Exposé for research
Audio speech data preprocessing using an AUX-VAE for downstream applications

Hussam Almotlak[1]

[1]Department of Informatics – University of Hamburg
Hamburg, Germany

8almotla@infromatik.uni-hamburg.de

Abstract. *Unsupervised learning is one of the main three categories of machine learning along with supervised learning and reinforcement learning. It is based on the idea of self-organization to find hidden patterns and features in the data without the need for labels. This thesis aims to apply unsupervised learning methods on speech audio data to extract a low dimensional representation for other tasks, that require one of the supervised learning approaches.*

1. Introduction

Supervised machine learning Methods require high amounts of labeled data, but data labeling is a very time-consuming process. For this reason, many developers tend to apply available unsupervised learning methods or to preprocess the data in an unsupervised manner to extract a low dimensional representation and then apply the supervised training on the new representation with fewer labeled data[**3**][**4**].

The process of reducing the number of dimensions (random variable) is known in many fields like statistics and machine learning as dimensionality reduction. It can also be considered as a preprocessing method for many other machine learning applications (classification and regression). Principle component analysis (PCA) is one of the first mathematically successful methods of dimensionality reduction, but it can't be applied on very high dimensional data. Autoencoders have shown a great job of capturing the most important features in the data. They are also, with their non-linear behavior, preferred over the other methods like PCA when the data is very high dimensional.

Lately, a lot of effort has been made on the processing of static data like images and clearly less on the processing of sequential data like audios and videos[**3**]. Sequential data normally holds instantaneous as well as long-term information. The long-term information describes certain characteristics in the sound (the voice), which don't change with time such as health status, age, and emotion...etc.

2. Related work

The authors in [**3**] applied an architecture based on a variational autoencoder to process the audio data. They used only one latent variable and a small dataset consisting of only 123 utterances in Spanish. According to the authors, their model did deliver slightly better results than the RMB (Restricted Boltzmann Machine), but not good enough for a downstream task. Later in 2019, The authors in [**3**] developed a model that distinguishes global-timescale from local-timescale speech characteristics in audio data. They extended the variational autoencoder with an auxiliary variable h for capturing the global-time scale features. This created an architecture with four probability density distributions. The new architecture is shown in (Figure 1, taken from the original paper) and consists of the following subnetworks.

1. global-timescale network: takes as input the preprocessed speech audio data and outputs a global latent variable h, which represents information extracted from the entire input.
2. local-timescale network: takes as input a concatenation of the same input to the global-timescale network and the output of the global-timescale network. Its output is the local latent variable z. These two networks form the encoder.
3. The decoder network: takes the local latent variable to predict the next speech frame.
4. The predictor network: process the local latent variable z to extract information from the global-timescale latent variable h hidden inside the local variable z. its output is the global-timescale variable h.

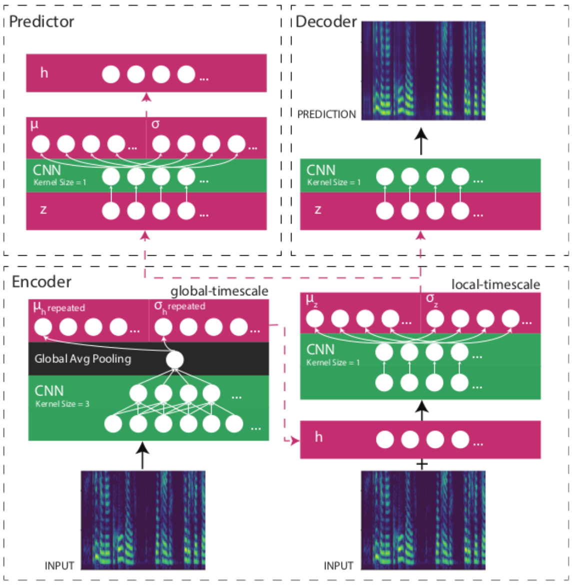

Figure 1. Figure 1: shows the entire architecture of the AUX-VAE introduced by [3] including the four networks

The model has shown great results of recognizing the gender of the speaker as well as making speaker identification. There is also more related work, which could be referred to later in the final thesis.

3. Research questions and objectives

The objective of the research is to preprocess speech audio data with an auxiliary autoencoder before feeding it to supervised learning applications like the classification of emotions in audio speech signals. Therefore the question to be answered in this work is whether it is possible to extract a sufficient low dimensional representation from the sequential audio data with an AUX-VAE for the next primary supervised task or not?

4. Methodology

4.1. Preprocessing

This part is still open for experimentation. The main goal here is to achieve an effective spectogram from the audios. Operations that can be used are short-time Fourier transform, Mel filter bank, and discrete cosine transform.

4.2. Autoencoders

Autoencoders are one of the advanced approaches of unsupervised learning. They are deep neural networks, that try to find a low dimensional representation (latent representation) for complex high dimensional data like images, audios, and videos learning by backpropagation throw reconstructing the same input from the latent representation. According to[**5**], autoencoders tries to find the weight matrixes E, D that minimize the mean square error:

$$E, D = argmin(X - (DoE)X)^2$$

arg min indicates the arguments (the two weight-matrixes) that correspond to the minimum square error. (DoE)X: is a symbol for D(E(X)).

4.3. Variational Autoencoders

Many variations of autoencoders have recently been introduced such as sparse autoencoders, denoising autoencoders, and contractive autoencoders... etc. But there are some particularly unique versions of autoencoders, the variational autoencoders. They are generative models, that apply probability modeling into neural networks to regularize the latent space, which gives the autoencoders a generative feature. In this work, I am planning to use a productive auxiliary variational autoencoder like the one introduced by[**3**] to get a low dimensional representation of the voice-data. After that, I am going to build a classifier on top of the representation to classify the emotion of the speaker. An appropriate dataset for the task could be the OMG emotions dataset introduced by the University of Hamburg.

4.4. OMG emotion dataset

This dataset consists of about 8 hours of emotional monolog youtube videos. The videos were separated based on utterances and labeled separately by humans. The labels are the arousal and the valence of clip (utterance) as well as the overall emotion, which will be mainly used in the end to display the representation of the autoencoder as well as for training the classifier. The used emotions are anger, disgust, fear, happy, neutral, sad and surprise [**2**].

5. Time plan

Starting on the tenth of November, there would be two weeks of research for Literature and relevant papers. Thereafter one or two weeks for collecting and preprocessing more speech audio data that contains emotions for the unsupervised learning process. Once the dataset is ready, there would be around three months to try different architectures of AUX-VAEs like the one introduced by [**3**] and trying also different classifiers upon the AUX-VAE. The rest of the time which is about 5-6 weeks could be dedicated to the writing of the thesis.

References

[1] Blaauw, Merlijn and Bonada, Jordi. *Modeling and transforming speech using variational autoencoders.*. 2016

[2] Barros, Pablo and Churamani, Nikhil and Lakomkin, Egor and Sequeira, Henrique and Sutherland, Alexander and Wermter, Stefan *The OMG-Emotion Behavior Dataset*. 2018 International Joint Conference on Neural Networks (IJCNN) 2018

[3] Springenberg, Sebastian and Lakomkin, Egor and Weber, Cornelius and Wermter, Stefan. *Predictive Auxiliary Variational Autoencoder for Representation Learning of Global Speech Characteristics.* Proc. Interspeech 2019. Pages934–938.

[4] Hsu, Wei-Ning and Zhang, Yu and Glass, James. *Learning latent representations for speech generation and transformation* 2017

[5] Wikipedia: Autoencoders,
`https://en.wikipedia.org/wiki/Autoencoder`