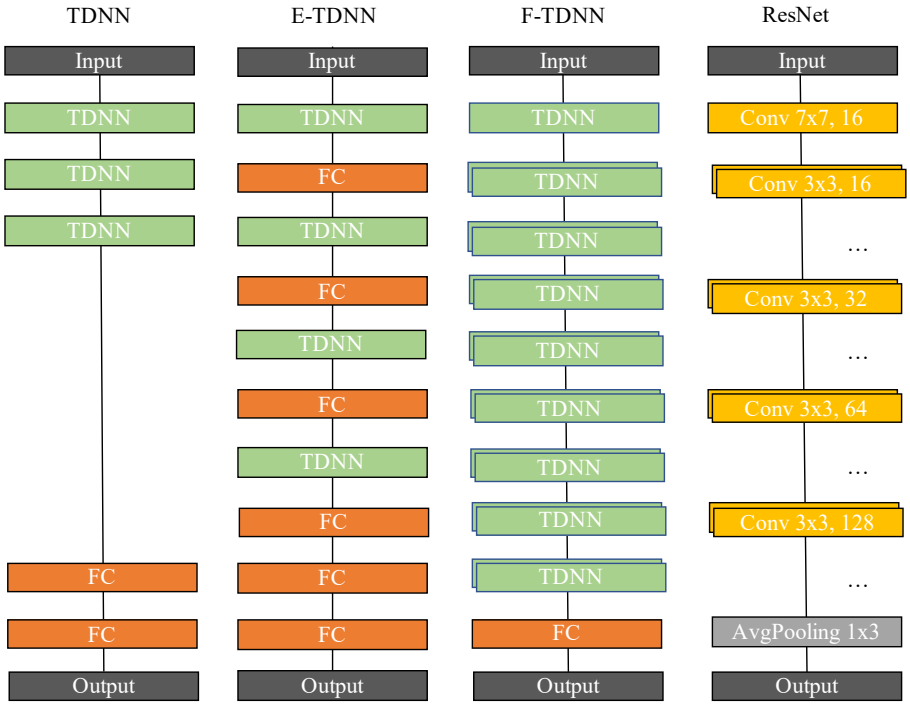


Figure 4.4 : Architectures of Encoder Networks

of three Time Delay layers, followed by two Fully Connected layers.

- *E-TDNN*: The Extended TDNN architecture (E-TDNN) is an extention of TDNN, with slightly wider temporal context and interleaves dense layers in between the Time Delay layers. Embeddings extracted from such architecture are very effective in both speaker recognition and speaker diarization [29].

- *F-TDNN*: F-TDNN [30] is a factored form of TDNN, with the same structure of Time Delay layers followed by Fully Connected layers. The difference is that the Time Delay layers in F-TDNN are compressed using Singular Value Decomposition (SVD) to reduce the number of parameters. The architecture gives substantial improvements over TDNNs in the area of speech recognition.

- *ResNet*: The ResNet Encoder Network is a variant of the x-vector system where the Time Delay layers are replaced by a residual network with 2D convolutions [**?**].

4.2.2 Utterance-Level Embedding Extractor

Frame-level features extracted from neural networks are concatenated and processed to form an utterance-level feature. Instead of averaging all frames equally, the attentive statistics pooling proposed by [26] is utilized to enable embeddings to focus on important frames and to obtain utterance-level representations with higher discriminative power. This is accomplished by giving different weights to different frames and generates not only weighted means but also weighted standard deviations.

Predicting discrete values is typically considered as classification problems with cross-entropy objective. However, the major drawback is that every class is penalized equally during classification and the relations between different estimates are lost. Since this is also a cardinality estimation task, regression based objective functions are also considered. For a similar task of age estimation [8], a combined classification-regression objective function is utilized by minimizing cross-entropy and mean squared error (MSE). By introducing MSE, the relations between different categories are kept and the penalization is conducted based on the deviation from ground truth. While the method works well for age estimation, it is not in line with our task based on the observation that humans are more discriminative when a speech overlaps on an "*smaller*" speech mixture. Adding a speaker to a big crowd, on the other hand, makes little difference since the background is already blurred.

Above example indicates that relative error is what matters in estimating how human perceive a speech mixture. To this end, I use a classification-regression objective

that jointly minimizes cross-entropy and Mean Square Percentage Error (MSPE) loss [10], where MSPE is to capture the relative deviation from ground truth.

$$\mathsf{L}_{\mathsf{Cross\ Entropy}} = -\frac{1}{n}\sum_{i=1}^{n} t_i \cdot \log(y_i) \tag{4.1}$$

$$\mathsf{L}_{\mathsf{MSPE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|z_i - t_i|}{|t_i|}\right)^2 \tag{4.2}$$

$$\mathsf{Loss} = \alpha \cdot \mathsf{L}_{\mathsf{Cross\ Entropy}} + \beta \cdot \mathsf{L}_{\mathsf{MSPE}} + \frac{\lambda}{2}\parallel \boldsymbol{W}\parallel_2^2 \tag{4.3}$$

where $\alpha, \beta \in [0, 1]$ represent weights of classification and regression objectives, t_i is the ground truth, y_i and z_i the output of classification layer and regression layer respectively, $\boldsymbol{W}$ denotes all trainable network parameters, and λ is the weight decay, which is applied to avoid over-fitting.

4.3 Back-end scoring and domain adaptation

To develop a system that works in unconstrained scenarios, the requirement is that unseen samples should be recognized without having to redesign the system. As mentioned in above sections, the embedding extractor is trained on a synthesized dataset, there must be a significant mismatch between the development dataset and application dataset in terms of speaker characteristics and channel conditions. Therefore, in this section, a back-end system is introduced to output the estimation result and generalize the algorithm.

As shown in Figure 4.5, the back-end system consists of three blocks, Correlation Alignment (CORAL) [31], Linear Discriminant Analysis (LDA) and Interpolated Probabilistic LDA. For scoring, a similarity score between embedding pairs is com-

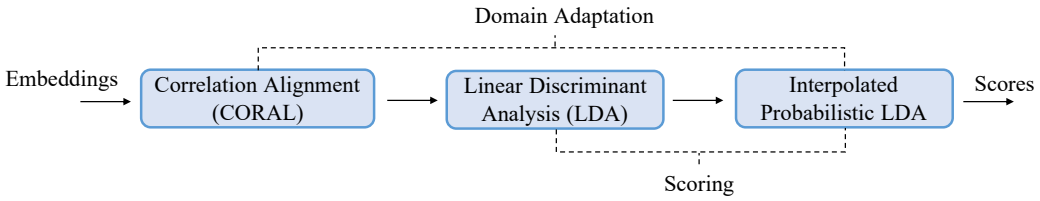


Figure 4.5 : Overview of the backend system

puted to determine whether an embedding belongs to class i or not. The key is to find a subspace to maximize inter-class differences and minimize the intra-class differences, which is accomplished by Linear Discriminant Analysis (LDA) and Interpolated Probabilistic LDA. For domain adaptation, a recently introduced domain adaptation algorithm called CORAL [31] is leveraged to align the distributions of out-of-domain and in-domain features in an unsupervised way.

4.4 Evaluation Metrics

Mean Percentage Error (MPE) is utilized to evaluate the performance to capture the relative deviation from ground truth, which is in line with the loss function I leveraged in the embedding extractor. Different from commonly used metrics like Mean Absolute Errors or Equal Error Rate, MPE is a better choice for our problem given two reasons below: Firstly, it is able to reflect the relative error, which keeps the relations between different class. Secondly, MPE in some way mimics humans perception system that are more capable of differentiate smaller mixtures. MPE for class k is calculated as:

$$\mathsf{MPE}(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{|\hat{t}_i - t_i|}{|t_i|} \tag{4.4}$$

where n_k is the number of testing samples for class k, $\hat{t}_i$ and t_i are the prediction and the ground truth respectively.

Therefore the averaged MPE is:

$$\mathsf{MPE} = \frac{1}{N}\sum_{k=0}^{N}\mathsf{MPE}(k) \tag{4.5}$$

Chapter 5

Performance Evaluation

In this section, the performance of AmbianceCount is assessed. First, the best model is selected with the combination of features and network architectures. Next, several experiments are conducted on the selected model to obtain results on LibriSpeech-test-clean, under different scenarios and on additional datasets.

5.1 Procedures and Experimental Settings

Pydub [32] is leveraged to synthesize the dataset and to pre-process the audio. The same data mixture strategy is applied to LibriSpeech [33], TIMIT [22], Voxceleb [34] and THCHS [23] for model development and validation. For model development, I combined three LibriSpeech subsets, clean-360, clean-100 and other-500, and get 960 hours of speech in utterances from 2338 speakers (1228 female speakers and 1210 male speakers). So under different scenario settings I have a training dataset of $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, others\}$ forming a total of 462,000 mixtures, with the duration of $T = 5$ seconds. In addition to the training dataset, I generated several fully separated validation dataset of 79,098 samples using a different set of speakers from LibriSpeech dataset. Similarly, I synthesize 22,000 mixture samples using TIMIT dataset, 16,000 samples using Voxceleb speakers and 10,800 mitures from THCHS for back-end training and evaluation.

I standardize all data into WAVs format and choose 16 kHz as our sample rate.

Voice Activity Detection is then applied using [35] and Kaldi toolkit [36]. Our models are developed based on asv-subtools [37] and Backends-for-SRE19 [38], which leverage Kaldi for acoustic feature extraction and back-end scoring, PyTorch for building neural networks and model training. The model is trained using the AdamW [39] (learning rate=0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.99$) with WarmRestarts [40], using a batch-size of 128. The training takes 20 epochs with the help of several NVIDIA 2080 Ti GPUs to accelerate the process.

5.2 Model Comparison

In this chapter, I intend to finalize the model by exploring different network architectures as well as acoustic features introduced in Chapter 4. The decision is determined based on performances of different configurations on a subset of the LibriSpeech testing data. As there are many combinations for selecting different features and embedding extractor, only top 3 models are listed as well as a baseline algorithm proposed in [16].

Table 5.1 : Model Comparison

Models	Acoustic Features	Pooling Layer	Evaluation Metrics (MPE)
ResNet34	FBANK+Pitch	Attentive Statistics	5.2%
F-TDNN	MFCC+Pitch	Attentive Statistics	8.9%
ResNet34	FBANK+Pitch	Statistics	9.6%
Baseline [16]	STFT	NA	10.7%

As listed in Table 5.1, pitch information and attentive pooling layer prove to boost the performance. As pitch information is able to catch the chaotic level in

the mixture, which is a very important clue to differentiate classes, and the attentive pooling layer can help capture long-term information of the mixture. From the top 3 models I can see FBANK seems to go well with ResNet while MFCC works better with F-TDNN. One possible reason is, ResNet is less susceptible to highly correlated input and is able to capture the information from FBANK without the compression process of MFCC. Remember the extra step it takes from FBANK to MFCC is to decorrelate the coefficients between FBANK.

Additionally, all the top3 models outperform our baseline algorithm [16] when applied to our testing data that mimics an unconstrained environment with different levels of reverberation and noise. Note that the baseline system is developed to distinguish from 0 to 10 speakers, I evaluate its performance within their defined classes. Though the goal of our baseline system is not entirely the same with our task, it is close to our novel task and also the state-of-the-art algorithm in estimating between up to 10 speakers.

5.3 Evaluation Results

In this section, the selected model is first evaluated on LibriSpeech-test under different settings. Then extensive experiments are conducted on two additional datasets, TIMIT and THCHS-30, which consist of new sets of speakers and different speech content.

5.3.1 Results on LibriSpeech-test-clean

The confusion matrices for the selected model is illustrated in Figure 5.1. It is observed that the model is able to give accurate count estimation when speaker count is smaller than 5. The accuracy drops to 56%-69% for speaker count from 6 to 10, and

increases to 98% when larger than 10. One possible reason is that, the *classification* part of the objective function tend to contribute to the result in a "one or many" way [17]. Moreover, it is shown that most of the errors locate in the adjacent classes of ground truth, which indicates that the *regression* part of the objective function enables the model to keep the relations between different classes during prediction. Overall, the model is more accurate for smaller speaker counts due to the percentage error penalty introduced. By penalizing more strictly on smaller speaker counts, the algorithm is able to mimic humans perception that is more discriminative when there are fewer sources in a mixture.

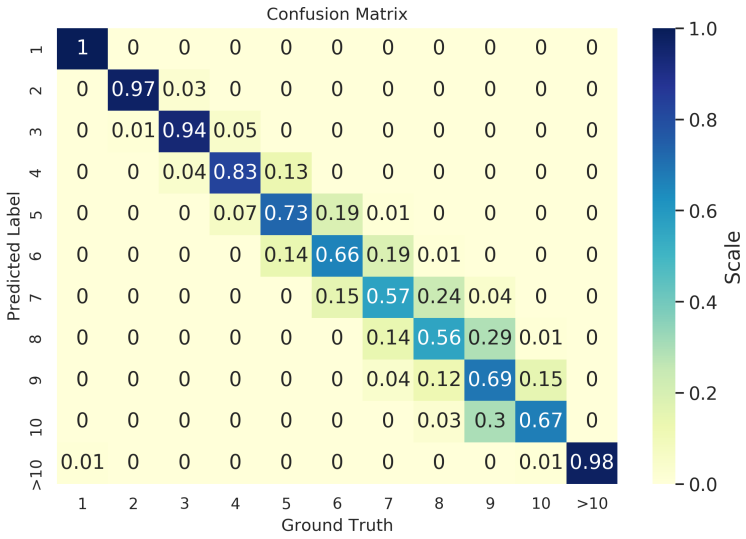


Figure 5.1 : Confusion matrices on LibriSpeech-test-clean

5.3.2 Results under different scenarios

Next, the model is evaluated on speech mixture combined with sound effects like noise and reverberation.

Previous works [41] [42] have reported noise and reverberation levels for specific scenarios, which provides an idea for us to simulate real-world scenarios. In this section, I evaluate the performance of our selected model under 10 categories(see Figure 5.2), from clean to noisy as 0 dB, and from no reverberation to high reverberation. Note that signal means the overlapped speech I intend to estimate, and category 0 represents the ideal situation where no noise and reverberation are added to the overlapped speech.

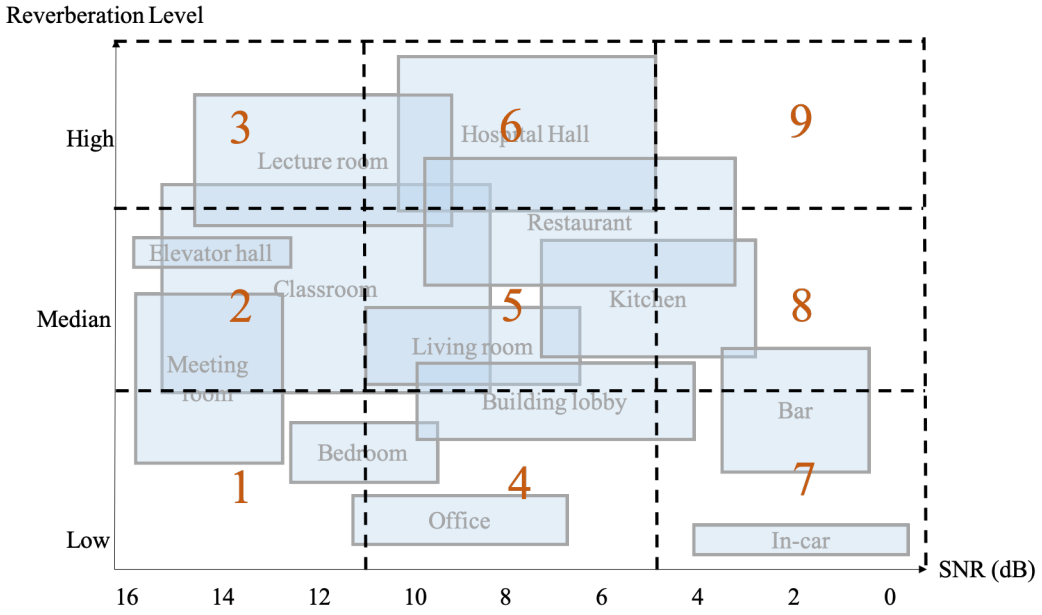


Figure 5.2 : Categories and real-world scenario examples

The detailed breakdown of the error counts for each category is illustrated in Figure 5.3. Results are averaged over five repeated experiments. AmbianceCount outperforms the baseline by a large margin, which illustrates that by introducing attentive pooling layer and an effective back-end, our model is able to capture the important information from the mixture and is robust in noisy environments. The

effect of noise and reverberation is also shown in Figure 5.3. MPE drops when SNR is smaller and when reverberation is higher, which means the algorithm is more accurate in scenarios like meeting rooms, offices compared with kitchens and bars. I also notice that reverberation seems to have a stronger influence on the performance than noise does. One possible reason is that reverberation creates a time-delayed speech signal so the system tends to overestimate the number of speakers in the mixture, as the number of over-estimated samples increases 22% from low to high reverberation.

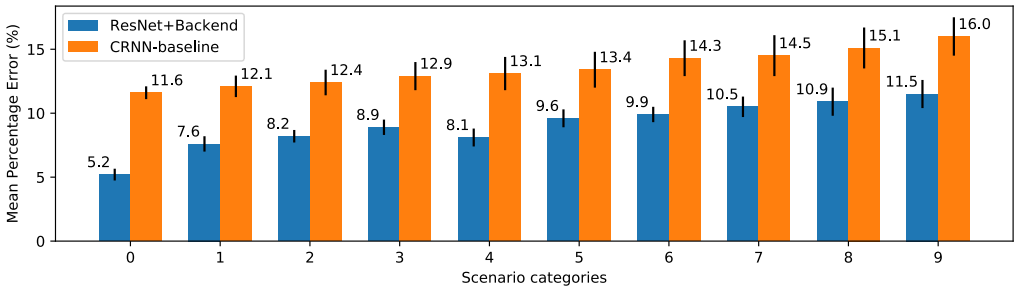
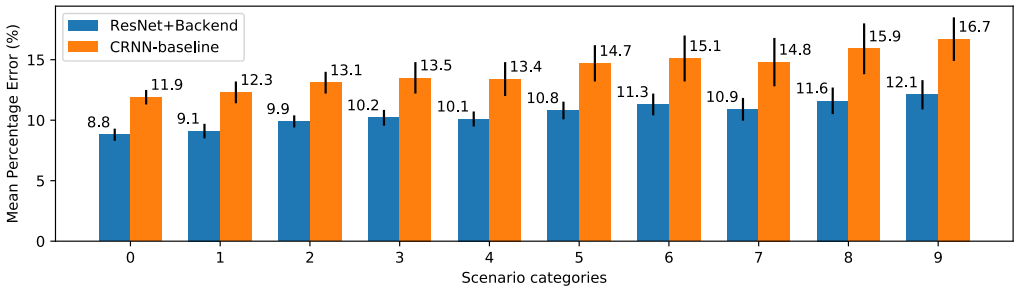


Figure 5.3 : Mean Percentage Error for overlapped speech synthesized from LibriSpeech under different scenario categories (averaged over five repeated experiments)

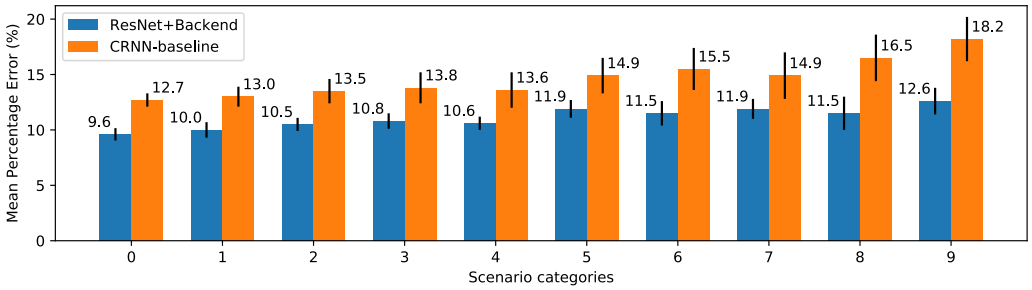
5.3.3 Results on additional datasets

To explore whether our speaker count algorithm is able to generalize from training data to unseen data, I test on speech mixtures created from TIMIT corpus and a Mandarin language THCHS-30 dataset. The results under different scenario categories are shown in Figure 5.4(a) and Figure 5.4(b). Mean percentage error increases slightly on TIMIT and THCHS-30, but overall, the algorithm is able to generalize well on datasets with different speakers, speech content and even in language. The result is consistent with [16] which suggests that the trained model is speaker and

language independent.



(a)



(b)

Figure 5.4 : (a) Mean Percentage Error for overlapped speech synthesized from TIMIT under different scenario categories (averaged over five repeated experiments). (b) Mean Percentage Error for overlapped speech synthesized from THCHS-30 under different scenario categories (averaged over five repeated experiments).

Chapter 6

Mental Health and Social Ambiance

To demonstrate the utility of our speaker count estimation in mental health, we apply AmbianceCount to explore the relation between social ambiance and mental status as well as personality traits. Three social ambiance levels are extracted from audio recordings, as shown in Figure 6.1:

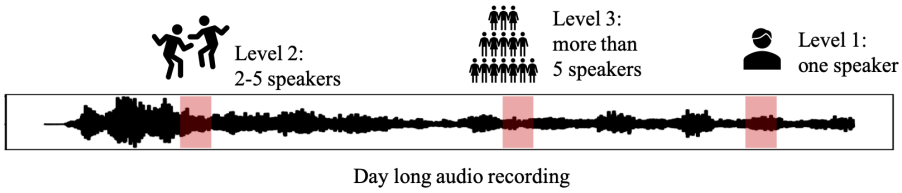


Figure 6.1 : An example of tagging social ambiance levels on day long audio recordings

While level 1 represents the participant is either involved in or around a conversation or monologue without interference, level 2 shows the participant is around a medium size crowd and level 3 captures those when one is around a bigger crowd of people.

Moreover, social ambiance patterns can be observed by calculating the distribution and variance of these levels throughout a day.

6.1 Dataset

I apply AmbianceCount to a sociability pilot, which captures multi-modal social interactions from depression and psychosis patients as well as age-matched healthy

controls. As shown in Figure 6.2, three types of data were collected using HealthSense [43]: (i) day-long audio recordings under unconstrained environments collected from wristband audio recorders, (ii) GPS locations and remote social interactions, i.e., phone calls and text message via continuous phone logging, and (iii) participants' responses to daily EMA questions, e.g.mood and sociability level, through HealthSense [43]. Besides, psychometric scores and personality measures were collected before the trial using Patient Health Questionnaire (PHQ-9) [44], General Anxiety Disorder-7 (GAD-7) [45], and Mini-IPIP Personality [46].

With multi-modal data collected from a clinical population, I am able to explore group level differences in terms of social ambiance patterns, as well as its association with psychometric scores and diverse personalities. In the following sections, I am going to introduce (i) the pipeline of extracting social ambiance patterns from unconstrained audio recordings, (ii) group-level differences in terms of social ambiance patterns and its correlation with clinical observations, (iii) association between social ambiance patterns and psychometric scores as well as personality traits.

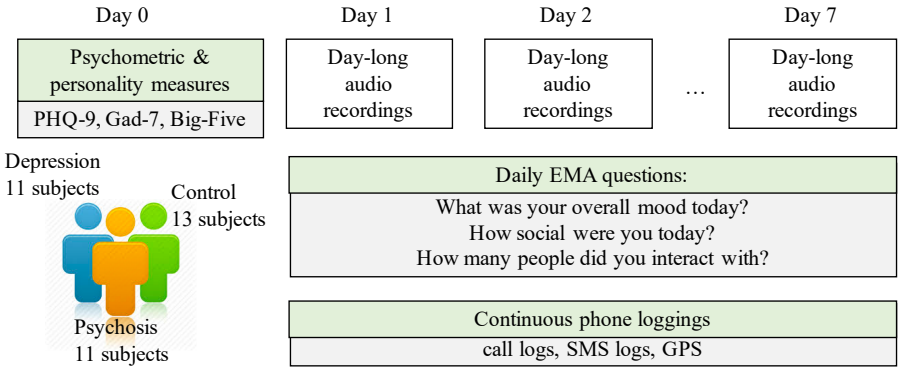


Figure 6.2 : Data and timeline of Sociability trial

6.2 Extract social ambiance patterns from audio

Social ambiance patterns are captured from unconstrained day-long audio with three steps. First, speech clips are extracted from unconstrained audio recordings using voice activity detection [35]. Next, the number of concurrent speakers in each clip is estimated using AmbianceCount. Finally, with the number of speakers mapped to three levels defined above, social ambiance pattern is estimated by calculating the distribution and variance of the levels.

While phone calls conversations were also recorded by the wristbands, the duration of phone calls makes only 15.8% of the recording for the control group, 10.2% for depression group and 16.4% for psychosis group. It should be safe to assume that what I captured from the recordings represents the perceived social ambiance patterns.

6.3 Correlation between group-level social ambiance patterns and clinical observations

Apart from above three social ambiance levels, consider no speech detected as social ambiance level 0. Entropy of social ambiance per day can be utilized as a proxy for environment variance. the Figure 6.3 shows the results for participants from three groups. It is observed that most psychosis patients have very small variance compared to other two groups, which is consistent with clinical observation.

Next, I compare across groups and analyze the difference in terms of the frequency of social ambiance levels.

By conducting one-way ANOVA test for the percentage of speech detected, I observe that there is a significant difference between three groups (p <0.1). It is shown in Figure 6.4(a) that the depression and psychosis patients have significantly

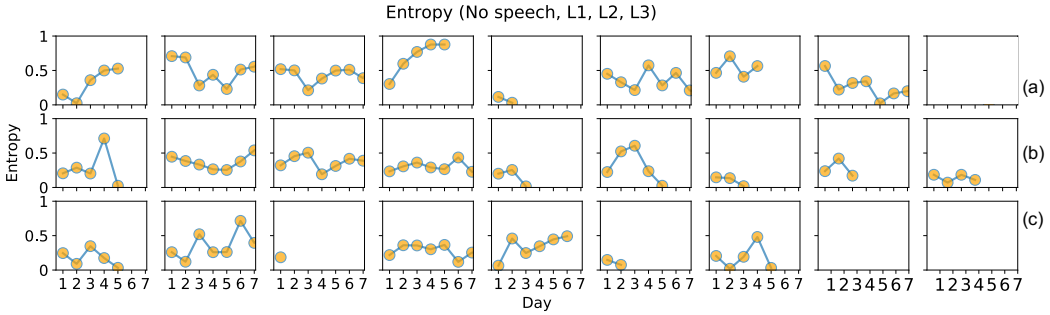
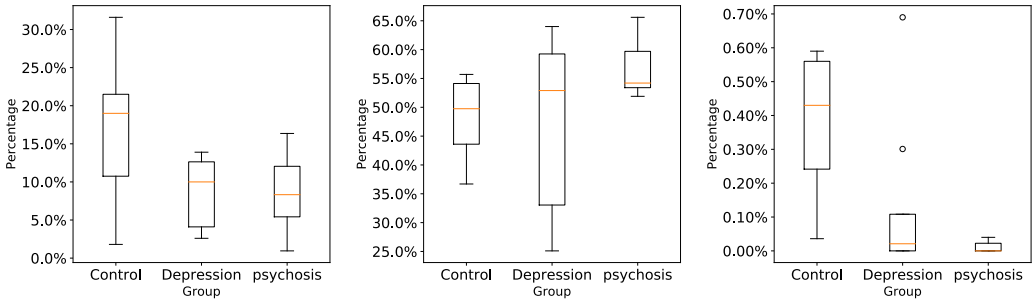


Figure 6.3 : Variation of social ambiance levels per day from (a) control group, (b) depression group, and (c) psychosis group.

reduced percentage of speech compared to the control subjects, which is consistent with clinical observations.



(a) (b) (c)

Figure 6.4 : One-way ANOVA test for (a) the percentage of speech detected, (b) the percentage of Level 1 , and (c) the percentage of Level 3

Also, the percentage of level 1 is compared across three groups. Note that level 1 represents there is only one active speaker detected, which means in these clips the participant is either in or around the conversation/monologue without interference. Results are shown in Figure 6.4(b). There are two observations that are consistent

with clinical observations. Firstly, depression patients tend to have a wide span over the percentage of Level 1. Compared to control subjects, individual differences play an important role in depression group. Secondly, participants from psychosis group have a small range in terms of Level 1 compared to other two groups. One-way ANOVA tests are conducted for three groups and results show that there is a significant difference between control group and psychosis group (p <0.1).

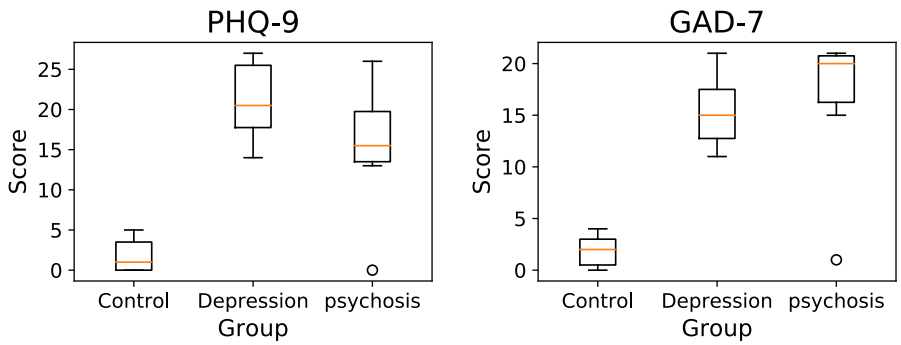
Another factor that differentiate three groups is the percentage of Level 3. Since Level 3 represents that there are more than five concurrent speakers detected in the audio, it represents how often our subjects were exposed to big and socially active crowds. The result of one-way ANOVA test shows that there is a significant difference between three groups (p <0.05). It is observed that the depression and psychosis patients have significantly reduced percentage of Level 3 compared to the control subjects, and psychosis group has the smallest chance to be exposed in bigger crowds.

6.4 Correlation between social ambiance patterns and personality scores

In section 6.3, it is observed that social ambiance patterns are significantly different across three groups, which is consistent with clinical observations. Meanwhile, individual differences are noticed within each group, which might reflect their mental status and diverse personalities.

IPHQ-9 [44], GAD-7 [45] are used as our psychometric and personality ground truth, which objectifies and assesses degree of depression and anxiety severity via questionnaire respectively.

Figure 6.5 shows the distribution of the PHQ-9 and GAD-7 scores across three groups. I conduct ANOVA tests and find that there are significant differences between three groups in terms of PHQ-9 (p <0.01) as well as GAD-7 (p <0.01).



(a) (b)

Figure 6.5 : Distribution of psychometric scores across control, depression and psychosis group: (a) PHQ-9 (p= 2.13e-5), (b) GAD-7 (p=2.98e-5)

Mini-IPIP Personality is utilized to measure [46] as our personality ground truth, which has five personality traits: extraversion, agreeableness, conscientiousness, neuroticism and intellect or imagination. Figure 6.6 shows the distributions of the five personality traits of participants from three groups. The distribution illustrates that there are significant differences across three groups in terms of Agreeableness, Conscientiousness and Neuroticism.

To investigate the association with ambiance patterns and above factors, for each group, I conduct multiple linear regression analyses of social ambiance patterns with psychometric scores and personality measures. The value of the coefficient indicates the direction and the strength of the association and the p-value indicates the probability that the coefficient is statistically significant.

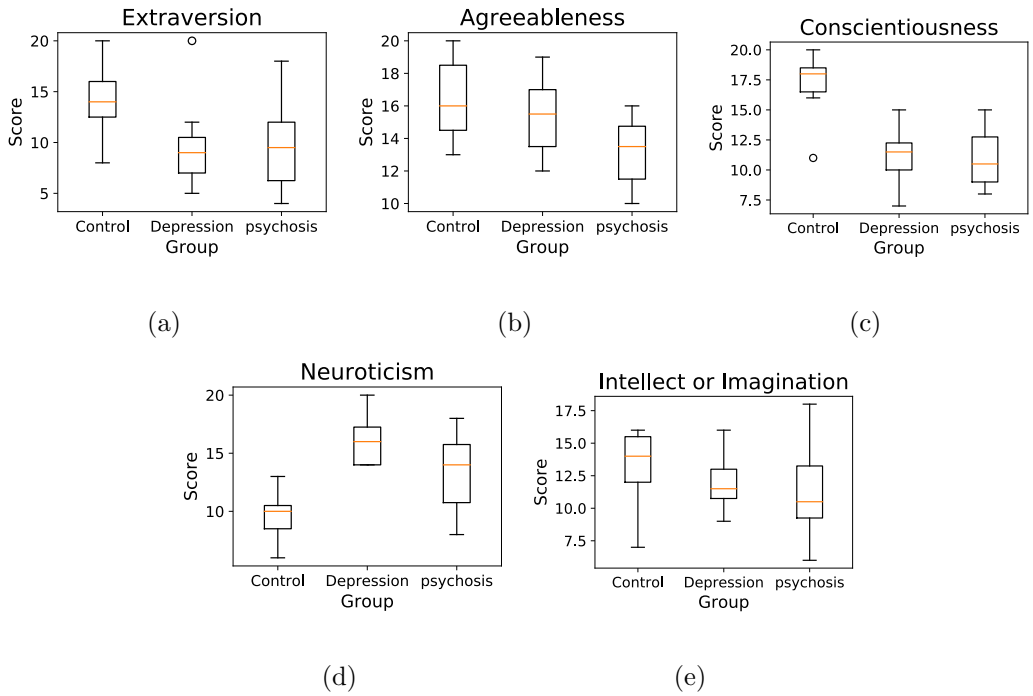


Figure 6.6 : Distribution of personality traits across control, depression and psychosis group: (a) Extraversion (p=0.152), (b) Agreeableness (p= 0.032) (c) Conscientiousness (p= 7e-4) (d) Neuroticism (p= 8e-4) and (e) Intellect or Imagination (p= 0.578)

The results in 6.4 indicate that participants who are more depressed are less likely to be around larger crowds, reflected by the negative coefficient between 'Percentage of level 3' and 'PHQ-9' (p <0.1). Compared to personality traits, the degree of depression severity seems to be more associated with the manifestation of social ambiance patterns, which suggests that sociability of depression patients is strongly affected by the severity of the depression symptoms.

Table 6.2 shows the multiple linear regression results for the psychosis group. It is observed that there is a positive association (p <0.01) between percentage of speech and Neuroticism, one of the personality traits. The results suggest that psychosis

Table 6.1 : Multiple linear regression analyses of social ambiance patterns with psychometric scores and personality measures from Depression group.

	PHQ-9	GAD-7	mini-IPIP				
			Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination
Percentage of speech	-0.51	0.10	-0.44	-0.33	0.23	-0.05	-0.11
Percentage of level 1	0.37	0.04	0.19	0.14	-0.02	-0.02	0.12
Percentage of level 3	**-19.84***	-5.42	3.46	0.60	-7.47	6.43	1.63

Figures are unstandardized regression coefficients. * p <0.1, **p <0.05, and ***p <0.01.

Table 6.2 : Multiple linear regression analyses of social ambiance patterns with psychometric scores and personality measures for Psychosis group

	PHQ-9	GAD-7	mini-IPIP				
			Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination
Percentage of speech	1.47	1.37	0.02	-0.05	-0.43	**0.74*** **	-0.23
Percentage of level 1	0.17	0.79	0.73	0.31	0.04	0.17	-0.11
Percentage of level 3	-14.35	-1.55	-5.12	3.26	0.04	-0.85	-12.26

Figures are unstandardized regression coefficients. * p <0.1, **p <0.05, and ***p <0.01.

patients who are more neurotic are more likely to be around people, where higher percentage of speech can be detected.

For control group, I also conduct a multiple linear regression and find a positive relation between social ambiance patterns and Agreeableness, one of the personality traits. Results in Table 6.3 indicate that more agreeable participants are more likely to be around bigger crowds (p <0.1).

Table 6.3 : Multiple linear regression analyses of social ambiance patterns with psychometric scores and personality measures for Control group.

	PHQ-9	GAD-7	mini-IPIP				
			Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination
Percentage of speech	0.21	0.10	-0.24	0.11	-0.21	-0.04	0.21
Percentage of level 1	0.06	-0.04	0.11	0.09	0.26	-0.04	0.68
Percentage of level 3	47.7	2.69	-6.84	**10.92***	1.55	-8.18	3.99

Figures are unstandardized regression coefficients. * p <0.1, ** p <0.05, and *** p <0.01.

Table 6.4 : Multiple linear regression analyses of social ambiance patterns with psychometric scores and personality measures from all participants.

	PHQ-9	GAD-7	mini-IPIP				
			Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination
Percentage of speech	-0.29	-0.27	-0.32	-0.01	0.03	-0.01	0.01
Percentage of level 1	-0.027	0.01	0.17	0.15	0.06	-0.04	0.08
Percentage of level 3	-8.91	-1.33	-0.72	**5.21****	-1.64	0.68	1.63

Figures are unstandardized regression coefficients. * p <0.1, ** p <0.05, and *** p <0.01.

Chapter 7

Conclusion

7.1 Summary

In this thesis, I propose AmbianceCount to objective measure social ambiance by estimating the number of concurrent speakers from unconstrained audio recordings. The proposed AmbianceCount is robust against noise and reverberation, and is able to generalize well on unseen data. To simulate human perception in terms of speech mixtures, I devise a classification-regression objective function which keeps the ordering information and conducts penalty based on percentage deviation. Finally, I apply the algorithm to Sociability pilot, which captures multi-modal social interactions from a clinical population. The result shows that participants from three groups exhibits significant differences in social ambiance patterns. Additionally, the correlation between objectively captured social ambiance with personality measure is consistent with clinical observations.

7.2 Discussion

7.2.1 Adapting AmbianceCount across devices

Given that a realistic dataset of fully overlapped speakers is not available, AmbianceCount now is developed and evaluated on speech mixtures synthesized from LibriSpeech, TIMIT and THCHS-30, which are all read speech datasets based on audio books. Next I will adapt the model across difference devices in order to improve the

robustness of AmbianceCount,

7.2.2 Jointly analyzing social ambiance patterns with smartphone data

Currently, I extract social ambiance patterns from unconstrained audio recordings and explore their correlation with psychometric scores and personality measures. In the Sociability pilot, smartphone loggings and phone usage patterns are also collected to track remote social interactions and mobility patterns. Next step I will jointly analyze extracted social ambiance patterns with smartphone data, e.g., GPS location, call and text logs. smartphone. By combining both smartphone sensor data and audio recordings, I will have a better understanding of social ambiance and its impact on mental health.

Bibliography

[1] O. Dictionary., "http://bit.ly/1day80r,"

[2] J. Collins, B. M. Ward, P. Snow, S. Kippen, and F. Judd, "Compositional, contextual, and collective community factors in mental health and well-being in australian rural communities," *Qualitative health research*, vol. 27, no. 5, pp. 677–687, 2017.

[3] E. Kafadar, V. A. Mittal, G. P. Strauss, H. C. Chapman, L. M. Ellman, S. Bansal, J. M. Gold, B. Alderson-Day, S. Evans, J. Moffatt, *et al.*, "Modeling perception and behavior in individuals at clinical high risk for psychosis: support for the predictive processing framework," *Schizophrenia Research*, 2020.

[4] R. Wang, M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, *et al.*, "Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 886–897, 2016.

[5] H. Wang, D. Lymberopoulos, and J. Liu, "Local business ambience characterization through mobile audio sensing," in *Proceedings of the 23rd international conference on World wide web*, pp. 293–304, 2014.

[6] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in

ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5796–5800, IEEE, 2019.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[8] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end deep neural network age estimation.," in *INTERSPEECH*, pp. 277–281, 2018.

[9] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors.," in *Odyssey*, pp. 105–111, 2018.

[10] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," *arXiv preprint arXiv:1809.03006*, 2018.

[11] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "Streetscore-predicting the perceived safety of one million streetscapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 779–785, 2014.

[12] V. Ordonez and T. L. Berg, "Learning high-level judgments of urban perception," in *European conference on computer vision*, pp. 494–510, Springer, 2014.

[13] D. Quercia, N. K. O'Hare, and H. Cramer, "Aesthetic capital: what makes london look beautiful, quiet, and happy?," in *Proceedings of the 17th ACM confer-*

ence on Computer supported cooperative work & social computing, pp. 945–955, 2014.

[14] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[15] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guérin, "High-resolution speaker counting in reverberant rooms using crnn with ambisonics features," *arXiv preprint arXiv:2003.07839*, 2020.

[16] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2018.

[17] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped speech detection and competing speaker counting—humans versus deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 850–862, 2019.

[18] S. Chen, J. Epps, E. Ambikairajah, and P. N. Le, "An investigation of crowd speech for room occupancy estimation.," in *INTERSPEECH*, pp. 324–328, 2017.

[19] N. Krishnamurthy and J. H. Hansen, "Babble noise: modeling, analysis, and applications," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1394–1407, 2009.

[20] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, IEEE, 2017.

[22] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[23] D. Wang and X. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[24] M. Thorogood, J. Fan, and P. Pasquier, "Soundscape audio signal classification and segmentation using listeners perception of background and foreground sound," *Journal of the Audio Engineering Society*, vol. 64, pp. 484–492, 08 2016.

[25] D. Wang, L. Li, Z. Tang, and T. F. Zheng, "Deep speaker verification: Do we need end to end?," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 177–181, IEEE, 2017.

[26] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[28] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, *et al.*, "State-of-the-art

speaker recognition for telephone and video speech: The jhu-mit submission for nist sre18.," in *Interspeech*, pp. 1488–1492, 2019.

[29] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5796–5800, IEEE, 2019.

[30] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, pp. 3743–3747, 2018.

[31] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *arXiv preprint arXiv:1511.05547*, 2015.

[32] J. Robert, M. Webbie, *et al.*, "Pydub," 2018.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[34] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[35] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5214–5218, IEEE, 2018.

[36] "Kaldi."

[37] "asv-subtools," 2020.

[38] "Backends-for-sre19," 2019.

[39] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.

[40] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[41] D. Ribas, E. Vincent, and J. R. Calvo, "A study of speech distortion conditions in real scenarios for speech processing applications," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 13–20, IEEE, 2016.

[42] K. Smeds, F. Wolters, and M. Rung, "Estimation of signal-to-noise ratios in realistic sound scenarios," *Journal of the American Academy of Audiology*, vol. 26, no. 2, pp. 183–196, 2015.

[43] A. Curtis, A. Pai, J. Cao, N. Moukaddam, and A. Sabharwal, "Healthsense: Software-defined mobile-based clinical trials," in *The 25th Annual International Conference on Mobile Computing and Networking*, pp. 1–15, 2019.

[44] R. L. Spitzer, K. Kroenke, J. B. Williams, P. H. Q. P. C. S. Group, *et al.*, "Validation and utility of a self-report version of prime-md: the phq primary care study," *Jama*, vol. 282, no. 18, pp. 1737–1744, 1999.

[45] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[46] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality.," *Psychological assessment*, vol. 18, no. 2, p. 192, 2006.