Audio Similarity Detection

Siddharth Harsh Y. Malhotra
CSE Department, Institute of Technology
Nirma University
Ahmedabad, India
23mce021@nirmauni.ac.in

Sapan H. Mankad
CSE Department, Institute of Technology
Nirma University
Ahmedabad, India
sapanmankad@nirmauni.ac.in

***Abstract*—Audio similarity detection plays a crucial role in various applications, such as music recommendations, speaker recognition, copyright infringement detection, and content-based retrieval. Various categories of similarity types such as Audio signal similarity, audio content similarity, audio perceptual similarity, speaker voice similarity, and audio caption similarity are covered by the researchers to identify similarity in audio signals. We have performed various experiments to investigate the appropriate audio embeddings for this task and to identify which feature and distance metrics are more appropriate for identifying speaker and content similarity. Various short-term spectral features have been extracted from the audio signal for feature extraction tasks which are then used for calculating a similarity matrix using Euclidean similarity and Cosine similarity. The performance of the fused features is better than the standalone features. Also, with the use of pre-trained features which are extracted from YAMNet, the model's performance has been improved.**

***Index Terms*—Audio Similarity, Speaker-speaker similarity, Euclidean Distance, Cosine similarity**

I. Introduction

Audio similarity detection is the process of identifying and quantifying the similarities between two or more audio signals. It can be applied in a variety of fields, including voice casting, speaker recognition, audio categorization, audio information retrieval, and audio captioning. Various characteristics of the audio signals, such as the signal waveform, the content or meaning, the perceptual quality, the speaker identity, or the caption text, can be used to determine audio similarity. Audio similarity can be measured and assessed using a variety of techniques and metrics, depending on the purpose and the data at hand.

Finding reliable and significant features that capture relevant aspects of similarity in audio signals is one of the challenges in audio similarity detection.

Algorithms that utilise deep learning have been proposed to automatically learn audio properties from data in order to get over such limitations. Multiple layers of nonlinear transformations make up deep learning models, which are able to develop intricate and sophisticated representations of the audio signals. These representations can serve as compact, fixed-length vectors called embeddings, which condense the properties of the audio signals. Simple metrics of similarity or distance, like contrastive loss, cosine similarity, or Euclidean distance, can be used to compare embeddings. Without using intermediary embeddings, deep learning models can also be trained end-to-end to directly optimise the similarity metric.

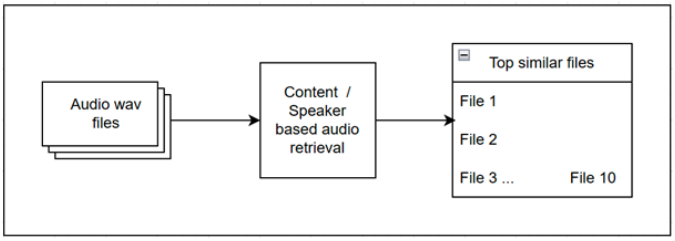

Fig. 1. Audio Similarity Model

Since the similarity of audio can be based on a lot of features such as speakers' voice, similar content, similar audio signals, perceptual similarity, similar melody, rhythm, beats in case of music, a lot of features extraction methods and features have been proposed by the researchers. The main objective of this study is to find which features are more suited to find which type of similarity in audio signals. In this study, we have experimented to find which features and distance metrics for calculating similarity are suited to find audio speaker and audio content similarity.

II. Related Works

Purwins et al. [1] provide an overview of the various deep learning methods for audio signal processing, considering the current explosion in deep learning research and development. The domains of speech, music, and environmental sound processing are compared side by side to identify similarities and distinctions, as well as to highlight common approaches, issues, important sources, and opportunities. A review is conducted on the popular feature representations used including log-mel spectra and raw waveform, as well as other deep learning methods, such as convolutional neural networks (CNNs), long short-term memory (LSTM) architecture variants, and more audio-specific neural network models. Along with that, well-known deep learning application domains using audio similarity are discussed including audio recognition (autonomous speech recognition, music information retrieval, ambient sound detection, localization, and tracking) and synthesis and transformation (source separation, audio improvement, generative analysis). The review discussed various methods

for appropriate feature representation including mel frequency cepstral coefficients (MFCC), log-mel spectrum and spectograms along with their advantages. Various models such as CNNs, Recurrent Neural Networks (RNNs), Sequence-to-Sequence Models and Generative Adversarial Networks (GANs) are discussed along with appropriate loss functions to be used with them and phase modeling for the audio signal similarity detection.

Peng et al. in [2] suggest a novel method for ranking audio clips similarity using graph modelling and matching. A new approach, Segment-based similarity is suggested as a substitute for frame-based or salient-based criteria in assessing the acoustical similarity of audio clips. This approach uses segment-based representation, then it uses a similarity measure and ranking system based on four different types of similarity characteristics. Segments retain and display the temporal order and change relationship between audio components in addition to capturing the changing properties of the audio clip in segment-based representation. There are four types of similarity factors shared which are acoustical, granularity,temporal order and interference which are measured by optimal matching and dynamic programming progressively and jointly. A bipartite graph computed by Optimal matching method has been extended and used for similarity of audio with dynamic programming. After calculation of interference factor, Audio clip similarity is measured jointly by the degree of acoustical and granularity similarity, the temporal order of matching, and interference factor.

P. Manocha et al. [3] introduce the Contrastive Deep Perceptual Audio Metric (CDPAM), a revolutionary deep learning-based approach to voice processing techniques. This method improves generalisation to a wider range of audio disturbances by gathering human opinions on triplet comparisons. It has been demonstrated that CDPAM correlates well with human reactions in nine different datasets. The study also shows that, as evaluated by both objective and subjective testing, the addition of this metric to current speech synthesis and augmentation techniques results in a notable improvement. The CDPAM method uses limited human-labeled data to supplement it with multi-dimensional, self-supervised learning. The training architecture shared in the paper is as follows. First an audio encoder is trained using contrastive learning, then trained the loss-net on JND data. Finally, the loss-net is fine tuned on the newly collected dataset of triplet comparisons.

K. Seyerlehner et al. in [4] present a novel multi-level vector quantization (ML-VQ) approach for modelling audio content based on local spectral properties in this research. The proposed method, ML-VQ method has multiple benefits and performs similarly to the most advanced frame-level audio similarity techniques. By implying a normed vector space, the suggested histogram intersection distance makes it possible to use more potent search techniques and to enable content-based music retrieval for even bigger music archives. This paper also indicates that the generated acoustic similarity space does not suffer from the hub problem that affects Gaussian Mixture Models of songs.

P. Avgoustinakis et al. in [5] propose a unique method for audio-based near-duplicate video retrieval called Audio Similarity Learning (AuSiL) in this paper. By extracting representative audio-based video descriptors by transfer learning from a CNN trained on a large dataset of audio events, the method identifies temporal patterns of audio similarity between video pairings. This CNN network extracts the temporal structures in its content is fed the similarity matrix that is created by comparing these descriptors pairwise. After generating triplets, the network is trained by optimising the triplet loss function. Its model is as follows. Each video's spectrogram is fed into the feature extraction procedure, which pulls out the feature vectors for every audio clip. Next, the dot product between the feature vectors of the two videos is used to create a similarity matrix. To capture the temporal patterns of the segment-level within-video similarities, the resulting matrix is sent to Audio Similarity CNN. Chamfer Similarity is used to aggregate the final similarity score.

X. Yu et al. [6] present an innovative method for measuring audio similarity based on Renyi's quadratic entropy in the paper. Since, noise interference occurs during audio processing, standard distance metrics are not reliable for determining audio similarity. To represent each audio in their method, MFCCs are extracted. Subsequently, the probability density function (pdf) of MFCCs is computed using the entropy of audio samples, which can be approximated using the Parzen window. The main process in this paper's application of Renyi's Quadratic Entropy include feature extraction, similarity measure and smoothening parameter. Their results show that the proposed algorithm exceeds the Euclidean distance in accuracy.

Q. Wu et al. in [7] explore the retrieval of perceptually similar audio, focusing on finding sounds according to human perceptions which investigates the retrieval of perceptually comparable audio. Comparatively speaking, this method is more "human-centered" than earlier audio retrievals that sought to identify similar sounds. The authors measure perceptual similarity by utilising a wide range of auditory features. Their results show that Line Spectral Pairs (LSP), MFCC, and Perceptual Linear Prediction (PLP) are the three most effective acoustic features. Moreover, the optimal combination of features can improve the accuracy of similarity classification compared with the best performance of a single acoustic feature.

M. K. Singh et al. in [8] used the MFCC feature extraction approach and the Euclidean distance to measure the similarity between speakers in this study. This approach is used in forensic statistics for voice similarity measurement. Euclidean distance is used to quantify the speaker voice similarity between two distinct sets of speakers: same speakers and dissimilar speakers. The voice of the speaker has a significant impact on the speech feature coefficient. When calculating speaker similarity, Euclidean distance is derived from the MFCC mean. The similarity index between same speakers and dissimilar speakers is found using the Euclidean distance difference. Their results show that the Euclidean distance is minimum for the same speakers who speak the same text

or message and maximum when different speakers speak the same text or message.

A. Gresse et al. [9] discussed a novel approach to voice casting, a process in dubbing where the original voice in a source language is replaced by a new one in a target language. The authors have proposed a Siamese Neural Networks-based approach to model a voice similarity metric that captures all voice characteristics, including the observers' receptive interests. The paper hypothesizes that pairwise relationships between different voices can be used to learn a similarity metric for professional acted voices. The experimental results show that Siamese Neural Networks can model this abstract notion of similarity and provide better generalization results on unseen voices compared to classic architectures. The learnt metric can discriminate target and non-target pairs on new speakers/characters, highlighting the abstract information sought.

J. H. Jensen et al. in [10] covered a technique for utilising an anchor space to calculate similarity in this paper. Spectral clustering is used in this method to classify audio recordings into semantic groups based on shared low-level properties. Using the data corpus, the spectral clustering method first creates an affine matrix that compares the similarity of each pair of data points, in this example, one-second audio segments. After extracting the eigenvectors, the original data is mapped into a low-dimensional space that is simple to cluster using a Singular Value Decomposition (SVD). The authors suggest a simplification technique in which the mean vector of each audio document's feature vectors is used to represent it in order to lessen this. After that, the collection of mean vectors calculated for the complete data set is subjected to spectral clustering, and the resulting clusters are chosen to serve as anchors. The study also covers an estimating method based on the eigen-gap that spectral clustering suggests for figuring out how many clusters need to build.

J. H. Jensen [11] presented a MIDI-based test framework for analysing music similarity measures in this paper. The framework is used to investigate how sensitive a music similarity measure is to transpositions and how dependent it is on a song's instrumentation as opposed to its melody. Three software programmes are analysed for their music similarity measures: Marsyas, MA toolbox, and Intelligent Sound Processing toolbox. Although the investigated timbral similarity measures are sensitive to transpositions and variations in sound fonts, they demonstrate good performance in instrument recognition. Sometimes the same melody played on different instruments is not recognised by the beat/rhythm/melody similarity metrics. According to the study, timbral similarity metrics did not translate well to other sound typefaces. In cases where certain smooth spectral adjustments, like changing the bass and treble, do not significantly impact the impression of timbre, it is hypothesised that timbral similarity measurements that additionally rely on the temporal envelope will better reflect human sound perception. The study also points out that melody recognition still needs work. Only a genre categorization experiment could not have produced the results that were obtained. The influence of melody and instrumentation have been successfully separated through the use of MIDI files, and a signal change called transposition has been included. The study makes the case that measuring how similarity metrics are affected by pace, instrument combinations, bandwidth, and audio compression. Various feature sets are tested for that in this paper which included Timbre, MFCC, Beat, Pitch, Sone, Spectrum histogram, Periodicity histogram, fluctuation pattern and multivariate autoregressive model.

Z. Li and P. Song [12] proposed an algorithm for detecting audio similarity based on a Siamese LSTM network is proposed in this paper. The feature matching model and the selection of audio signal features are the main technologies. LSTM is used in the Siamese network's fundamental network segment as part of the approach. In order to compute audio similarity, the method starts with the extraction of Filter banks characteristics from two audio signals. In order to compare the similarity of two samples, the Siamese LSTM network model outputs their high-dimensional space representation. The network is made up of two neural networks that share weights and have the same structure. The FBank feature of each audio segment is used as input, processed by LSTM, and then coupled to a multi-layer complete connection, replacing the base network of the Siamese neural network. In audio similarity calculations based on deep learning, the paper proves the usefulness of the FBank feature and claims that it outperforms the MFCC feature. The Siamese LSTM network and the FBank feature work together to precisely determine how similar two audio portions are.

S. Bhosale [13] presented in this study a new measure in order to assess Automatic Audio Captioning (AAC), a task that requires summarising an audio sample in natural language text. Acoustic semantics is needed for AAC in order to map natural language text to analogous sounds, in contrast to other natural language activities that use lexical semantic metrics for evaluation. The authors include these acoustic semantics into a unique metric that is based on Text-to-Audio Grounding (TAG). They point out that although tasks involving natural language, such as language translation and picture captioning, are anticipated to yield outputs that are semantically comparable, AAC activities do not follow this pattern. As an example, the terms "clock" and "car-turn indicator," which are semantically distant in the Word Embedding (WE) space, may sound identical and appear interchangeably in an audio caption. The research suggests a novel embedding space based on TAG that can map words that produce comparable sounds next to one another in order to fix this. The creation of an acoustic embedding (s2v) that produces similar embeddings for text corresponding to acoustically similar noises is the primary contribution. This metric is composed of two modules: one for Phrase Extraction (PE), which extracts phrases from text; and another for TAG, which creates s2v embeddings for each phrase. The trials show that the suggested evaluation metric outperforms other metrics that are currently in use in the research on AAC for natural language text and picture captioning.

W. Wang et al. [14] presented a Neural-network-based approach to measuring similarity between any two speaker embeddings that takes into account both past and future contexts in this research. For similarity measurement, a segmental pooling technique is suggested, which gives the embeddings more context and hence better performance. To further improve performance, the speaker embedding network's parameters are unfrozen and trained in conjunction with the similarity measurement goal. This joint training approach provides a generic framework for segment-level target-speaker voice activity detection (TS-VAD) and similarity measuring. The authors extended the system to TS-VAD and investigate the impact of various pooling sizes. The findings demonstrate that the MISS in the Diarization Error Rate (DER) is considerably decreased by the TS-VAD approach.

III. Proposed Approach

In this work, we aim to investigate impact of different audio embeddings in evaluating the speaker-speaker similarity and content similarity. To evaluate our proposed approach, we use a subset of ASVspoof 2017 v 2.0 dataset which includes only genuine speech data.

In Fig. 2, a general architecture of the proposed system with feature extraction module and similarity calculation module is shown. First of all various short term spectral features like MFCC, BFCC, CQCC etc. are extracted from audio wav files. On these extracted features, similarity is calculated with distance metrics- Euclidean distance and Cosine Similarity. The top similar files are returned as output to the user. The performance of these features is noted down.

In the next approach, the YAMNet model has been used to extract deep features from audio wav files. On these extracted features, similarity is calculated with distance metrics- Euclidean distance and Cosine Similarity. The top similar files are returned as output to the user. The performance of these deep features is compared with that of the previous approach. If both the models performance is not satisfactory, then both of the features can be combined and then used to calculate the similarity of the audio wav files.

1) Extract features from each wav file and represent it in form of 13- dimensional feature vector. We have used around 9-10 feature representations including MFCC etc. Total 3565 audio files are there.
2) Compute Euclidean distance between each feature vector. This gives a 3565x3565 distance matrix.
    a) Further looking into and analysing this matrix shows us some insights.
    b) We found top 10 closest audio pairs for each feature.
    c) We also find which two speakers are the most similar as per all the features.
    d) Observing them showed that most of the features focus on speaker similarity rather than content similarity.
3) We repeat the same experiment with Cosine similarity based approach.

A. Features

We use short-term spectral features to represent the input. We use them as standalone first, and then we see the impact of fused features for the underlying task.

Pre-trained features are extracted from wav files with the help of the deep YAMNet model [15]. It is a pre-trained deep learning learning model which is trained on AudioSet corpus and can predict around 521 audio events. This model is available on TensorFlow Hub and can give 3 outputs which are class scores, embeddings and Log Mel Spectrograms.

B. Algorithm

Algorithm 1 Feature Extraction

```
Set file path
Require features not to be null
procedure FeatureExtraction
    for each metric do
        Take one feature
        Generate feature matrix
        Find top most similar files based on feature matrix
    end for
end procedure
```

IV. Experimental Results and Discussion

A. Datasets used

To test the effectiveness of our proposed approach, we used ASVSpoof 2017 version 2.0 dataset [16]. This dataset comprises bonafide and spoofed audio samples from ten differnt speakers and phrased through repeated trials under various recording, and playback conditions.

There are 10 common phrases which are spoken by 42 individual speakers. Since, we have details about content spoken and speaker in the dataset, we can use that to identify which features are able to identify similar content or similar speaker voice.

We focus on examining speaker-speaker similarity and speech-content similarity through these experiments.

B. Metrics used

To determine the distance between two input audio files, we used Euclidean distance and Cosine similarity measure.

1) Euclidean Distance: The Euclidean distance between two feature vectors $\mathbf{u}$ and $\mathbf{v}$ can be computed using the following formula:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$$

where:

- $\mathbf{u} = (u_1, u_2, \ldots, u_n)$
- $\mathbf{v} = (v_1, v_2, \ldots, v_n)$
- n is the number of dimensions in the feature vectors.

TABLE I
LITERATURE SUMMARY TABLE

Author	Year	Title	Key methodology used
Y. Peng et al	2006	Audio similarity measure by graph modeling and matching	graph modeling and matching segment-based similarity
J. H. Jensen et al.	2007	A framework for analysis of music similarity measures	Mel frequency cepstral coefficient Multiple signal classification
K. Seyerlehner et al	2008	Frame level audio similarity-a codebook approach	multi-level vector quantization approach
L. Lu and A. Hanjalic	2008	Unsupervised anchor space generation for similarity measurement of general audio	Spectral clustering anchor space
X. Yu et al.	2010	Audio similarity measure based on Renyi's quadratic entropy	Renyi's Quadratic Entropy Mel Frequency Cepstral Coefficients
Q. Wu et al	2012	Perceptual similarity between audio clips and feature selection for its measurement	acoustic features perceptual similarity
M. K. Singh et al.	2019	Speaker's Voice Characteristics and Similarity Measurement using Euclidean Distances	Mel-frequency cepstral coefficient Euclidean distance
A. Gress et al.	2019	Similarity Metric Based on Siamese Neural Networks for Voice Casting	Siamese Neural Networks
H. Purwins et al.	2019	Deep Learning for Audio Signal Processing	Deep learning Computational modeling
P. Manocha et al.	2021	CDPAM: Contrastive Learning for Perceptual Audio Similarity	Contrastive Deep Perceptual Audio Metric
P. Avgoustinakis et al.	2021	Audio-based Near-Duplicate Video Retrieval with Audio Similarity Learning	Convolutional neural networks spectrogram
Z. Li et al.	2021	Audio similarity detection algorithm based on Siamese LSTM network	Siamese LSTM network
W. Wang et al	2022	Similarity Measurement of Segment-Level Speaker Embeddings in Speaker Diarization	neural-network-based similarity segmental pooling strategy
S. Bhosale et al	2023	A Novel Metric For Evaluating Audio Caption Similarity	Text-to-Audio Grounding s2v embeddings

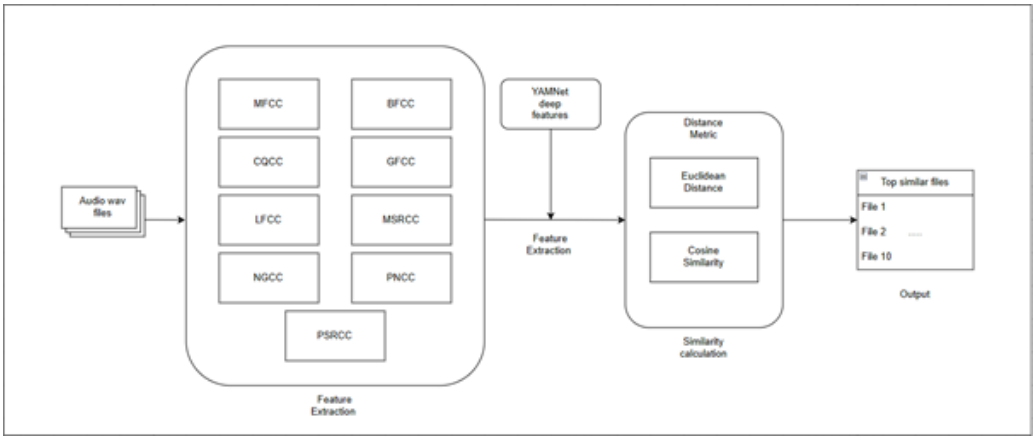

Fig. 2. Architecture of audio similarity detection system

2) Cosine Similarity: The cosine similarity between two feature vectors $\mathbf{A}$ and $\mathbf{B}$ is calculated using the following formula:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Here:

- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors $\mathbf{A}$ and $\mathbf{B}$.
- $\|\mathbf{A}\|$ is the magnitude (or norm) of vector $\mathbf{A}$.
- $\|\mathbf{B}\|$ is the magnitude (or norm) of vector $\mathbf{B}$.

The dot product $\mathbf{A} \cdot \mathbf{B}$ is given by:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{n} A_i B_i$$

The magnitude of a vector $\mathbf{A}$ is given by:

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} A_i^2}$$

Similarly, the magnitude of vector $\mathbf{B}$ is:

$$\|\mathbf{B}\| = \sqrt{\sum_{i=1}^{n} B_i^2}$$

Putting it all together, the formula for cosine similarity is:

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

This formula measures the cosine of the angle between two vectors in a multi-dimensional space, which indicates how similar the two vectors are. The value ranges from -1 (exactly opposite) to 1 (exactly the same), with 0 indicating orthogonality or no similarity.

C. Discussion

After getting top similar files for each file with previously mentioned distance metrics, the following observations were noted:

- When using Euclidean distance as distance metric, with all of the feature extraction techniques it was able to identify same speaker but not content/common phrases spoken.
- When using cosine similarity as distance metric, it was able to identify similar content in some of the wav files with features extraction techniques- MFCC, CQCC, MSRCC and NGCC.
- With CQCC feature extraction technique, it was able to identify same content in top similar files.
- For both distance metrics and all feature extraction techniques, it was able to find the same speakers in top most similar files.
  With the use of YAMNet deep features, the model is able to find more audio wav files with the similar speakers.

V. Conclusion and Future Work

The field of audio similarity detection is broad and diverse, and a lot of cutting-edge techniques are being created. These techniques, which have many uses, including as speaker identification and music retrieval, are always developing in tandem with advances in computational techniques and technology.

In this project, a study was done specifically to measure how effective these features are in identifying similar speaker voice and content similarity. It was concluded that with the distance metrics of Euclidean distance and Cosine similarity, these features identified similar speaker voice, but could not identify content similarity significantly. It was also concluded that deep learning models are more effective in capturing similarity as when also including features extracted from YAMNet to current model, performance improved in comparison to using short term spectral features.

References

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[2] Y. Peng, C.-W. Ngo, C. Fang, X. Chen, and J. Xiao, "Audio similarity measure by graph modeling and matching," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 603–606. [Online]. Available: https://doi.org/10.1145/1180639.1180763

[3] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 196–200.

[4] K. Seyerlehner, G. Widmer, and P. Knees, "Frame level audio similarity-a codebook approach," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08)*, 2008, p. 31.

[5] P. Avgoustinakis, G. Kordopatis-Zilos, S. Papadopoulos, A. L. Symeonidis, and I. Kompatsiaris, "Audio-based near-duplicate video retrieval with audio similarity learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5828–5835.

[6] X. Yu, X. Pan, W. Yang, W. Wan, and J. Zhang, "Audio similarity measure based on renyi's quadratic entropy," in *2010 International Conference on Audio, Language and Image Processing*, 2010, pp. 722–726.

[7] Q. Wu, X. Zhang, P. Lv, and J. Wu, "Perceptual similarity between audio clips and feature selection for its measurement," in *2012 8th International Symposium on Chinese Spoken Language Processing*, 2012, pp. 387–391.

[8] M. K. Singh, N. Singh, and A. K. Singh, "Speaker's voice characteristics and similarity measurement using euclidean distances," in *2019 International Conference on Signal Processing and Communication (ICSC)*, 2019, pp. 317–322.

[9] A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre, "Similarity metric based on siamese neural networks for voice casting," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6585–6589.

[10] L. Lu and A. Hanjalic, "Unsupervised anchor space generation for similarity measurement of general audio," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 53–56.

[11] J. H. Jensen, M. G. Christensen, and S. H. Jensen, "A framework for analysis of music similarity measures," in *2007 15th European Signal Processing Conference*. IEEE, 2007, pp. 926–930.

[12] Z. Li and P. Song, "Audio similarity detection algorithm based on siamese lstm network," in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2021, pp. 182–186.

[13] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "A novel metric for evaluating audio caption similarity," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.

[15] C. Malmberg, "Real-time audio classification onan edge device: Using yamnet and tensorflow lite," 2021.

[16] T. Kinnunen, M. Sahidullah, E. Héctor Delgado, E. Massimiliano Todisco, E. Nicholas Evans, J. Yamagishi, and K. A. Lee, "The 2nd automatic speaker verification spoofing and countermeasures challenge (asvspoof 2017) database, version 2," 2018.