Age Classification Based on Voice Using Mel-Spectrogram and MFCC

Tariq AL-Maashani
Graduate School of Science and Tech.
Kumamoto University
Kumamoto, Japan
tariqalmaashani8@gmail.com

Israel Mendonça
Faculty of Advanced Science and Tech.
Kumamoto University
Kumamoto, Japan
israel@cs.kumamoto-u.ac.jp

Masayoshi Aritsugi
Faculty of Advanced Science and Tech.
Kumamoto University
Kumamoto, Japan
aritsugi@cs.kumamoto-u.ac.jp

***Abstract*—Analyzing speech signals to learn the age behind the voice can be valuable for different purposes including voice authentication or advertisements formulated on age groups. Applying diverse machine learning algorithms to reveal the main features behind age can be challenging since it requires a deep understanding of voice aging; thus, to solve this problem, we developed two methods that focus on classifying a person's age into a group (teens, twenties, thirties, forties, fifties, and sixties) based on his/her voice. The first method was using the Mel-Spectrograms images and feeding them into a Convolutional Neural Network (CNN) model as an image classification task. The second method was extracting important acoustic features including Mel-Frequency-Cepstral-Coefficients (MFCCs), and other features like, Spectral Contrast, Spectral Roll-Off, and Spectral Bandwidth. Then, classifying those extracted features using different machine learning algorithms namely K-Nearest Neighbors (KNN) and Label Propagation (LP). They achieved an accuracy of 95%. Finally, a combination of the two methods was implemented to enhance the model's robustness. We were able to attain an overall accuracy of 97% which is the highest in the literature for the age classification task on the Mozilla Common Voice dataset.**

***Index Terms*—CNN, MFCC, KNN, LP, Spectral Contrast, Spectral Roll-Off, Spectral Bandwidth.**

I. Introduction

Speech is a mean of interaction between individuals or a group of people [1]. The usability of speech as a way of verification is becoming more important these days, regardless of business advertisements, and demographic awareness. However, interpreting audio classification from the perspective of machine learning and computer vision is challenging since it requires preprocessing steps which are essential to perform the classification task. Biologically, voice is generated by the combination of the vibration of vocal cords, breathing process, and resonance [2]. Similarly, a voice could be defined as an analog signal that is converted into a digital signal by using a recorder and saved in audio formats such as mp3, or wav; thus, we can process it.

In this study, we proposed two models to solve the age classification issue by dividing ages into six groups (teens, twenties, thirties, forties, fifties, and sixties):

- The log mel-spectrograms are saved as images and fed into the CNN to solve the classification task.
- A set of features were extracted for each audio wav file. Those features are 46 features per file, and 40 of them are called Mel-Frequency-Cepstral-Coefficients (MFCCs).
- Lastly, we combined the two methods and tested on the same test set.

The remainder of this paper is structured as follows: Section II provides related work. Section III introduces the suggested methodology. Section IV explains the performed experiments. Section V presents results and discussion. Finally, Section VI represents the conclusion and future work.

II. Related work

Multiple techniques were applied to determine the age of a person from his/her voice. A study was conducted by Tursunov et al. [3], in which they proposed a Convolutional Neural Network (CNN) model with a specialty-designed Multi-Attention Module (MAM). They tried to classify age and gender using speech spectrograms, and evaluated their proposed model using the Common Voice dataset. Their average result was 72% for classifying age into a group of six (teens, twenties,.., sixties). We have used this study as a baseline named "MAM" in Results section. In another study demonstrated by Kuchebo et al. [4], they designed a CNN to classify the gender first, and then, based on those results, they classified the age. They have used the mel-spectrogram as an input for the CNN and Mozilla Common Voice as a dataset. Their results were 43.46% for male-age prediction and 43.18% for female-age prediction. This study was used as a second baseline named "CNNER" in Results section. Another study was conveyed by Koli [5], in which he used five different pre-trained CNN models (Xception, inception-V3, VGG-16, VGG-19, and ResNet50). The best result for classifying the age from the voice was 22% for the Mozilla Common Voice dataset by VGG-19, and 52% for the Speech Accent Archive dataset by VGG-16.

A different study performed by Zaman et al. [1] focused on classifying age from voice based on 20 statistical features that were analyzed with Frequency Spectrum Analysis (FSA). Those features were chosen as they claimed that they can capture a good amount of information about the signal like, Fundamental Frequency (F0), Spectral Flatness, and median frequency. They used the Mozilla Common Voice dataset to assess their results and they used different ten prediction

979-8-3503-3959-8/23/$31.00

models including: Random Forest, and CatBoost. However, they grouped the age into three groups: "Young", "Adult", and "Old". Their best result for classifying age was 70.4% which was obtained by the Random Forest algorithm. In our study, we highlight the important role of using MFCCs which contain valuable information regards hoarseness and tonal information. Moreover, mel-spectrograms might increase the performance better than normal spectrograms since they mimic the way how human ears perceive sounds; thus, we take this into consideration and show how much it differs from the existing literature.

III. Methodology

Our solution to classify age from voice consists of two different approaches. The first one is to use a set of 46 features that can be extracted from every audio file. The second approach is to obtain the log mel-spectrogram and solve audio classification as an image classification task while taking advantage of CNN models that outperform many techniques in terms of classification.

A. Age Classification using log mel-spectrograms and CNN

We have used a Convolutional Neural Network (CNN) model to tackle the classification task. Nevertheless, some steps are required to transform the wav files into images. Log mel-spectrogram is a favorable representation for audio classification since it can transcribe human auditory perception. To generate the log mel-spectrogram, the following processes took place:

1) A signal was sampled at a rate of 48 kHz was used, and a Hanning window was applied with a time frame of 43 ms. Furthermore, the hop length was 512, which means an overlap of 75% with the default frame length of 2048.
2) Discrete Fourier Transform is applied to each frame separately, and by doing so, we replaced the amplitude with frequency while maintaining the time on the x-axis.
3) Applying the Mel-spaced filter bank to the transformed signal. Mel is a scale based on the way how human ears perceive sound. In point to the fact that Mel-scale is logarithmic above 1 kHz, and approximately linear below that [6]. In other words, humans are able to distinguish changes in pitch at all frequencies; nevertheless, in higher frequencies, the distance in frequencies needs to be bigger. In order to convert the frequency to Mel-scale, the following equation by [7] is applied:

$$mel = 2595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

4) The mel-spectrogram can be generated by applying a dot product between every window and its corresponding mel; thus, a mel-spectrogram is a spectrogram with mel-bins at the y-axis rather than frequencies.
5) Converting the mel-spectrogram from power to decibel is required to obtain log mel-spectrogram.

B. Age Classification using a set of 46 features

Every audio file was analyzed and we extracted a set of features. We chose these features based on their contribution to detecting age information. Moreover, they represented the vocal cord change as we get old. Furthermore, the number of features was selected based on the higher performance achieved in this study. These features are explained as follows:

- MFCCs: a set of coefficients obtained after applying the Discrete Cosine Transform (DCT). MFCCs are a satisfactory set of features that represents the whole signal by a small amount of coefficients. Nevertheless, the process of extracting them contains the same steps 1-3 from the previous Sub-Section. An additional step is added, which is applying DCT to transform mel-frequency coefficients to cepstral coefficients, and a total of 40 coefficients are extracted by the end of this step.
- Spectral Centroid: the center mass of the spectrum [8].
- Spectral Bandwidth: the difference in frequencies (Upper Frequency - Lower Frequency) at a specific time frame.
- Spectral Contrast: the difference in decibels between valleys and peaks in the spectrum [9]. High contrast generally indicates a clear signal; however, low contrast means broad-band noise [10].
- Spectral Flatness- or tonality coefficient: used to analyze how much a sound is similar to tone-like more than noise-like [11].
- Spectral Roll-Off: a predefined percentage where the concentration of the magnitude distribution of the spectrum is below [8]. In this study, we have used 85% as a roll percentage.
- Zero Crossing Rate: a rate at which a signal crosses the zero in the time domain within one second [12], or it could be defined as the number of times a signal is changing from negative to positive, and vice versa, divided by the frame length [8].

IV. Experiment

In this study, we used the Mozilla Common Voice dataset [13]. This dataset is open source and it is powered by anonymous contributors that can record themselves through the Mozilla website and participate in the verification of the validity of recordings as well. In addition, Common Voice is multilingual and it supports over a hundred languages; nonetheless, we used the English language subset. It contains validated (two up-votes), invalidated (two down-votes), and other recordings (up-votes = down-votes). Moreover, the valid clips and other clips are divided into three subsets: train, dev, and test. We have used the valid clips which are already split for training, development, and testing. Each subset contains a .csv metadata file that contains information about audio clips such as, filename, age, gender, accent, and clip duration. Furthermore, the ages are divided into nine groups, each group for a decade as it was stated in the dataset documentation. A notable point of this dataset is that a speaker who appears in the training subset, he/she is not included in the other two

subsets as well as for the text. Furthermore, only the first six groups were used and some clips were not labeled with age; hence, they were eliminated. A total of 74885 audio clips in mp3 format were extracted from this dataset as shown in Table I.

TABLE I
MOZILLA COMMON VOICE DATASET - AGE DISTRIBUTION AMONG TRAIN, DEV AND TEST SETS

Age Group (labels)	**Range** (years)	**Train**	**Dev**	**Test**
Teens	<19	5441	113	117
Twenties	19 -29	23003	487	466
Thirties	30-39	18303	345	389
Forties	40-49	11100	244	236
Fifties	50-59	9466	203	205
Sixties	60-69	4584	95	88
Total		71897	1487	1501

Only the filename and the corresponding age group were saved. After that, audio clips were converted from mp3 to wav format using Pydub [14] and FFmpeg [15]. Afterward, data cleaning was performed by removing any silence from audio files using Librosa [16], and Noise Reduce [17] with a predefined threshold. In the last step of the preprocessing, the wav files were mapped with their age groups, and new three csv files were generated for train, dev, and test subsets, correspondingly.

For age classification using log mel-spectrograms and CNN, audio files were transformed into log mel-spectrograms and saved as PNG images with the size of (256×256) pixels due to model performance. We used the model builder provided by PyTorch [18] to instantiate an EfficientNet_V2_M model [19]. Furthermore, we changed the classifier layer to be of dimension (1280×6), while maintaining the dropout layer with a probability of 0.3 to prevent the over-fitting problem. Additionally, the model was trained with 30 epochs with a batch size of 8, a learning rate of 0.0001, CrossEntropyLoss as a loss function, and Adam as an optimizer. For age classification using a set of 46 features, all acoustic features were normalized and saved into three different csv files with their corresponding age group for train, dev, and test subsets. We used different machine learning algorithms using LazyPredict which contains 28 classifiers. We suggested enhancing the performance by combining the predictions of the two methods and taking advantage of their diversity. We used a weighted average of predictions as in the equation follows:

$$Predictions = \alpha \times A + \beta \times B \tag{2}$$

where: α = scale, A = CNN model predictions, β = scale, and B = 46-features model predictions. The values of α and β are chosen from a set of nine values based on the achieved performance.

V. RESULTS AND DISCUSSION

We tested the robustness of our suggested models separately on the test set provided by the Mozilla Common Voice dataset, and then join both models at the end. We compared all results with the MAM [3], and CNNER [4] which did not provide results per decade. We have used the test set provided by the Mozilla Common Voice team for two reasons: 1) to keep the comparison consistent by using the same test set. 2) All speakers and texts in the test set are not included in the train, and validation sets which ensure the generalization of the suggested models.

A. Results using 46 Features

We ran different experiments by using Label Propagation (LP), and K-Nearest Neighbors (KNN), as shown in Table II. Based on the results, Label Propagation attained the best performance using 46 features. Furthermore, Fig. 1 shows the confusion matrix for age classification using the Label Propagation classifier. The best-predicted group was the "Sixties" group with 88 correct out of 88, and the worst was the "Teens" group with 107 correct out of 117.

TABLE II
TEST RESULTS FOR AGE CLASSIFICATION WITH 46 FEATURES USING KNN, AND LP ON COMMON VOICE DATASET INCLUDING F1-SCORE, AND ACCURACY COMPARED WITH BASELINES. BOLD NUMBERS INDICATE TOP RESULTS FOR EVERY AGE GROUP

Age Group	F1-Score				**Speakers**
	CNNER	**MAM**	**KNN**	**LP**	
Teens	-	0.58	0.92	**0.93**	117
Twenties	-	0.73	0.93	**0.95**	466
Thirties	-	0.7	0.93	**0.95**	389
Forties	-	0.7	0.92	**0.94**	236
Fifties	-	0.81	0.93	**0.95**	205
Sixties	-	0.76	0.94	**0.96**	88
Accuracy	0.43	0.72	0.93	**0.95**	
Weighted Accuracy	-	0.72	0.93	**0.95**	1501
Unweighted Accuracy	-	0.71	0.93	**0.95**	

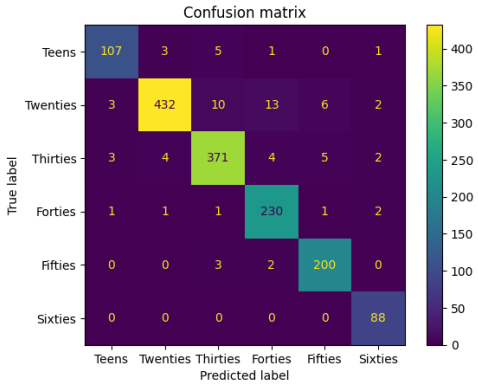

Fig. 1. Confusion matrix attained by LP for age classification using 46 features

B. Results using EfficientNet_v2_m

Alternatively, we tested the EfficientNet_v2_m model by using the log mel-Spectrograms. We ran ten different experiments using different seeds (0, 1, 2, 3, 5, 7, 10, 20, 42, 100)

to ensure the random initialization for the CNN. Our proposed model results in an overall accuracy of 95%. Table III shows the results, and Fig. 2 depicts the confusion matrix.

TABLE III
TEST RESULTS FOR AGE CLASSIFICATION WITH CNN USING EFFICIENTNET_V2_M (BEST, AVERAGE (AVG) AND STANDARD DEVIATION (STD)) INCLUDING F1-SCORE, AND ACCURACY, WEIGHTED ACCURACY, AND UNWEIGHTED ACCURACY COMPARED WITH BASELINES. BOLD NUMBERS INDICATE TOP RESULTS FOR EVERY AGE GROUP

Age Group	F1-Score					**Speakers**
	CNNER	**MAM**	EfficientNet_v2_m			
			BEST	**AVG**	**StD**	
Teens	-	0.58	**0.95**	0.90	0.02	117
Twenties	-	0.73	**0.95**	0.95	0.01	466
Thirties	-	0.7	**0.96**	0.94	0.01	389
Forties	-	0.7	**0.96**	0.94	0.01	236
Fifties	-	0.81	**0.97**	0.96	0.01	205
Sixties	-	0.76	**0.99**	0.97	0.01	88
Accuracy	0.43	0.72	**0.96**	0.94	0.01	
Weighted Accuracy	-	0.72	**0.96**	0.94	0.01	1501
Unweighted Accuracy	-	0.72	**0.96**	0.94	0.01	

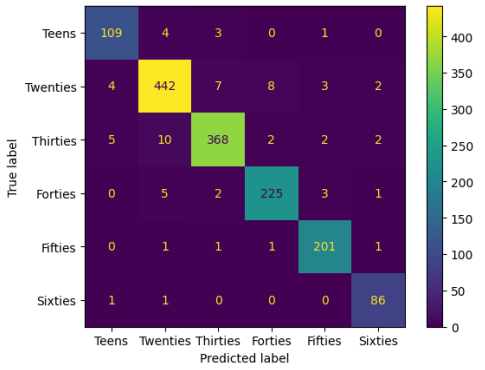

Fig. 2. Confusion Matrix achieved by our model among 10 experiments

C. Results by Combining the two models

A combination of the two proposed models was joined to elaborate our solution to the classification task. We ran experiments by using different scales including [0.1,0.2,...,0.9] to determine values of α and β from equation 2. An improvement was achieved in terms of all age groups by using α = 0.4, and β = 0.6. Results for our Model are shown in Table IV. Fig. 3 depicts the corresponding confusion matrix.

Our proposed model attained better results compared with the baselines conducted by Tursunov et al. [3] and Kuchebo et al. [4], and there are many reasons behind that. First, we used a complex convolutional neural network architecture which required more computation resources. Thus, the model generalizability outperforms the baselines. In addition, we chose our hyper-parameters carefully among a set of different ones. The results of such parameters were not included in this study due to paper length limitations. As well, a deep study was conducted on feature engineering for both models including the number of features which seems to give a better performance by selecting them and discarding others, and the quality of images by choosing the right dimensions and applying well preprocessing steps. Nevertheless, acoustic features that might work with a specific recognition system might not be applicable to a different one [20].

TABLE IV
TEST RESULTS FOR AGE CLASSIFICATION USING OUR MODEL INCLUDING F1-SCORE, ACCURACY, WEIGHTED ACCURACY, AND UNWEIGHTED ACCURACY COMPARED WITH BASELINES. BOLD NUMBERS INDICATE TOP RESULTS FOR EVERY AGE GROUP

Age Group	F1-Score			**Speakers**
	CNNER	**MAM**	**Our Model**	
Teens	-	0.58	**0.93**	117
Twenties	-	0.73	**0.97**	466
Thirties	-	0.7	**0.98**	389
Forties	-	0.7	**0.97**	236
Fifties	-	0.81	**0.98**	205
Sixties	-	0.76	**0.99**	88
Accuracy	0.43	0.72	**0.97**	
Weighted Accuracy	-	0.72	**0.97**	1501
Unweighted Accuracy	-	0.72	**0.97**	

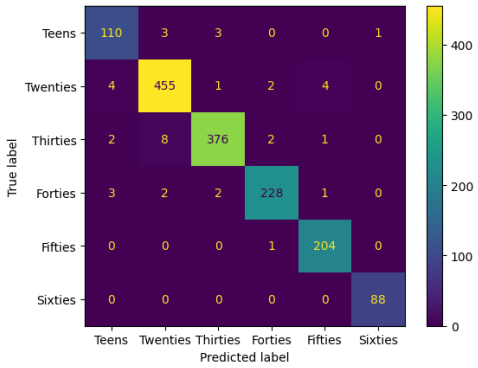

Fig. 3. Confusion Matrix attained by our Model

VI. CONCLUSION

In this study, we proposed two approaches to classify age from voice. The first method was to extract vital acoustic features which carry information about age. Our method showed a high performance with an average accuracy of 95%. The second approach was to generate log-mel-spectrograms which is an extension of spectrograms with mel-scale instead of frequencies, and the third dimension is decibels instead of power. By using this approach and taking advantage of CNN architecture, we were able to have an improvement with an average accuracy of 95% compared with the baselines [3], and [4] where their average accuracy was 72%, and 43%, respectively. Lastly, a combination of the two methods was carried out which yielded an average accuracy of 97%. Our future work will include applying such approaches to classify different demographic parameters; for example emotion, and accent. We are planning to use the RAVDESS dataset for classifying emotions [21].

REFERENCES

[1] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: Gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 338–343. [Online]. Available: http://doi.org/10.1109/COMPSAC51774.2021.00055

[2] K. Makiyama, S. Hirano *et al.*, *Aging voice*. Springer, 2017.

[3] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, 2021. [Online]. Available: http://doi.org/10.3390/s21175892

[4] A. V. Kuchebo, V. V. Bazanov, I. Kondratev, and A. M. Kataeva, "Convolution neural network efficiency research in gender and age classification from speech," in *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2021, pp. 2145–2149. [Online]. Available: http://doi.org/10.1109/ElConRus51938.2021.9396365

[5] R. N. Koli, "Classification of speaker's age, gender and nationality using transfer learning," Master's thesis, Dublin, National College of Ireland, 2021. [Online]. Available: https://norma.ncirl.ie/5177/1/rohannarayankoli.pdf

[6] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937. [Online]. Available: https://asa.scitation.org/doi/pdf/10.1121/1.1915893

[7] D. O'Shaughnessy, *Speech Communication: Human and Machine*, ser. Addison-Wesley series in electrical engineering. Addison-Wesley Publishing Company, 1987. [Online]. Available: https://books.google.co.jp/books?id=mHFQAAAAMAAJ

[8] T. Giannakopoulos and A. Pikrakis, "Chapter 4 - audio features," in *Introduction to Audio Analysis*, T. Giannakopoulos and A. Pikrakis, Eds. Oxford: Academic Press, 2014, pp. 59–103. [Online]. Available: https://doi.org/10.1016/B978-0-08-099388-1.00004-2

[9] J. Yang, F.-L. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Communication*, vol. 39, no. 1, pp. 33–46, 2003. [Online]. Available: https://doi.org/10.1016/S0167-6393(02)00057-2

[10] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, 2002, pp. 113–116 vol.1. [Online]. Available: https://doi.org/10.1109/ICME.2002.1035731

[11] S. Dubnov, "Generalization of spectral flatness measure for non-gaussian linear processes," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 698–701, 2004. [Online]. Available: https://doi.org/10.1109/LSP.2004.831663

[12] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Chapter 3 - features for content-based audio retrieval," in *Advances in Computers: Improving the Web*, ser. Advances in Computers. Elsevier, 2010, vol. 78, pp. 71–150. [Online]. Available: https://doi.org/10.1016/S0065-2458(10)78003-7

[13] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019, https://www.kaggle.com/datasets/mozillaorg/common-voice, Accessed online [2022-09-13].

[14] J. Robert, M. Webbie *et al.*, "Pydub," 2018. [Online]. Available: http://pydub.com/

[15] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[17] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[19] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106. [Online]. Available: https://arxiv.org/abs/2104.00298

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980. [Online]. Available: http://doi.org/10.1109/TASSP.1980.1163420

[21] B. Vimal, M. Surya, Darshan, V. Sridhar, and A. Ashok, "Mfcc based audio classification using machine learning," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–4. [Online]. Available: http://doi.org/10.1109/ICCCNT51525.2021.9579881