Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych

Abstract

The effectiveness of audio signal processing systems depends on several factors, including the choice of neural network architecture, the training data, and their initial representation. The trained neural networks of appropriate architectures (e.g., convolutional networks or autoencoders) can act as additional feature detectors, complementing the initial data representation and thus improving their discriminative capabilities. This can also improve the performance of the systems that directly use data in the time domain form (end-to-end learning). Additionally, enhancing the initial training data with additional contextual data of a different modality further improves the effectiveness of these systems. Therefore, proper data modelling and representation are crucial for audio signal processing. This dissertation is based on a series of publications focusing on methods for enhancing the effectiveness of audio signal processing systems using the abovementioned approaches.

In this dissertation, we hypothesise that including sound events in speech corpora, along with reverberation and background sounds, enhances the robustness of speech processing models to such disturbances. We also demonstrate that using various lengths of time frames of the Scattering Wavelet Transform (SWT) with Convolutional Neural Networks (CNN) helps improve the recognition of early symptoms of heart diseases. Finally, we validate the hypothesis that using the extended version of the standard Big Five model for representing user personality helps reduce the error made by music recommendation systems compared to the standard model.

The presence of unexpected sound events in audio recordings containing speech can disrupt intelligibility and cause the recording to be excluded from further processing due to its low quality. In the paper "Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions", we emphasised the significance of the lack of a suitably diversified training set that reflects real acoustic conditions for training

systems aimed at improving speech intelligibility. The paper proposes a solution that generates short audio recordings containing Polish speech against background sounds, with reverberation and unexpected sound events. Our experiments with this data and training deep neural networks confirm their usefulness and support the hypothesis regarding improving the models' robustness to the occurrence of such additional sound events.

In the next paper, "Beyond the Big Five Personality Traits for Music Recommendation Systems", we proposed enriching the input data for music recommendation systems with the additional personality factors of listeners. These factors are represented by an extended Big Five model, which measures three additional aspects of each of the five main personality traits. Incorporating this extended model allowed us to reduce the recommendation error. The paper "Music Recommendation Systems: a Survey" discusses the advancements in music recommendation systems, including those involving deep neural networks. It also presents the challenges of personalising music recommendation systems and outlines future research directions.

An essential aspect of effective recommendation is the accurate classification of songs by genre or mood, which improves the quality of the music recommendation. The effectiveness of the SWT in representing music data was confirmed in the work titled "Pre-trained Deep Neural Network Using Sparse Autoencoders and Scattering Wavelet Transform for Musical Genre Recognition". The paper utilised the autoencoder architecture for pre-training the final neural network for music genre classification. The experiments highlighted the high efficiency of SWT in representing music data and provided insight into the behaviour of training the networks pre-trained with autoencoders. The results obtained in this work contributed to the subsequent work titled "Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation". This work addressed extracting additional features of the heartbeat signal using SWT and CNN. The convolution operations on several SWT time frames enabled the detection of dependencies between frames, improving the efficiency of recognising the first symptoms of heart diseases from noisy data.

To summarise, we verified all hypotheses of this dissertation.