Paulo Moura, William Liaw

Analyzing and Extending Time Series Kernels based on Nonlinear Vector AutoRegressive Delay Embeddings

This report is based on the paper (FELICE; GOULERMAS; GUSEV, 2023), published at NeurIPS 2023.

Palaiseau
2024

Paulo Moura, William Liaw

Analyzing and Extending Time Series Kernels based on Nonlinear Vector AutoRegressive Delay Embeddings

This report is based on the paper (FELICE; GOULERMAS; GUSEV, 2023), published at NeurIPS 2023.

Advisor: Prof. Florence D'Alché

Palaiseau
2024

ABSTRACT

In this work we study the...

Keywords: KEYWORD. KEYWORD.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\Delta(h)$ Assinatura diádica

CONTENTS

1 INTRODUCTION

Here you should give the context, justifications...

Do yourself a favor and follow the structure guidelines in the file *Research_structure_guidelines*
It should make your life easier.

In this template I will leave examples on how to cite, reference chapters, tables, figures, use math symbols along the text, write equations, label them for further referencing, use cases in equations, write tables, include figures, use special math formatting and symbols, use proof environments for theorems, matrix environments, etc. Remember to build the main file *thesis_main.tex* to visualize the updated pdf.

Example of how to reference a chapter: This dissertation is structured in... chapters. In Chapter 2 we present a ...

2 ANALYSIS

2.1 Datasets

The selection of appropriate datasets plays a crucial role in the development, training, and evaluation of emotion recognition models. In this work, we utilize two widely recognized datasets that are frequently employed in the domain of speech and audio-visual emotion classification: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Berlin Database of Emotional Speech (Emo-DB). Both datasets contain emotionally expressive performances from professional actors, providing high-quality labeled data that is indispensable for benchmarking and advancing the capabilities of emotion classification systems. The following subsections provide a detailed description of each dataset, highlighting their key features, structure, and the rationale for their use in this study. All the datasets' time series were downsampled to with a sampling rate of 8kHz for the purposes of this study.

2.1.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (LIVINGSTONE; RUSSO, 2018) is a widely recognized benchmark dataset utilized extensively in audiovisual emotion classification research (ANUSHA et al., 2021; VIMAL et al., 2021; ABDULLAH; AHMAD; HAN, 2020). The dataset consists of short audio and video recordings that feature both spoken and sung performances, enacted by a cohort of 24 actors (12 male and 12 female). Each recording is labeled with one of the following emotion categories: *angry*, *calm*, *disgust*, *fearful*, *happy*, *neutral*, *sad*, and *surprised*.

To promote consistency and reproducibility, each actor delivers two predefined phrases in English: "Kids are talking by the door" and "Dogs are sitting by the door." Apart from the neutral category, all emotions are expressed at two distinct intensity levels (normal and strong), with each instance repeated twice. These structured variations in emotional intensity, repetition, and diversity of vocal expressions make RAVDESS an invaluable asset for the development and validation of emotion recognition models in a

wide range of applications.

For the audio-only subset of the dataset which we employ for further analysis in the present work, there are a total of 1440 speech recordings and 1012 song recordings. It is worth noting that the singing subset is slightly smaller, as one actor's data is missing, and the emotions *sad* and *surprised* are not included for singing performances.

Despite its favorable reception within the academic community, as demonstrated by its widespread adoption, evidence suggests that the application of RAVDESS in real-world scenarios may lead to underwhelming results (CHURAEV; SAVCHENKO, 2021). One possible explanation for this discrepancy is the issue of data leakage. Specifically, an overlap of similar samples between the training and validation sets may result in unintended information sharing, thereby artificially inflating performance metrics. This overestimation does not accurately reflect the generalizability and practical effectiveness of models when deployed in real-world environments.

2.1.2 Berlin Database of Emotional Speech

The Berlin Database of Emotional Speech (Emo-DB) (BURKHARDT et al., 2005), akin to RAVDESS, is a well-regarded dataset for speech emotion classification tasks (SINITH et al., 2015; KOTTI; KOTROPOULOS, 2008; YING; ZHANG, 2010). It comprises short spoken audio recordings performed by 10 professional actors (5 male and 5 female), each enacting various grammatical phrases in German, as detailed in Table 1. Each recording is annotated with one of the following emotion categories: *anger*, *anxiety/fear*, *boredom*, *disgust*, *happiness*, *neutral*, and *sadness*.

To ensure the quality and reliability of the dataset, these samples underwent evaluation by a significant number of listeners, who assessed the naturalness of the emotional expressions. In total, the dataset comprises 535 speech files.

Table 1: Grammatical phrases in the Emo-DB dataset

German	English
Der Lappen liegt auf dem Eisschrank.	The cloth is on the refrigerator.
Das will sie am Mittwoch abgeben.	She will deliver it on Wednesday.
Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there next to the piece of wood.
In sieben Stunden wird es soweit sein.	In seven hours it will be time.
Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags that are under the table?
Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going back down.
An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	On weekends, I now always went home and visited Agnes.
Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just take this away and then go have a drink with Karl.
Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always put it.

Source: Own authorship

3 RESULTS

4 CONCLUSION

In this work we have considered ...

REFERENCES

ABDULLAH, M.; AHMAD, M.; HAN, D. Facial expression recognition in videos: An cnn-lstm based model for video classification. In: **2020 International Conference on Electronics, Information, and Communication (ICEIC)**. [S.l.: s.n.], 2020. p. 1–3.

ANUSHA, R. et al. Speech emotion recognition using machine learning. In: **2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)**. [S.l.: s.n.], 2021. p. 1608–1612.

BURKHARDT, F. et al. A database of German emotional speech. In: . [S.l.: s.n.], 2005. v. 5, p. 1517–1520.

CHURAEV, E.; SAVCHENKO, A. V. Touching the limits of a dataset in video-based facial expression recognition. In: **2021 International Russian Automation Conference (RusAutoCon)**. [S.l.: s.n.], 2021. p. 633–638.

FELICE, G. D.; GOULERMAS, J. Y.; GUSEV, V. Time series kernels based on nonlinear vector autoregressive delay embeddings. In: **Thirty-seventh Conference on Neural Information Processing Systems**. [s.n.], 2023. Disponível em: <https://openreview.net/forum?id=UBUWFEwn7p>.

KOTTI, M.; KOTROPOULOS, C. Gender classification in two emotional speech databases. In: **2008 19th International Conference on Pattern Recognition**. [S.l.: s.n.], 2008. p. 1–4.

LIVINGSTONE, S. R.; RUSSO, F. A. The Ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in North American English. **PLOS ONE**, Public Library of Science, v. 13, n. 5, p. 1–35, 05 2018. Disponível em: <https://doi.org/10.1371/journal.pone.0196391>.

SINITH, M. S. et al. Emotion recognition from audio signals using support vector machine. In: **2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)**. [S.l.: s.n.], 2015. p. 139–144.

VIMAL, B. et al. Mfcc based audio classification using machine learning. In: **2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. [S.l.: s.n.], 2021. p. 1–4.

YING, S.; ZHANG, X. A study of zero-crossings with peak-amplitudes in speech emotion classification. In: **2010 First International Conference on Pervasive Computing, Signal Processing and Applications**. [S.l.: s.n.], 2010. p. 328–331.