A Review on Machine Learning for Audio Applications

Nagesh B[1]* andDr. M. Uttara Kumari[2]

[1,2]*R V College of Engineering, Bengaluru, India*
[1]*nageshb.ec17@rvce.edu.in,* [2]*uttarakumari@rvce.edu.in*

Abstract: *Audio processing is an important branch under the signal processing domain. It deals with the manipulation of the audio signals to achieve a task like filtering, data compression, speech processing, noise suppression etc. which improves the quality of the audio signal. For applications such as the natural language processing, speech generation, automatic speech recognition, the conventional algorithms aren't sufficient. There is a need for the machine learning or deep learning algorithms which can be implemented so that the audio signal processing can be achieved with good results and accuracy. In this paper, a review of the various algorithms used by researchers in the past has been described and gives the appropriate algorithm that can be used for the respective applications.*

Keywords: Machine learning, digital signal processor, deep learning, signal processing, neural networks.

1. Introduction

Different sorts of acoustic sounds, such as voice, music, and birdsong, have been retained as audio recordings since Edison introduced the phonograph in 1877 [8]. Essential aspects of numerous auditory signals have been disclosed enabling a detailed understanding of the mechanism for production, propagation, and perception after decades of investigation and analysis [2]. Acoustic signals contain a wealth of information that may be used to determine spatial location, determine the species of a sound source, determine the identity of a speaker, and determine the message being transmitted [9]. Machine learning has lately been used to create solutions for complex tasks such as picture and speech recognition, as well as tactile data processing, with judgement accuracy approaching that of a human [3].Machine learning algorithms have recently shifted their focus to more practical and emergent sectors such as speech/speaker recognition, speech synthesis, and acoustic event/scene recognition. Artificial neural networks have risen to prominence in three waves, driven by the perceptron algorithm, the backpropagation algorithm, and deep learning's success in speech recognition and picture classification, resulting in a deep learning renaissance [1].

In this paper, a review of the various machine learning algorithms that have been used for the audio signal processing applications have been given. Many researchers have found different ways of audio signal processing. Few researchers have also combined the existing models to obtain a hybrid models which would perform better for the required application. The use of neural network models and its variants in [1], [2] for machine learning models. The use of generative adversarial networks for the speech generation is also dealt with in [1] and [2] which will be very helpful in situation where there is scarcity of data. Implementation of the machine learning models on hardware would give better results, but the process of implementation of these models on an embedded system would be a cumbersome task. In [3], this task has been implemented in an elegant way by considering the tensorial kernel approach and the algorithm has been implemented on the Virtex-7 FPGA board. The process of feature selection would be a cumbersome task in the applications such

as audio signal processing. In [4], the constraint compensated Laplacian score has been used as a performance metric in order to extract the features form the audio signals. In machine learning applications, the feature extraction is done in order to reduce the amount of data being trained on and also discard the unwanted signal which may add up the error in the model. This is dealt with in [5], where the feature reduction is done for the automatic genre classification and the resulting model obtained form training with the reduced features is analysed. The machine learning models have to be optimized in order to obtain good results without much delay. This is done in [6] by implementing an optimized audio classification and segmentation algorithm by the use of the ensembled bagged trees. In [7], mixed stereo audio classification is doneusing a stereo-input mixed-to-panned level feature. The new measure called the speech to music ratio is also calculated here. Overall, when it comes to implementation of the machine learning models, there are several areas and aspects which needs to be looked into in order to obtain a good trained and tested model.

2. Review

Convolutional neural networks, versions of the long short-term memory architecture, and more audio-specific neural network models are discussed, as well as the prevalent feature representations. Audio recognition for automatic speech recognition, music information retrieval, environmental sound detection, localization, tracking and synthesis and transformation for source separation, audio enhancement, and generative models for speech, sound, and music synthesis are among the prominent deep learning application areas covered in [1]. The various audio feature extraction methods have been discussed such as the deep neural networks which are used to extract features while also achieving a goal, such as classification, MFCC is employed as the primary acoustic feature representation for audio analysis tasks. [1] also gives an overview of the convolution neural networks, recurrent neural networks, sequence-to-sequence models, generative adversarial networks. The loss functions such as the mean squared error, which can be employed are also discussed in [1]. When applied to huge training datasets, deep learning is known to be the most profitable. The paper also talks about acquiring the required data set, filtering them and preparing it for the use case and using them appropriately to obtain maximum efficiency of the system. For audio signal processing, there is not much data available. Hence the paper suggests to use the data generation and data augmentation. Data that resembles real data, with specified synthesis parameters and labels, can be generated for particular tasks. Understanding, troubleshooting, and enhancing machine learning methods is made easier by a controlled steady increase in the complexity of the generated data. By altering existing samples to span a wider range of possible inputs, data augmentation provides extra training data. Independently proposed pitch shifting and time stretching to modify audio extracts for automatic speech recognition, and simply resampling also helps. Linearly mixing training examples with their labels enhances generalisation for ambient noises. Models for source separation can be successfully trained using datasets created by blending separated sounds. These are some of the techniques that can be used for data generation in [1].

H. Purwins, et.al in [1] have also analysed the various applications of the audio signal processing. They are the speech, music, environmental sound and localization and tracking. The various algorithms that have been used in past till date have been explained in detail. The triphone-state Gaussian mixture model has been the standard for modelling speech for decades. Later, the Gaussian mixture models have been proposed to be replaced with hybrid models based on neural networks. On several speech recognition tasks, deep neural networks with millions of parameters trained on thousands of hours of data were found to drastically

reduce the word error rate. The recent algorithms such as the long short-term memory and gated recurrent unit have been shown to outperform feedforward deep neural networks. The advantages of using the recurrent neural network for the modelling of speech has also been discussed. For music applications, low-level analysis, rhythm analysis, harmonic analysis, high-level analysis and high-level comparison are among the major tasks. Each of these problems was previously addressed using hand-crafted algorithms or features mixed with shallow classifiers, but is now being addressed using deep learning. Environmental sound analysis offers a variety of applications, including context-aware devices, acoustic surveillance, and multimedia indexing and retrieval. In [1] three basic approaches have been given for environmental sound analysis. They are acoustic scene classification, acoustic event detection and tagging. Brief discussion for all the three methods has been given in [1].

Y. Zhao, et. al. in [2] provides an overview of current deep learning algorithms and their utility in acoustic signal processing. Under the categories of discriminative and generative algorithms, thorough investigations of several deep learning architectures are offered, including the most up-to-date Generative Adversarial Networks as an integrated model. A complete review of audio generation applications is presented. The explanation on how deep learning methods can assist the field of speech/acoustic signal synthesis and the potential challenges that need to be solved for future real-world scenarios based on what is learned from these approaches. [2] also given an overview of the perceptron, multi-layer perceptron, convolution neural networks, recurrent neural networks. The application involving the speech recognition has been discussed. Similar to [1], the evolution of the algorithms that have been used for the speech modelling have been discussed in detail. They are the gaussian mixture models, deep neural network models, mixture density networks, long shorth term memory, recurrent neural networks. The above algorithms have been discussed by mentioning the advantages and disadvantages the process of the evolution that have been taken place throughout. Two generative algorithms have been discussed in [2]. They are the variation autoencoder, deep belief network. The variational autoencoder is evolved from the original autoencoder, but with more stringent assumptions on latent variable distribution. An autoencoder is a sort of neural network that learns data coding unsupervised. Because of their ability to represent a set of data, autoencoders are most commonly used for dimensionality reduction. Although basic autoencoders have been successfully used, autoencoder development is hampered by a number of issues. One of the most basic is that the latent space for generation might not be continuous or easily interpolated. A discontinuous section from the latent space makes it harder to develop a generative model and produce variations in the input image. Deep belief networks are a sort of generative graphical model that can be thought of as a stack of basic restricted Boltzmann machines. A DBN has both directed and undirected edges between layers. Hidden units are linked to those from other layers, but not to those from the same layer. It's worth noting that DBNs can be used for supervised learning by adding a final layer of variables that reflect the intended outputs, as well as backpropagating error derivatives. A different approach is to use the weights from a trained DBN to train a neural network for classification jobs.

Y. Zhao, et. al. in [2] have discussed the GANs (generative adversarial networks) extensively. GANs (generative adversarial networks) are a type of unsupervised machine learning neural network. A conventional GAN is made up of two primary components: a generator G and a discriminator D, which are trained in a zero-sum game strategy by competing with each other. A detailed explanation of the working of the GANs have been discussed. Finally, the application of the above-mentioned algorithms has been put forward. They are the image generation and synthesis, audio synthesis, speech synthesis, extract the features of speech and data augmentation of the acoustic signal processing. The most widely used deep generative and discriminative algorithms were reviewed in this study. For readers

in the machine learning community, a detailed statement on the application of deep learning to speech/acoustic signal processing, particularly audio production are presented. In addition, the advantages of deep learning approaches are examined in depth, as well as a number of research concerns that need to be addressed.

A.Ibrahim, et.al in [3] have presented hardware architectures and implementation of a real-time machine learning method based on a tensorial kernel approach for multidimensional input tensors in this study. The two distinct hardware architectures are proposed and discussed. The first technique has been implemented on hardware with multidimensional tensorial inputs. The proposed implementation may be used to expand kernel-based algorithms such as SVM and K-ELM to multidimensional tensors while keeping the original data's underlying structure. For the tensorial kernel function, it provides two distinct implementations based on cascaded and parallel architectures. For a case study of classification of three input touch modalities providing real-time functionality for the target application, the implementations are investigated. There is a comparison of the proposed architectures. When different numbers of training tensors are studied, it reveals that parallel design outperforms cascaded architecture by reducing the percentage occupied area. The study also in [3] indicates that the constructed system is capable of real-time classification. For the Virtex-7 XC7VX980T FPGA device, the suggested parallel architecture achieves a peak performance of 302 G-ops while requiring only 1.14W, outperforming some state-of-the-art alternatives. The tensorial approach has been proposed in this paper as the experiments in many applications strongly advocate tensorial representation of data since it keeps the original data's underlying structure. The data is acquired in tensorial representation, with the 2-D sensor array representing the first two dimensions and time representing the third dimension of the tensor. Detailed explanation is given for the tensor unfolding, singular value decomposition, kernel computation and use of the support vector machine for classification. The paper also talks about the computational complexity of the above steps. The computational complexity of an algorithm is defined as the number of operations it must do.

A. Ibrahim et. al in [3] have also written a MATLAB script to calculate the number of operations required based on the tensor dimension. The work done has used the one-sided Jacobi algorithm. The implementation done is as follows: iteratively, the computations are organised. Each iteration includes n(n-1)/2 transformations that obliterate the off-diagonal elements. Matrix Symmetrization is the next step where the one-sided Jacobi algorithm requires symmetric square matrices as inputs. Matrix symmetrization transforms arbitrary rectangular matrices into symmetric square matrices. The phase solver’s implementation on hardware which executes the one-sided jacobi algorithm is explained. The Pre-rotation block is in charge of employing the one-sided Jacobi block to rotate the rows. The columns are rotated using the one-sided Jacobi block in the post-rotation phase, but this time only the columns of the matrices are affected. The applied rotations are based on the cosine and sine functions that the phase solver block has already calculated. The hardware implementation was written in RTL VHDL and synthesised with the 14.7 edition of the Xilinx ISE tool. A Virtex-7 XC7VX980T device was used to synthesise the various blocks.Kernel approaches' solid mathematical structure piqued their interest in a variety of applications. The hardware implementation, on the other hand, is primarily focused on linear tensor kernels. The hardware architectures and implementations of kernel functions for multidimensional tensors are discussed in this work. For the presented technique, the paper proposed and evaluated two hardware architectures. The results show that the proposed implementations for real-time classification are feasible, with a peak performance of 302 G-ops and a power consumption of 1.14W, outperforming state-of-the-art systems. Although the results are promising and demonstrate good real-time functioning, the power consumption is still too high for the target application, where the power budget is confined by the available battery's size and weight.

Yang, et.al in [4] gives an important front-end task in speech signal processing is audio classification, which involves categorising audio segments into broad categories such as speech, non-speech, and silence. For audio classification, many of features have been proposed. Unfortunately, these characteristics are not mutually exclusive, thus combining them does not increase classification accuracy. Feature selection is a powerful tool for determining which features are most useful and least redundant for categorization. This study proposes the Constraint Compensated Laplacian Score (CCLS), a semi-supervised feature selection approach that takes advantage of the local geometrical structure of unlabelled data as well as constraint information from labelled data. This strategy is tested against other known feature selection methods on an audio classification problem. The findings of the experiments show that CCLS improves the situation significantly. Splitting an audio stream into homogeneous content chunks is known as audio segmentation. The process of segmentation involves joint boundary detection and classification, resulting in the identification of segment regions as well as the classification of those regions, given a preset set of audio classes. The feature selection methods such as the Laplacian score, constraint score, constrained laplacian score have been discussed in great detail. The short coming of the constrained Laplacian score is put forward in this paper. The work done also provides the comparison of the models implementing the Laplacian score, constraint score and the constrained Laplacian score.

A semi-supervised filter-based feature selection method was introduced in [4] using the word done. The new CCLS approach combines unlabeled data location preservation with label consistency inside labelled data. For audio categorization, the suggested method outperforms Spec, ReliefF, LS, and CS, according to experimental data. In terms of the optimal number of features, CCLS did not perform as well as RelifF. This could imply that the optimum feature set selected by the CCLS approach contains some redundancy features. Several more research have looked into the effects of redundancy.

Multimedia database retrieval is becoming more popular, and this is reflected in the popularity of online retrieval systems. Searching, retrieving, and organising music information can be difficult due to large datasets. As a result, for arranging this music data into separate classifications based on specific viable information, a strong computerised music-genre categorization approach is required. For genre categorization, two key elements must be considered: audio feature extraction and classifier design. B. K. Baniya, et.al in [5] to correctly characterise the music content, have presented a variety of auditory features. Dynamic, rhythmic, spectral, and harmonic feature sets are divided into four categories. Five statistical parameters are chosen as representations from the features, including the fourth-order central moments of each feature and covariance components. In the end, lowest redundancy and maximum relevance are used to control insignificant representative parameters. This algorithm ranks all feature qualities based on their score levels. For genre classification, only high-score audio elements are taken into account. Furthermore, the audio attributes are used to determine which of the various statistical factors produced from them is most essential for genre classification. The statistical characteristics of the mel frequency cepstral coefficient, such as covariance components and variance, are chosen more frequently than the feature attributes of other groups. Unlike principal component analysis and linear discriminant analysis, this method does not change the original features. Other feature reduction techniques, such as locality-preserving projection and nonnegative matrix factorization, are also taken into account. The proposed system's performance is evaluated using reduced features from the feature pool obtained through various feature reduction approaches. The findings show that the overall classification is on par with current state-of-the-art frame-based techniques.

In [5], different auditory parameters (dynamic, rhythmic, spectral, and harmonic) were chosen for music genre classification in this work. Lower mean and standard deviation, as well as higher-order moments like skewness and kurtosis, were used to combine these traits. To improve the classification, covariance components were calculated as well. The trials were carried out in two stages, based on the feature statistics acquired. The purpose was to see how much classification accuracy improved when extra parameters (higher-order statistics and covariance components) were added to the lower-order statistics with decreased features. Lower-order statistics were evaluated for classification in the first stage. After that, all statistics (lower, higher, and covariance components) were added together to establish overall genre categorization results. The direct result of taking into account all statistics was a rapid increase in feature dimensions. As a result, multiple feature reduction approaches, such as Principal component analysis (PCA), Linear discriminant analysis (LDA), Locality preserving projection (LPP), Non-negative matrix factorization (NMF), and Minimum redundancy maximum relevance (MRMR), were used to regulate them. Using PCA, LDA, LPP, NMF, and MRMR reduction feature sets, the classification accuracies of the lower-order statistics were 75.0, 73.50, 77.25, 79.80, and 80.75 percent, respectively. Similarly, when all feature data were considered, 63 percent of features were unimportant. Using MRMR, LPP, LDA, PCA, and MNF feature reduction approaches, overall classification accuracies of up to 87.9, 85.20, 84.50, 78.2, and 75.20 were obtained from the remaining feature data. The performance of the MRMR-based feature reduction approach employing SVM was comparable to that of other modern methodologies. In addition, the MRMR algorithm selects other features that have the greatest relevance among all of them. As a result, it is clear which traits, as well as the statistics associated with them, had a substantial impact on genre classification in MRMR.

Audio segmentation serves as a foundation for multimedia content analysis, which is today's most essential and extensively utilised application. Saadia Zahid, et.al in [6] presents an efficient audio classification and segmentation method that divides an overlaid audio stream into four primary audio kinds based on its content: pure-speech, music, environment sound, and quiet. An algorithm is proposed that preserves key audio material while lowering the rate of misclassification without requiring a huge amount of training data, is noise-tolerant, and may be used in real-time applications. Noise is separated out of an audio stream as ambient sound. Bagged support vector machines and artificial neural networks are employed in a hybrid classification technique. The audio stream is first divided into speech and non-speech segments using bagged support vector machines; the non-speech segment is then further divided into music and environmental sound using artificial neural networks; and finally, the speech segment is divided into silence and pure-speech segments using a rule-based classifier. The training classifier is trained with the bare minimum of data; ensemble methods are utilised to reduce the misclassification rate, and about 98 percent accurate segments are obtained. A fast and efficient algorithm that can be employed with real-time multimedia applications has been developed.

Saadia Zahid, et.al in [6] portraits the importance of the Audio Classification and Segmentation Step to classify an audio clip into basic data kinds, a hybrid classification approach is given. A pre-classification step is performed before classification, which analyses each windowed frame of the audio sample separately. The feature extraction stage is then completed, yielding a normalised feature vector. The voice signal is overlay during the pre-classification process, indicating that a discussion is taking place in any location or party where there is music and a lot of noise. Blind source separation is the process of separating the source or desired segments inside an independent component analysis framework. By applying the Fourier transform at short time intervals, the mixed signal is first transformed to the time frequency domain, commonly known as the spectrogram of signal. The Hamming

window is employed, and all of these brief intervals are correlated. The next step is the feature extraction step where the features are extracted from the audio signal. The feature vectors contain information about the audio signal's temporal and spectral characteristics. The feature vectors are computed on a window-by-window basis. The features that are chosen have a big impact on how well audio segmentation systems work. The performance measures such as the Zero-Crossing Rate (ZCR), Short-Time Energy (STE), Mel-Frequency Cepstral Coefficients (MFCCs), Periodicity Analysis have been explained in great detail. An hybrid approach is proposed where the bagged support vector machine is combined with artificial neural network. Most of the environment data is misclassified as speech and music when only one classifier is used, which is not a smart method. To avoid misclassification, a hybrid technique is utilised, in which speech and nonspeech are separated first, and then nonspeech is separated further into music and environmental sound. There has been discussion of an efficient and rapid audio classification and segmentation strategy that does not require a significant quantity of training data but produces good discrimination results. An audio stream is divided into homogeneous sections and categorised into basic audio kinds such pure voice, music, environmental sound, and quiet in this study. The main goal is to create an audio segmentation algorithm that may be used in multimedia content analysis and audio recognition applications. For audio classification and segmentation, a hybrid technique was adopted. Using a bagged SVM classifier, audio samples are first separated into speech and nonspeech parts. ANN classifier is used to further categorise non-speech segments into environment sound and music. Using a rule-based classifier, speech segments are separated into silence and pure-speech segments. The technique has been shown to be particularly efficient for real-time multimedia applications in tests.

Due to the potential uses for broadcast and other media, many previous studies on speech/music discrimination have been undertaken; however, it is still conceivable to broaden the experimental scope to include samples of speech with varied quantities of background music. A. Chen et.al in [7] have discussed the development and evaluation of two measurements of the ratio between speech and music energy are the subject of this paper: a feature called the stereo-input mix-to-peripheral level feature (SIMPL), which is computed from the stereo mixed signal as an approximate estimate of SMR, and a reference measure called speech-to-music ratio (SMR), which is known objectively only prior to mixing. SIMPL is an objective signal measure determined using broadcast mixing procedures, in which vocals, unlike most instruments, are often positioned at stereo centre. SMR, on the other hand, is a hidden variable defined by the power of parts of audio assigned to speech and music. SIMPL has been demonstrated to be predictive of SMR and can be paired with cutting-edge characteristics to boost performance. This novel metric is used in speech/music (binary) classification, speech/music/mixed (trinary) classification, and a new speech-to-music ratio estimation issue for assessment. The proposed algorithms include the SMR and the SIMPL. The machine learning algorithms can be broadly classified into two different algorithms. They are the regression and the classification models. The SMR has been dealt with in both the models i.e., regression and classification. Speech and music can be separated from mixed audio signals from broadcast and other media sources. The estimation of an audio file's intrinsic SMR can be represented as a regression issue based on the concept of speech-to-music ratio. It has been shown that by quantizing the output and training a polychotomous classifier instead of real-valued regression models, real-valued regression models may frequently be developed more precisely in a variety of other machine learning applications. The capacity to quickly scale down to other types of categorization issues is a major advantage of this style of solution. Recording engineers have employed the inversion or phase cancellation procedure to make rough instrumental renditions of songs from mainly un-editable formats. In the past, music information retrieval and signal processing studies have

recorded and used this approach. Audio discrimination and SMR evaluation can be conducted using a variety of existing machine learning methods. Gaussian mixture models (GMM) are used in this work because they perform similarly to the best reported methods on similar regression and classification tasks and can be implemented using ordinary voice recognition software. The research carried out by A. Chen et.al in [7], established a simple audio categorization feature that approximates the ratio of energy present in recorded voice or speaking and instrumental portions based on their normal stereo mix placements. SIMPL demonstrated to be a valuable feature in speech/music discrimination applications, with 81.9 percent success rates for three-way classification and 97.9% for two-way classification. With the addition of many spectral features, three-way classification was improved to 92.6 percent accuracy, which compares favourably to previous studies. SIMPL was also tested in a novel application using mixed audio. Thousands of mixed audio signals were created by varying the amounts of pure speech and pure music files. Through a classification framework, machine learning algorithms were trained to automatically estimate the intrinsic speech-to-music ratio of these clips. When charting the data, a clear linear relationship between the real and predicted SMR was visible, indicating the potential of SMR estimation as a component algorithm in systems meant to improve broadcast media consumption and automatic speech recognition. SMR estimate using the SIMPL feature surpassed the best available methods that did not include SIMPL.

3. Conclusion

Various machine learning algorithms have been implemented by researchers. The models implemented also depend on the type of application being used for. When it comes to the process of obtaining a machine learning algorithm, the very first step is to identify the type of application for which the model is being used. The next would be to obtain the data for training and testing the model. The data has to be good enough and not have any randomization in it as it would fail to train the model accurately. In cases where there is scarcity of data, algorithms have to be used which generates data. For audio signal processing application, the Generative Adversarial Networks along with other algorithms have been used by researchers. The next step would be the feature extraction. Few researchers have done this manually and few other have written an automatic feature selection algorithm. Laplacian score, constraint score, constraintLaplacian score have been used as a semi-supervised feature selection algorithm. Few other researchers have used the statical characters such as the Mel frequency, cepstral coefficient, covariance components and variance of the audio signal. Feature reduction algorithms such as the principal component analysis, linear discriminant analysis, locality preserving projection, non-negative matrix factorization and minimum redundancy and maximum relevance algorithms have been used. For designing the model, the convolutional neural networks, recurrent neural networks, deep neural networks, deep belief networks, variational auto encoder, generative adversarial networks, Wasserstein generative adversarial networks have been used. The available data has to be split into the training and test data. An important partition could be the cross-validation data set where this data set would be used in between the training and the testing phase. This reduces the interdependencies of the features on the training parameters. A few other researchers have suggested the use of the tensorial approach as the data can be easily represented in the form of a 2-D matrix and the third dimension would be the time. The data handling can be done better and the algorithms can be easily implemented when the data is in the tensorial format. Finally, the trained model can be implemented on the embedded platforms. This would be a tricky and cumbersome task as the algorithm should support the target device and there are not many existing libraries which supports the implementation of the machine learning

algorithms on hardware. The implementation of the machine learning algorithms on the specific hardware devices such as the digital signal processor for audio processing applications would give good results as these processors have been designed for the application involving the audio signal processing. The future work could be to develop a platform where the implementation of the machine learning models on dedicated hardware can be achieved easily.

4. References

[1] *H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang and T. Sainath, "Deep Learning for Audio Signal Processing," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.*

[2] *Y. Zhao, X. Xia and R. Togneri, "Applications of Deep Learning to Audio Generation," in IEEE Circuits and Systems Magazine, vol. 19, no. 4, pp. 19-38, Fourthquarter 2019, doi: 10.1109/MCAS.2019.2945210.*

[3] *A. Ibrahim and M. Valle, "Real-Time Embedded Machine Learning for Tensorial Tactile Data Processing," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 11, pp. 3897-3906, Nov. 2018, doi: 10.1109/TCSI.2018.2852260.*

[4] *Yang, XK., He, L., Qu, D. et al. Semi-supervised feature selection for audio classification based on constraint compensated Laplacian score. J AUDIO SPEECH MUSIC PROC. 2016, 9 (2016). https://doi.org/10.1186/s13636-016-0086-9*

[5] *B. K. Baniya, J. Lee and Z. Li, "Audio feature reduction and analysis for automatic music genre classification," 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2014, pp. 457-462, doi: 10.1109/SMC.2014.6973950.*

[6] *Saadia Zahid, Fawad Hussain, Muhammad Rashid, Muhammad Haroon Yousaf, Hafiz Adnan Habib, "Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods", Mathematical Problems in Engineering, vol. 2015, Article ID 209814, 11 pages, 2015. https://doi.org/10.1155/2015/209814*

[7] *A. Chen and M. A. Hasegawa-Johnson, "Mixed Stereo Audio Classification Using a Stereo-Input Mixed-to-Panned Level Feature," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 2025-2033, Dec. 2014, doi: 10.1109/TASLP.2014.2359628.*

[8] *T. A. Edison, "The phonograph and its future," North Amer. Rev., vol.126, no. 262, pp. 527–536, 1878.*

[9] *D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Trans. Multimedia, vol. 17, no. 10, pp. 1733–1746, 2015.*