Investigating Frame Size Effects on Mental State Classification from the Androids Corpus

Vadim Sokolov

Department of Computer Science, University of Milan, Milan, Italy

Abstract

Mental state detection from speech is an important task in clinical settings, where non-invasive methods can support early diagnosis and monitoring. In this work, we investigate how varying frame sizes affect the accuracy of mental state prediction using audio from the Androids Corpus. We extract standard acoustic features (MFCCs, deltas, RMS) with varying temporal windows and evaluate performance using a linear classifier under a speaker-independent 5-fold protocol. Preliminary results show a substantial improvement in accuracy as frame size increases, confirming the hypothesis that mental states vary slowly. TODO: describe the results of further experiments with nonlinear models and additional acoustic descriptors.

1 Introduction

Mental health disorders such as depression affect millions worldwide. Automatic detection of such conditions from speech offers a non-invasive, scalable, and cost-effective screening tool. Speech contains both linguistic and paralinguistic cues that can correlate with psychological states.

In this study, we aim to explore how temporal framing in audio feature extraction affects classification performance. The assumption is that mental states change slowly over time, and thus longer frames might capture more relevant descriptors.

2 Related Work and Motivation

Previous work on the Androids Corpus [1] uses features extracted with OpenSMILE and evaluates classifiers using a speaker-independent 5-fold protocol. Other studies have employed deep learning, but often neglect the temporal resolution of acoustic features.

Our goal is to systematically explore different frame lengths to understand how temporal granularity affects classification. We hypothesize that longer frames improve performance by capturing more stable features.

3 Methodology

3.1 Dataset

We use the Androids Corpus, which contains recordings from interviews and reading tasks by individuals classified as either healthy controls or patients. The corpus includes:

- Interview-Task audio clips (874 files)

- Reading-Task recordings (112 files)
- Labels: condition (PT vs. C)

3.2 Feature Extraction

We extract the following features using `librosa`:

- 13 MFCCs
- Delta and Delta-Delta of MFCCs
- Root Mean Square (RMS) Energy

We experiment with various frame sizes: 30ms, 100ms, 250ms, 500ms, 1000ms, and 5000ms. All features are normalized using standard scaling.

3.3 Classification

We start with a logistic regression model and evaluate frame-level accuracy and F1 score. The dataset is split using a speaker-independent 5-fold division, consistent with the original baseline setup.

4 Experiments and Results

4.1 Evaluation Metrics

We use:

- Frame-level accuracy
- Frame-level F1 score
- Confusion matrix (TBD)

4.2 Results

Initial results show that performance improves with larger frame sizes, peaking around 500–1000ms. See Figure 1.

5 Discussion

Larger frames provide better performance, suggesting that mental state-related features are better captured over longer time spans. Short frames likely introduce variability and noise.

TODO:

- Train nonlinear models (Random Forest, SVM, MLP)
- Add fundamental frequency and harmonicity-based features
- Perform feature importance analysis across frame sizes

6 Conclusion

This paper presents an analysis of frame size on speech-based mental state classification. Our experiments show a clear performance trend in favor of longer frames, supporting the idea that slowly varying descriptors matter more. TODO: refine these findings.

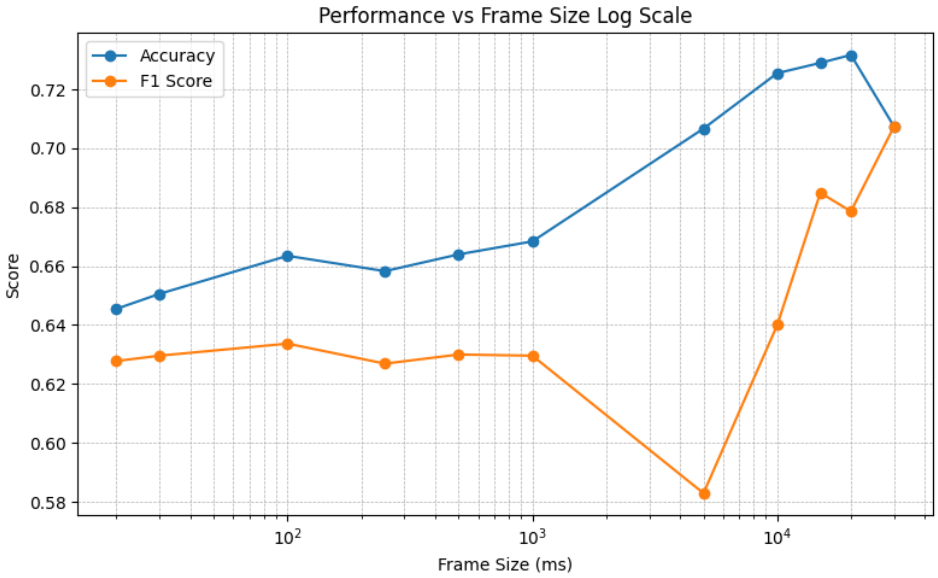

Figure 1: Accuracy and F1 Score vs Frame Size (log-scale)

References

[1] Alessandro Vinciarelli, University of Glasgow et al. *The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection*. Interspeech 2023.

[2] Brian McFee et al. *librosa: Audio and music signal analysis in Python*. Proceedings of the 14th python in science conference. 2015.

[3] Pedregosa et al. *Scikit-learn: Machine Learning in Python*. JMLR 2011.