EVALUATING PRE-TRAINED MACHINE LEARNING MODELS FOR AMBIENT ACOUSTIC EVENT CLASSIFICATION

ALEXANDRE THUROW BENDER[1]; RICARDO MATSUMURA ARAUJO[2]

[1]*Universidade Federal de Pelotas – atbender@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – ricardo@inf.ufpel.edu.br*

1. INTRODUCTION

In recent years, computational auditory analysis has gained significant attention as a research area (ROSENTHAL, 2021) due to a combination of factors: the production of sophisticated hardware in the form of GPUs (Graphics Processing Units) and tools to support them (BOYER, 2013) allowed notable advances in machine learning techniques and understanding (HOLZINGER, 2018). Combined with the crescent interest, followed by investment from the industry towards data in general (GILCHRIST, 2016), audio classification has become a field of research comprising several sub-areas of effort towards specific problems.

Acoustic Event Detection/Classification (AED/C) is a research sub-field stemming from computational auditory scene analysis (WANG, 2006) concerned with, in addition to processing acoustic signals, converting them into symbolic descriptors that match the perception of a human listener.

Even though speech is certainly the most important acoustic event information-wise, auditory analysis is in no way limited to speech detection. As a matter of fact, human social activity presents itself in a rich variety of acoustic events (TEMKO, 2006). Ambient sound events offer valuable insight when identified, for example, they can be used to improve the robustness of speech recognition systems by filtering non-speech signals. The Amazon Alexa home assistant technology can detect fire alarms and glass windows breaking and notify the user of a potentially dangerous situation that would otherwise go unnoticed.

Sound classification is, therefore, an interesting medium to better understand the human social activity. Its applications include music classification for recommendation systems, monitoring children and the elderly, automated surveillance systems, advanced multimedia retrieval systems, speech recognition, accessibility tools for the hearing impaired, and many others. There may also be relevant medical uses, particularly for the early diagnosis of certain conditions such as heart and lung disorders (YANG, 2019).

The current research originates in the context of a bigger effort, the IDEA-TECH Project (Identifying Depression Early in Adolescence), led by professor Dr. Christian Kieling from the Federal University of Rio Grande do Sul (UFRGS). Incorporating a global interdisciplinary consortium, the project aims to test the predictive utility of a risk calculator for future depression using information from the daily experience of adolescent volunteers. Said experience is registered passively through their smartphones in audio clips. These automatic recordings are expected to capture ambient audio samples depicting the environment of the volunteers as they go about their daily routine. One of the challenges in the project is automating the analysis of the audio clips without actually listening to them since they hold private information.

The current work proposes to identify and evaluate different machine learning approaches for acoustic event classification to validate an eventual use of these

technologies to represent audios semantically with symbolic descriptors. This can be understood as the task of extracting audio elements in the format of interpretable classes contained in the audios. Such symbolic audio representations are important not only for extracting information from audio clips but because they potentially allow for sensitive data analysis with an impersonal approach.

The IDEA-Tech Project captures ambient audio data from adolescents to verify whether they can be used for depression risk assessment. This study contributes by examining acoustic event classification methods to represent audio clips as classes in the hopes of eventually studying the relationship between these classes and the risk factor.

2. METHODOLOGY

In this work, we hypothesize current acoustic event classification methods are useful for audio classification. This entails converting the auditory signal into a semantic meaningful space of characteristics, functionally describing the contents present throughout the audio clip.

Our general objective in this study is to survey and map acoustic event classification methods, as well as evaluate the current relevant machine learning models for the task of audio classification. It is important to note the audio clips in this study are recorded passively and contain a collection of ambient sounds not well known in advance. This analysis explores the viability of employing this technology in automated pipelines for audio processing while studying the different technical approaches for tackling this problem.

It is expected this work yields a comprehensive collection of machine learning methods for acoustic event classification, including robust metrics depicting their applicability in symbolically extracting audio contents. Such ranking is a helpful guideline for deciding which approach is better equipped for this task in real-world scenarios.

To achieve our goals, we need to compile the current relevant machine learning pre-trained models in the context of acoustic event classification. Therefore, a significant portion of this work is to map the computational auditory analysis literature and select a subset of methods to review.

Not every acoustic event classification pre-trained model will translate well to ambient data collected by the IDEA-Tech project. One of the challenges is the fact ambient data is polyphonic (contains audio from several sources). Note this entails a multi-label problem, multiple non-exclusive labels. Furthermore, each pre-trained model will have classes of its own, according to the task it was trained to do. Investigating such classes is important to evaluate their intersection with our problem.

Ultimately, acoustic event classification can be understood as feature extraction, followed by a classification process. Usually, the most popular features include the frequency-domain, time-domain, and Mel-Frequency Coefficients. For representative comparison results, the feature extraction process should match the features used in each pre-trained model. This is challenging on its own, as each model added for analysis implies an additional layer of complexity in the form of its feature extraction process. Another concern is whether the models are available for public use. Thus, it is expected some pre-trained models with potential in ambient sound classification to be out of reach for the present analysis.

For this study, we possess an annotated dataset representing the target task, which originated from the IDEA-Tech project. However, there is a lack of data in this format, and validating the results on such a small dataset may yield non-representative results. The literature provides bigger curated datasets which might be necessary for a robust analysis, complementing the IDEA-Tech dataset.

3. RESULTS AND DISCUSSION

Preliminary results (seen in Table 1) take into account a subset of auditory scene analysis, which is the task of voice activity detection in the following models: YouTube-8M, GPV-F, and GPV-B (CHEN, 2020). All models were pre-trained on Audioset (GEMMEKE, 2017), contemplating annotated audio segments of YouTube videos. Audioset is, to our present knowledge, the most comprehensive audio dataset for the task of acoustic event classification, therefore suitable for our purposes.

Table 1: Experiment results.

Model	Accuracy	Precision	Recall	F1-Score
YouTube-8M	0.62	0.61	0.57	0.56
GPV-F	**0.78**	**0.77**	**0.77**	**0.77**
GPV-B	0.63	0.72	0.56	0.49

YouTube-8M was, in fact, designed for video classification but accepts audio inputs as well. The model was trained using a traditional frame-level approach. Its performance could suggest it does not translate well into the domain of our task. This alludes to a larger, generalized problem in machine learning called domain shift. The problem of domain shift can be understood as a mismatch in the distribution of data used to pre-train a model and data in its actual real-world application. In computation listening, this is often caused by differences in recording devices and overall acoustic conditions.

GPV-F and GPV-B were trained in a weak-supervised fashion, with clip-level annotations. This approach comes with a few advantages, one of them which addresses domain-shift. Speech in the wild often comes with significant amounts of unpredictable noise, making frame-level training non-representative of real-world scenarios. By training models with clip-level annotations, researchers can train models with noisy data closer to practical scenarios. This is unrealistic to accomplish using frame-level labels, as annotating data obtained in non-controlled environments is costly. Such an approach can be understood as a form of domain-adaptation, as the model is trained on data closer to its target domain.

The main difference between GPV-F and GPV-B is that the latter was trained only on detecting speech, while the other was trained on detecting several additional classes. For this reason, it is expected that GPV-F outperforms GPV-B in speech detection, as the former actively models the other classes instead of clustering all non-speech classes together.

4. CONCLUSION

In this work, we discussed the relevance and potential of acoustic event detection in assessing the risk factor of clinical depression in individuals. We addressed some of the key challenges in the area, including model evaluation, label mapping, and domain adaptation. Furthermore, we have shown preliminary experiment results, achieving 0.77 F1-score using the GPV-F model, suggesting clip-level annotations are better suited for real-world applications in machine listening.

We hope this study promotes a better understanding of machine learning models for auditory analysis and their potential for acoustic event classification in the wild. This is relevant for better handling noisy data and automating pipelines.

Future studies may focus on improving the results with several methods. One such approach consists of using multiple models to form an ensemble. Other possibilities include parameter tuning or training the model on a small subset of the IDEA-Tech data.

5. REFERENCES

BOYER, V.; EL BAZ, D. Recent advances on GPU computing in operations research. **IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum**, p.1778–1787, 2013.

CHEN, Y.; DINKEL, H.; WU, M.; YU, K. Voice Activity Detection in the Wild via Weakly Supervised Sound Event Detection. **Interspeech**, p.3665–3669, 2020.

GEMMEKE, J. F. et al. Audio set: An ontology and human-labeled dataset for audio events. **IEEE International Conference On Acoustics, Speech And Signal Processing** (ICASSP), 2017. p.776–780, 2017.

GILCHRIST, A. Introducing Industry 4.0. **Industry 4.0.** Springer, 2016. p.195–215.

HOLZINGER, A.; KIESEBERG, P.; WEIPPL, E.; TJOA, A. M. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. **International Cross-Domain Conference for Machine Learning and Knowledge Extraction**, p.1–8, 2018.

ROSENTHAL, D. F.; OKUNO, H. G.; OKUNO, H.; ROSENTHAL, D. Computational Auditory Scene Analysis: Proceedings of the Ijcai-95 Workshop. **CRC press**, 2021.

TEMKO, A. et al. CLEAR evaluation of acoustic event detection and classification systems. **International Evaluation Workshop on Classification of Events, Activities And Relationships**, p.311–322, 2006.

WANG, D.; BROWN, G. J. Computational auditory scene analysis: Principles, algorithms, and applications. **Wiley-IEEE press**, 2006.

YANG, R.-Y.; RAI, R. Machine auscultation: enabling machine diagnostics using convolutional neural networks and large-scale machine audio data. **Advances in Manufacturing**, v.7, n.2, p.174–187, 2019.