Datastore Design for Analysis of Police Broadcast Audio at Scale

Ayah Ahmad
University of California, Berkeley
Department of Electrical Engineering & Computer Sciences
Berkeley, California, USA
ayahahmad@berkeley.edu

Christopher Graziul (advisor)
University of Chicago
Department of Comparative Human Development
Chicago, Illinois, USA
graziul@uchicago.edu

Margaret Beale Spencer (advisor)
University of Chicago
Department of Comparative Human Development
Chicago, Illinois, USA
mbspencer@uchicago.edu

ABSTRACT

With policing coming under greater scrutiny in recent years, researchers have begun to more thoroughly study the effects of contact between police and minority communities. Despite data archives of hundreds of thousands of recorded Broadcast Police Communications (BPC) being openly available to the public, a closer look at a large-scale analysis of the language of policing has remained largely unexplored. While this research is critical in understanding a "pre-reflective" notion of policing, the large quantity of data presents numerous challenges in its organization and analysis.

In this paper, we describe preliminary work towards enabling Speech Emotion Recognition (SER) in an analysis of the Chicago Police Department's (CPD) BPC by demonstrating the pipelined creation of a datastore to enable a multimodal analysis of composed raw audio files.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Information systems** → **Temporal data**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

temporal data, datastores, audio analysis, speech emotion recognition, feature extraction

ACM Reference Format:
Ayah Ahmad, Christopher Graziul (advisor), and Margaret Beale Spencer (advisor). 2021. Datastore Design for Analysis of Police Broadcast Audio at Scale. In *St. Louis '21: The International Conference for High Performance Computing, Networking, Storage, and Analysis, November 14–19, 2021, St. Louis, MO.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.****/*****.*******

1 INTRODUCTION

In this section, we discuss the data that we operate on, relevant literature that influenced datastore design choices, and the framework that forms the foundation for this project.

St. Louis '21, November 14–19, 2021, St. Louis, MO

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.****/*****.*******

1.1 Data

The data we operated on consisted of a public archive of 160,000 30-minute audio files of the CPD's BPC. Each audio file is 30 minutes long and approximately 3.5 MB, totaling approximately 4.8 million minutes and 560 GB for the raw archive. Associated with each audio file was metadata extracted from the name of the file, including dispatch zone and date, and metadata extracted using Voice Activity Detection (VAD), including timestamps of non-silent slices of audio.

1.2 Literature Review

To survey the research in Speech Emotion Recognition, we relied heavily on [1] to contextualize and summarize pertinent SER models. In this search, we sought out prevalent research that focused on, or was bound by, the following constraints:

(1) Examined **elicited emotions** during improvised conversations, as opposed to acted emotions
(2) Was **robust to noise**, particularly at a level close to human speech
(3) Used a **3-dimensional model of emotion**, instead of a categorical model

These constraints fundamentally allowed us to look closer at the research more applicable to the data we were analyzing. While no model that we examined fulfilled all three requirements, [3] was perhaps the closest. There, they created a 3-D Convolutional Recurrent Neural Network (CRNN) for SER. To avoid adding potential bias, we decided to extract audio features and use those as inputs, instead of using human-labeled data.

1.3 PVEST Framework

The Phenomenological Variant of Ecological Systems Theory [7] serves as the theoretical framework supporting this project. In this context, the framework seeks to identify adaptive and maladaptive coping mechanisms of police officers' stress responses. Thus, we seek to understand how normal communication can contribute to increasing or decreasing the likelihood of an adverse encounter between police and the general public—and more specifically, Law Enforcement Officers (LEOs) and Male Minority Youth (MMY).

2 CHALLENGES

This section will discuss challenges associated with the creation of a database, due to the scale, temporality, and silence of the data.

2.1 Scale

In expanding the audio into discrete samples, we ended up with approximately 40 million data points. From there, we extracted 26 temporal features, at approximately 183,000 samples per feature—based on the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [4]—using openSMILE [5], resulting in approximately 690,000 data points per file. Using Praat-Parselmouth [2, 6] to extract intensity, harmonicity, and pitch for each file resulted in approximately 230,000 data points per feature, per file. Scaling upwards, to include all 160,000 audio files results in approximately 7.2 trillion data points for the raw audio, GeMAPS, and Parselmouth files.

2.2 Temporality

When extracting different features from individual files, the default sample rate varies from one program to another. Thus, for each audio file, we have both raw audio data for each 22kHz sampling period and features extracted at differing periods of time.

2.3 Silence

Since some files contain silent slices, clustering on data that contains silence could lead to an inherently binary model, explained by one dimension—silence or sound.

3 DATABASE DESIGN AND IMPLEMENTATION

In searching for a database management system (DBMS) that was scalable, and ACID-compliant, extensible, with high levels of concurrency, we decided to use PostgreSQL.

Operating under a 1TB constraint meant that we could not store our raw and extracted data directly in the database because this composition of features exceeded 30 TB. This was in addition to the constraint set by the misalignment in temporality. Thus, we determined to design a datastore, such that raw and extracted features could simultaneously be accessed and preprocessed as inputs to a CRNN. Each file is stored in a specified location, with the locations used instead of the files in the database. Thus, for any feature stored as a column in the database, a script would extract the file locations, and feed them into a clustering algorithm that would parse the given file and cluster the data accordingly. Similarly, for the SER model, a script would perform the same parsing of the files and use the parsed data as inputs to the model.

4 CONCLUSION

In this project, we created a framework that enabled easy interoperability with statistical methods for an unbiased large-scale analysis of police broadcast audio for SER, allowing us to do large-scale pre-processing, clustering and PCA on the dataset.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number R01MD015064.

REFERENCES

[1] Berkehan Akçay and Kaya Oguz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (01 2020). https://doi.org/10.1016/j.specom.2019.12.001

[2] Paul Boersma and David Weenink. 2021. Praat: Doing Phonetics by Computer [Computer program]. http://www.praat.org/ retrieved 22 July 2021.

[3] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 2018. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters* 25, 10 (2018), 1440–1444. https://doi.org/10.1109/LSP.2018.2860246

[4] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Phuong Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE transactions on affective computing* 7, 2 (April 2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417 Open access.

[5] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. https://doi.org/10.1145/1873951.1874246

[6] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[7] Margaret Spencer. 2007. *Phenomenology and Ecological Systems Theory: Development of Diverse Groups*. Vol. 1. https://doi.org/10.1002/9780470147658.chpsy0115