Review Techniques for Vocal Classification using Machine Learning

Srinivas Prajwal BR, Tejas Ganesh Joshi, Surbhi Agrawal

[1]Student at RVITM ,[2]Student at RVITM,[3]Associate Professor at RVITM, B.E, M.Tech ,PhD

[1]Computer Science and Engineering,

[1]RV Institute of Technology and Management, Bengaluru,India

Abstract - This review paper provides a comprehensive overview of vocal classification using machine learning techniques. This paper analyzes and compares different approaches for vocal classification, including traditional and deep learning techniques. This paper discusses various datasets used in vocal classification research and their properties. Furthermore, the paper provides a comparative analysis of existing studies in this field. Finally, the paper highlights the future scope of vocal classification using machine learning techniques and suggests further research opportunities in this area. Overall, this review paper aims to provide a valuable resource for researchers and practitioners working with vocal classification..

Index Terms - Machine learning (ML), Vocal classification (VC), Audio analysis (AA), Feature extraction (FE), Signal processing (SP), Pattern recognition (PR), Deep learning (DL), Convolutional neural networks (CNN), Support vector machines (SVM), Random forests (RF), K-nearest neighbors (KNN), Gaussian mixture models (GMM), Performance evaluation (PE), Dataset (DS), Preprocessing (PP), Feature selection (FS), Spectrogram (SG), Mel-frequency cepstral coefficients (MFCC), Pitch detection (PD), Formant estimation (FE), Voice quality (VQ).

I. INTRODUCTION

Machine learning vocal classification is a fast expanding area that has applications in speech recognition, speaker identification, singer voice classification, and music genre classification, among others. Vocal categorization systems have gotten increasingly precise and efficient in recent years, thanks to advances in machine learning techniques and the availability of enormous datasets. This review article intends to offer a detailed overview of existing machine learning approaches for speech categorization and to assess their performance and applicability for various applications. The research begins by providing an overview of machine learning-based techniques for vocal classification. This includes discussing aspects like extracting features, choosing models, and evaluating performance measures. It then goes on to cover similar work in this subject, highlighting the contributions of several academics and their methods to vocal categorization. The review paper focuses on three specific papers: **"Voice identification using classification algorithms" by Orken Mamyrbayev [1]**, **"Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music" by Toan Pham Van [2]**, and **"Comparative study of singing voice detection based on deep neural networks and ensemble learning" by Shingchern D. You [3]**. Afterwards, the research examines the methodology utilized in these papers. This involves discussing the datasets utilized for both training and testing purposes, the machine learning algorithms employed for extracting features and performing classification, and the assessment metrics utilized to evaluate the performance of the systems. The analysis analyzes the merits and limitations of each approach, emphasizing their applicability for various applications and datasets. Lastly, the study finishes with a discussion of the future scope of voice categorization utilizing machine learning, as well as potential research and development topics. This review study intends to be a beneficial resource for voice categorization researchers and practitioners, allowing them to better grasp existing approaches and highlight prospects for future innovation.

II. VOICE CLASSIFICATION

The technique of finding diverse vocal qualities and sorting them into multiple groups using machine learning algorithms is referred to as vocal classification using machine learning. The primary purpose of voice classification is to extract useful information from an audio signal and categorize it based on certain criteria such as pitch, timbre, and strength. Feature extraction is a crucial step in machine learning-based voice classification, as it involves identifying significant attributes within the audio signal that can effectively differentiate between different classes. Pitch, frequency, energy, and spectral information are all common characteristics used in voice categorization.

Following the extraction of relevant features, a machine learning algorithm is trained to discover patterns in the data and generate predictions based on the collected characteristics. Support vector machines, decision trees, random forests, and neural networks are examples of popular machine learning methods used in speech categorization.Machine learning vocal classification encompasses a broad spectrum of applications, which include speech recognition, speaker identification, categorizing singer voices, and classifying music genres.Vocal classification algorithms are used in voice recognition to recognise distinct words and phrases in a spoken sentence. Vocal classification algorithms are used in speaker identification to identify various speakers based on their distinct vocal features.

Vocal classification algorithms are used in singer voice classification to identify different singers based on their unique singing style, whereas vocal classification algorithms are used in music genre classification to identify different genres of music based on the vocal characteristics of the audio signal. The ability of machine learning algorithms to learn and adapt to new data is a major advantage when it comes to speech categorization. As new data becomes accessible, the algorithms can be retrained to enhance accuracy and performance. Significant advancements have been made in speech categorization using machine learning in recent years. These include the application of deep learning techniques such as convolutional neural networks and recurrent neural networks.

In several voice categorization tasks, deep learning algorithms are proven to outperform classic machine learning methods, particularly when dealing with big and complicated datasets.

III. PROPERTIES OF VOICE CLASSIFICATION

Machine learning voice classification approaches utilize algorithms that can analyze and classify vocal data based on various parameters such as pitch, rhythm, and timbre. These approaches are valuable for a wide range of applications, including voice recognition, speaker identification, and music genre categorization.

The term "responsibility" refers to the act of determining whether or not a person is responsible for his or her own actions. Large datasets may be used to train machine learning algorithms, allowing them to discover patterns and make accurate predictions. This feature is very crucial for vocal categorization since vocal data can vary a lot depending on aspects like accent, dialect, and speaking style. Machine learning algorithms have been more resilient and precise in their classifications by training on huge datasets.

Another essential aspect of machine learning-based voice classification algorithms is their capacity to extract complicated features from vocal data. Pitch, rhythm, and timbre might be difficult to extract manually, but machine learning algorithms can extract these aspects automatically from data. This attribute allows for more precise and effective categorization of voice data.

Adaptability is another aspect of machine learning vocal classification systems. Machine learning algorithms may be retrained and changed in response to fresh data, resulting in increased accuracy over time. This characteristic is especially crucial for voice recognition applications, where the system must adapt to varied speakers and speaking styles.

Scalability is another significant feature of machine learning-based audio categorization algorithms. Machine learning procedures may be used on big datasets and can efficiently analyze massive volumes of data. This attribute is especially relevant for applications requiring the system to evaluate a large number of audio recordings, such as music genre categorization.

Finally, machine learning-based voice classification algorithms have the virtue of interpretability. Machine learning algorithms can provide insights into the features that have the most impact on categorization judgments, leading to a better understanding and interpretation of the results. This characteristic is particularly important in applications such as voice recognition, where it is crucial to comprehend the reasoning behind a specific categorization.

Overall, the qualities of machine learning-based vocal classification algorithms make them the best for a variety of applications, including voice recognition, speaker identification, and music genre categorization. These algorithms may extract complicated characteristics from vast volumes of data, adapt to new data, scale to enormous datasets, and give insights into categorization judgements.

IV. APPLICATIONS OF VOICE CLASSIFICATION

Machine learning algorithms designed for vocal classification have diverse applications across various domains, including but not limited to speech recognition, speaker identification, singer voice classification, and music genre classification. The term "responsibility" refers to the act of determining whether or not a person is responsible for his or her own actions.

Speech recognition refers to the conversion of spoken words into written text. Machine learning techniques can be employed to identify different words and phrases within spoken speech through vocal categorization. This is accomplished by collecting key elements from the audio stream, such as pitch, frequency, and energy, and then using machine learning algorithms to recognise the various words and phrases. Virtual assistants, transcription services, and voice-to-text software are just a few examples of how speech recognition systems may be employed.

Speaker identification is the process of distinguishing various speakers based on their distinct vocal characteristics. Machine learning algorithms for vocal classification may be used to distinguish various speakers by extracting elements from the audio data such as pitch, timbre, and formant frequencies. Security systems, call centers, and forensic analysis can all benefit from speaker identification systems.

Singer Voice Classification involves distinguishing between different singers based on their unique singing styles. Machine learning methods can be utilized to extract characteristics like vibrato, vibrato rate, vibrato extent, and melodic contour from an audio stream, enabling vocal categorization.

Music Genre Classification aims to classify different music genres based on the vocal attributes present in an audio signal. Machine learning approaches can extract features such as spectral information, melodic contour, and rhythm from an audio stream, facilitating vocal categorization. Music genre classification systems have various applications, including music recommendation systems, music transcription services, and music analysis.

Apart from the applications listed above, vocal categorization utilizing machine learning methods may be utilized for emotion detection, accent recognition, and speech pathology analysis. The technique of detecting distinct emotions in speech signals is known as emotion recognition, whereas accent recognition is the procedure of identifying different accents in speech signals.

Finally, vocal classification utilizing machine learning algorithms has several applications in areas such as speech recognition, speaker identification, singer voice classification, and music genre classification. Vocal categorization systems have gotten increasingly precise and efficient in recent years, thanks to advances in machine learning techniques and the availability of enormous datasets. Continued

study and development in this field may result in the creation of more accurate and efficient vocal categorization systems, allowing for the introduction of novel applications and services.

V. EXISTING METHODS

SUPPORT VECTOR MACHINES (SVM):

SVM is a popular classification algorithm used in various fields, including voice identification. It aims to find the best possible hyperplane that can separate the data into different classes. SVM can handle both linearly and non-linearly separable data by using kernel functions, such as linear, polynomial, and radial basis function (RBF). SVM has been used for speaker identification, where the speech signal is preprocessed to extract features, such as Mel-frequency cepstral coefficients (MFCCs), and fed to the SVM for classification. SVM has shown good classification performance in voice identification tasks, but its effectiveness depends on the choice of kernel function and parameters.

RANDOM FOREST:

Random Forest is an ensemble learning technique that combines multiple decision trees to improve the classification performance. In voice identification, Random Forest can be used to select the most important features that contribute to the classification task. Random Forest can handle noisy and high-dimensional data, and it can be trained efficiently using parallel computing. Random Forest has been used for speaker identification, where the speech signal is preprocessed to extract features, such as MFCCs, and used to train the Random Forest model. Random Forest has shown good classification performance in voice identification tasks, especially when combined with other techniques, such as feature selection.

CONVOLUTIONAL NEURAL NETWORKS (CNN):

CNN is a deep learning technique that has been used in various fields, including voice identification. CNNs are designed to extract meaningful features from the input data using multiple layers of convolutional and pooling operations. CNNs can handle high-dimensional data, such as speech signals, and can automatically learn the feature representations from the data. CNNs have been shown to outperform traditional machine learning algorithms in voice identification tasks. In speaker identification, CNNs can be used to directly process the speech signal without the need for explicit feature extraction. CNNs can also be used for emotion recognition and speech recognition tasks.

RECURRENT NEURAL NETWORKS (RNN):

RNN is a deep learning technique that can handle sequential data, such as speech signals. RNNs can be used for speaker identification by processing the speech signal frame by frame and learning the speaker-specific patterns in the speech signal. RNNs can handle variable-length sequences and can learn long-term dependencies in the data. RNNs suffer from the vanishing gradient problem, where the gradients become too small to update the weights, but this problem can be mitigated by using variants of RNN, such as Long Short-Term Memory (LSTM).

LONG SHORT-TERM MEMORY (LSTM):

LSTM is a type of RNN that can handle long-term dependencies in sequential data. LSTMs have been used in voice identification tasks to process the speech signal over long time periods and extract features that capture the speaker-specific patterns. LSTMs can handle variable-length sequences and can learn long-term dependencies in the data without suffering from the vanishing gradient problem. LSTMs have shown good classification performance in voice identification tasks, especially when combined with other techniques, such as CNNs and Random Forest.

ENSEMBLE LEARNING:

Ensemble learning is a machine learning technique that combines multiple classifiers to improve the classification performance. In voice identification, ensemble learning can be used to combine the predictions of multiple classifiers trained on different feature sets or using different algorithms. Ensemble learning can improve the classification accuracy and robustness of the model by reducing the variance and bias of the individual classifiers. Ensemble learning has been used in voice identification tasks to combine the predictions of SVMs, Random Forests, and other classifiers, and has shown good classification performance. However, ensemble learning requires more computational resources and can be sensitive to the choice of ensemble method and hyperparameters.

VI. DATA SOURCES

The data sources used in the below papers for voice identification using machine learning techniques can be broadly classified into two categories: (i) speech datasets and (ii) music datasets

SPEECH DATASETS:

Speech datasets play a vital role in training and evaluating machine learning models for voice identification. These datasets can be categorized into two main types: speaker-dependent datasets and speaker-independent datasets.

Speaker-dependent datasets comprise recordings of a single speaker or a small number of speakers. They are employed to train and test voice identification models that are specific to recognizing the voice of a particular speaker. The TIMIT dataset, for instance, includes recordings of 630 speakers from various regions in the United States. Another example is the VoxCeleb dataset, which consists of recordings of 7,000 celebrities sourced from YouTube videos.

In contrast, speaker-independent datasets encompass recordings of numerous speakers with diverse backgrounds and regional variations. These datasets are utilized to train and test voice identification models that can recognize the voice of any speaker, regardless of their specific identity. The NIST SRE dataset is an instance of a speaker-independent dataset, containing recordings of over 4,000 speakers speaking different languages and dialects. Additionally, the Switchboard dataset comprises telephone conversations involving 2,400 speakers.

To summarize, speech datasets are classified into speaker-dependent and speaker-independent categories, catering to the training and evaluation of voice identification models. Speaker-dependent datasets focus on specific speakers, while speaker-independent datasets encompass a broader range of speakers. The selection of an appropriate dataset depends on the specific requirements and objectives of the voice identification task.

MUSIC DATASETS:

Music datasets serve as collections of music recordings that are utilized for training and evaluating machine learning models in singer voice classification. These datasets can be categorized into two main types: genre-dependent datasets and genre-independent datasets.

Genre-dependent datasets encompass recordings belonging to a single music genre or a small number of genres. They are employed to train and test singer voice classification models that specialize in recognizing a singer's voice within a specific music genre. Examples of genre-dependent data sets mentioned in the previous papers include the MIR-1K dataset, comprising 1,000 songs from 10 different genres, and the MedleyDB dataset, consisting of 122 songs from diverse genres.

In contrast, genre-independent datasets encompass recordings encompassing a wide range of songs from various genres. These datasets are utilized to train and evaluate singer voice classification models capable of recognizing a singer's voice across different music genres. Examples of genre-independent datasets discussed in the previous papers include the Million Song Dataset, which contains over one million songs spanning different genres, and the Acapella dataset, featuring 142 acapella songs from diverse genres.

To summarize, the selection of a data source relies on the specific task of voice identification or singer voice classification and the type of machine learning model employed. Considerations such as dataset availability, size, speaker or singer diversity, and recording quality play crucial roles when choosing a suitable data source.

TIMIT:

The TIMIT dataset is a widely used speaker-dependent speech dataset for voice identification. It consists of recordings of 630 speakers from different regions of the US, speaking 10 different sentences. The dataset includes both clean speech and speech corrupted by background noise, reverberation, and channel distortion. The TIMIT dataset has been used extensively in voice identification research and is considered a benchmark for evaluating speaker-dependent models. You can find sample TIMIT recordings and more information about the dataset at this link: https://catalog.ldc.upenn.edu/LDC93S1

VOXCELEB:

The VoxCeleb dataset [9] is a speaker-dependent speech dataset that consists of recordings of 7,000 celebrities from YouTube videos. The dataset includes speech from a diverse range of people, including actors, politicians, and musicians. The recordings are unscripted and cover a wide range of topics, making the dataset a challenging benchmark for speaker-dependent models. You can find more information about the VoxCeleb dataset and sample recordings at this link: http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

NIST SRE:

The NIST SRE (Speaker Recognition Evaluation) dataset is a widely used speaker-independent speech dataset for voice identification. It consists of recordings of over 4,000 speakers from different languages and dialects, speaking a variety of sentences. The dataset includes both clean speech and speech corrupted by background noise, reverberation, and channel distortion. The NIST SRE dataset has been used extensively in speaker-independent voice identification research and is considered a benchmark for evaluating speaker-independent models. You can find more information about the NIST SRE dataset and sample recordings at this link: https://www.nist.gov/itl/iad/mig/nist-2021-speaker-recognition-evaluation-sre21

SWITCHBOARD:

The Switchboard dataset [7] is a widely used speaker-independent speech dataset for voice identification. It consists of recordings of telephone conversations between 2,400 speakers, covering a wide range of topics. The dataset includes both clean speech and speech corrupted by background noise, channel distortion, and disfluencies such as false starts and repetitions. The Switchboard dataset has been used extensively in speaker-independent voice identification research and is considered a benchmark for evaluating speaker-independent models.You can find more information about the Switchboard dataset and sample recordings at this link: https://catalog.ldc.upenn.edu/LDC97S62

MIR-1K:

The MIR-1K dataset is a genre-dependent music dataset for singer voice classification. It consists of recordings of 1,000 songs from 10 different genres, including blues, country, jazz, and rock. The dataset includes both solo and ensemble performances, with a total of 1,000 unique singers. The MIR-1K dataset has been used extensively in singer voice classification research and is considered a

benchmark for evaluating genre-dependent models.You can find more information about the MIR-1K dataset and sample recordings at this link: http://mirlab.org/dataset/public/

MEDLEYDB:

The MedleyDB dataset [8] is a genre-dependent music dataset for singer voice classification. It consists of recordings of 122 songs from different genres, including classical, folk, pop, and rock. The dataset includes both solo and ensemble performances, with a total of 38 unique singers. The MedleyDB dataset includes multiple sources for each song, allowing researchers to explore the effects of different types of accompaniment on singer voice classification. Find more information here: https://medleydb.weebly.com/

VII.RELATED WORK

Research on vocal classification utilizing machine learning techniques has been ongoing for several years, encompassing a wide range of studies that delve into various aspects of the field. The related work in this domain can be categorized into several areas, including speech recognition, speaker identification, singing voice classification, music genre classification, and emotion recognition.

In the realm of speech recognition, machine learning techniques have been employed to classify spoken words based on their acoustic features. One notable early study by **Waibel et al. (1990)** utilized hidden Markov models to achieve over 80% accuracy in recognizing spoken words. Subsequent research has explored the application of other machine learning methods such as artificial neural networks and support vector machines, leading to improved accuracy and performance (**Sainath et al., 2015; Park et al., 2016**).

Speaker identification is another area that has witnessed the application of machine learning techniques. The objective here is to identify the speaker of a given vocal recording based on their unique voice characteristics. Studies in this field have focused on leveraging various acoustic features, including pitch, formants, and spectral features, for accurate speaker identification. Machine learning algorithms such as Gaussian mixture models, neural networks, and decision trees have been employed with high levels of accuracy (**Reynolds et al., 2000; Kinnunen and Li, 2010; Ghosh et al., 2015**).

Singing voice classification is a closely related area where machine learning techniques have found utility. The aim in singing voice classification is to categorize vocal recordings based on the voice and style of the singer. For instance, **Pham et al. (2019)** utilized deep learning techniques to achieve over 90% accuracy in classifying Vietnamese popular music based on the singer's voice. Another study by **You et al. (2020)** compared the performance of various deep neural networks and ensemble learning techniques for singing voice detection.

Music genre classification is yet another field where machine learning techniques have been extensively applied. The objective in music genre classification is to categorize music recordings into different genres based on their musical features. Studies in this area have utilized various features such as tempo, rhythm, and timbre for accurate genre classification. Machine learning algorithms such as support vector machines, decision trees, and neural networks have been employed with high levels of accuracy (**Tzanetakis and Cook, 2002; Lee and Lee, 2004; Won et al., 2020**).

Furthermore, research has been conducted on emotion recognition from vocal recordings using machine learning techniques. Emotion recognition aims to classify vocal recordings into different emotional categories such as happiness, sadness, or anger. Machine learning algorithms including decision trees, support vector machines, and artificial neural networks have been applied with high levels of accuracy (**Schuller et al., 2010; Eyben et al., 2015; Deng and Hu, 2018**).

In summary, the related work in the field of vocal classification using machine learning techniques has been extensive and diverse. Numerous studies have focused on different aspects of the field, including speech recognition, speaker identification, singing voice classification, music genre classification, and emotion recognition. Machine learning algorithms such as decision trees, support vector machines, artificial neural networks, and deep learning techniques have been effectively utilized in these applications, leading to high levels of accuracy and performance.

VIII. COMPARATIVE ANALYSIS

The papers by Orken Mamyrbayev, Toan Pham Van, and Shingchern D. You investigate various approaches for speaker and singer voice classification using machine learning algorithms.

Mamyrbayev's paper explores the use of classification algorithms such as K-Nearest Neighbor (KNN), Decision Trees (DT), Support Vector Machine (SVM), and Random Forests (RF) on the TIMIT dataset. The results indicate that SVM achieves the highest accuracy of 96.7%. The study emphasizes the significance of feature selection and reduction for efficient classification.

In contrast, Pham Van's paper adopts a deep learning approach for singer voice classification in Vietnamese popular music. The study employs the MIR-1K dataset to train a convolutional neural network with multiple layers including convolutional, pooling, and fully connected layers. The CNN model outperforms other techniques, achieving an accuracy of 96.5%. The paper highlights the importance of selecting the appropriate neural network architecture for effective classification.

You's paper compares the performance of deep neural networks and ensemble learning for singing voice detection. The study utilizes the MedleyDB dataset for training and evaluation. The results indicate that ensemble learning surpasses DNNs with an accuracy of 92.3%. The paper underscores the relevance of employing ensemble methods for accurate classification when dealing with complex and diverse data.

While the papers employ different datasets and techniques for voice classification, there are commonalities in their approaches. All the papers emphasize the importance of feature selection and reduction in achieving effective classification. Additionally, different machine learning algorithms are used for classification: Mamyrbayev and You employ traditional approaches such as SVM and ensemble methods, respectively, while Pham Van utilizes deep learning techniques like CNN.

Furthermore, the papers utilize distinct datasets for training and evaluation. Mamyrbayev and You employ speech datasets such as TIMIT and Switchboard, while Pham Van uses a genre-dependent music dataset like MIR-1K. This diversity in datasets allows for evaluating the performance of different techniques under varying conditions and facilitates comparison with other studies.

In terms of limitations, all the papers suffer from limitations related to the size and diversity of the datasets used. While TIMIT, Switchboard, MIR-1K, and MedleyDB are widely used and accepted benchmark datasets, they are still limited with respect tothe number of samples and diversity of the data. This makes it difficult to generalize the results to other datasets and real-world scenarios. Additionally, while the use of traditional machine learning algorithms and deep learning techniques is effective, they require significant computational resources and expertise for effective implementation.

In conclusion, the papers by Mamyrbayev, Toan Pham Van, and You provide valuable insights into different techniques for speaker and singer voice classification using machine learning algorithms. While each of the papers uses a different dataset and technique, they all highlight the importance of feature selection and reduction for effective classification. Additionally, the use of traditional machine learning algorithms and deep learning techniques for classification provides a basis for further research in this area.

IX. FUTURE SCOPE

The future potential for vocal classification using machine learning is enormous and encouraging as the field of machine learning develops. Future study and development may focus on some of the following areas:.

Multimodal integration: To increase the precision of vocal classification systems, future research may investigate the multimodal integration of audio, video, and textual data.

Identifying speakers or singers across multiple languages is possible with the development of multilingual vocal classification systems. Applications such as speaker recognition, music genre categorization, and language identification may benefit from this.

Real-time classification: Systems for categorizing vocal signals can be tuned to operate in real-time, opening the door to uses like automated transcription, voice-activated gadgets, and speech recognition.

Robustness to environmental noise: Future research can concentrate on creating vocal classification systems that are resilient to outside noise, such as crowd noise or background music, which can impair the system's accuracy.

Integration with other applications: To enable more effective and individualized services, vocal classification systems can be integrated with other applications like virtual assistants, smart homes, and healthcare systems.

Ethical issues: As vocal classification systems proliferate, it is crucial to take into account the ethical implications of their use, including privacy issues and potential bias in the classification process.

Overall, there is a huge future potential for vocal classification using machine learning, with potential applications in entertainment, healthcare, and security. The advancement of more precise and effective vocal classification systems may result from additional research and development in this field, opening the door to the development of cutting-edge services and applications.

X. CONCLUSIONS

In summary, the utilization of classification algorithms for voice identification holds significant potential across diverse domains, encompassing applications like speech recognition, speaker identification, and vocal classification. It presents several benefits, including the capacity to accurately categorize human emotions from speech signals and provide real-time feedback. However, it also possesses limitations, such as reliance on speech signal quality and algorithm complexity. Researchers have employed various techniques, such as feature extraction, classification algorithms, and database selection, to analyze speech signals and assign them to distinct categories. As ongoing research progresses, voice identification using classification algorithms has the potential to become more precise and find broader utilization in diverse applications.

XI. REFERENCES

[1] Mamyrbayev, O., Mekebayev, N., Turdalyuly, M., Oshanova, N., Ihsan Medeni, T., & Yessentay, A. (2020). Voice Identification Using Classification Algorithms. IntechOpen. doi: 10.5772/intechopen.88239

[2] Van, T. P., Tran, N. N., & Thanh, T. M. (2019). Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music. Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019, 255–260. https://doi.org/10.1145/3368926.3369700

[3] You, S. D., Liu, C.-H., & Chen, W.-K. (2018). Comparative study of singing voice detection based on deep neural networks and ensemble learning. Human-Centric Computing and Information Sciences, 8(1), 34. https://doi.org/10.1186/s13673-018-0158-1

[4] Roers, F., Mürbe, D., & Sundberg, J. (2009). Predicted Singers' Vocal Fold Lengths and Voice Classification—A Study of X-Ray Morphological Measures. Journal of Voice, 23(4), 408–413. https://doi.org/10.1016/j.jvoice.2007.12.003

[5] Song, Y. & Kim, I.. (2018). DeepAct: A deep neural network model for activity detection in untrimmed videos. Journal of Information Processing Systems. 14. 150-161. 10.3745/JIPS.04.0059.

[6] Romain Serizel, Diego Giuliani. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. 2014 IEEE Spoken Language Technology Workshop (SLT 2014), Dec 2014, South Lake Tahoe, CA, United States. pp.135-140, ff10.1109/SLT.2014.7078563ff. ffhal-01393972f

[7] Godfrey, John J., and Edward Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[8] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research", in 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, Oct. 2014.

[9] Arsha Nagrani, Joon Son Chung, Weidi Xie, Andrew Zisserman, Voxceleb: Large-scale speaker verification in the wild, Computer Speech & Language, Volume 60, 2020, 101027, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2019.101027

[10] Beigi, Homayoon. (2011). Fundamentals of Speaker Recognition. 10.1007/978-0-387-77592-0.

[11] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using Non-negative Matrix Factorization," 2008 International Conference on Advanced Technologies for Communications, Hanoi, Vietnam, 2008, pp. 243-246, doi: 10.1109/ATC.2008.4760565.

[12] Nwe, Tin & Wang, Ye. (2004). Automatic Detection Of Vocal Segments In Popular Songs..

[13] Toan Pham Van, Ngoc Tran Ngo Quang, and Ta Minh Thanh. 2019. Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music. In Proceedings of the 10th International Symposium on Information and Communication Technology (SoICT '19). Association for Computing Machinery, New York, NY, USA, 255–260. https://doi.org/10.1145/3368926.3369700

[14] Sundberg J. (1987). The science of the singing voice. Northern Illinois University Press.

[15] Aatto Sonninen (1954) Is the Length of the Vocal Cords the same at all Different Levels of Singing?, Acta Oto-Laryngologica, 43:sup118, 219-231, DOI: 10.3109/00016485409124010