biomimetics

MDPI

Article

Assessment of Pepper Robot's Speech Recognition System through the Lens of Machine Learning

Akshara Pande and Deepti Mishra *

Educational Technology Laboratory, Intelligent System and Analytics Group, Department of Computer Science (IDI), Norwegian University of Science and Technology, 2815 Gjøvik, Norway; akshara.pande@ntnu.no
* Correspondence: deepti.mishra@ntnu.no

Abstract: Speech comprehension can be challenging due to multiple factors, causing inconvenience for both the speaker and the listener. In such situations, using a humanoid robot, Pepper, can be beneficial as it can display the corresponding text on its screen. However, prior to that, it is essential to carefully assess the accuracy of the audio recordings captured by Pepper. Therefore, in this study, an experiment is conducted with eight participants with the primary objective of examining Pepper's speech recognition system with the help of audio features such as Mel-Frequency Cepstral Coefficients, spectral centroid, spectral flatness, the Zero-Crossing Rate, pitch, and energy. Furthermore, the K-means algorithm was employed to create clusters based on these features with the aim of selecting the most suitable cluster with the help of the speech-to-text conversion tool Whisper. The selection of the best cluster is accomplished by finding the maximum accuracy data points lying in a cluster. A criterion of discarding data points with values of WER above 0.3 is imposed to achieve this. The findings of this study suggest that a distance of up to one meter from the humanoid robot Pepper is suitable for capturing the best speech recordings. In contrast, age and gender do not influence the accuracy of recorded speech. The proposed system will provide a significant strength in settings where subtitles are required to improve the comprehension of spoken statements.

Keywords: Pepper robot; speech recognition; audio features; evaluation metrics; K-means clustering

Citation: Pande, A.; Mishra, D. Assessment of Pepper Robot's Speech Recognition System through the Lens of Machine Learning. *Biomimetics* **2024**, *9*, 391. https://doi.org/10.3390/biomimetics9070391

Academic Editors: Weiwei Wan, Yisheng Guan and Li He

Received: 24 April 2024
Revised: 5 June 2024
Accepted: 24 June 2024
Published: 27 June 2024

1. Introduction

With the advancement of technology, social robots have become more capable, revolutionizing human lives. They have exceeded their roles and become vital societal partners, including in workplaces and households. These machines are equipped with sensors that help them perceive their environment and make them proficient in social interactions. They can participate in natural and meaningful conversations, understand context, and respond accordingly. This enriched human–robot interaction is transforming various domains, such as healthcare, education, and customer service, where intelligent assistance is required. Speech is one of the fields in which intelligent assistance is highly demanded. However, the complete understanding of speech can be challenging due to various reasons. New language learners face one common problem in that they may struggle to comprehend when native speakers communicate fluently. Additionally, online meetings can sometimes make it hard to thoroughly grasp the speaker's message. In crowded settings, such as large gatherings, background noise can make it problematic for people to hear one another. Moreover, hearing impairments can also affect voice understanding. Another method of communication could be through texts; however, it can sometimes be difficult to see texts properly due to several reasons, such as long distances and dim-light environments. One potential solution to these issues is to combine speech and subtitles together since only displaying subtitles is not enough due to the possibility of ambiguity in written words. By using this multimodal technique, which combines both speech and text, comprehension can be improved and cognitive burdens can be reduced. In such scenarios, integrating

the humanoid robot Pepper can be an effective solution as Pepper has speech recognition capability and a display screen. Social robots are popular for establishing human–robot interaction (HRI) in various domains such as healthcare, education, and the service industry. The purpose of robots in healthcare can be related to the engagement of older persons and identifying their stress [1], as well as for the therapy of autistic children [2]. Abdollahi et al. [1] utilized the social robot Ryan for the emotional recognition of elders and employed two modalities for this purpose: facial expression and speech sentiment. Further, they prepared the robot to respond accordingly and demonstrated that this interaction had a positive effect. The involvement of social robots in the education field could enhance students' engagement and learning [3,4]. Past studies showed the employment of social robots in the hotel industry for interaction purposes [5] and as receptionists [6]. The speech recognition of social robots plays an important role in establishing proper interaction between humans and robots. Pepper, developed by Softbank Robotics, is an outstanding example of a social robot (https://www.aldebaran.com/en/pepper, (accessed on 20 August 2023)). It is famous for its communicating expertise by comprehending and responding to spoken language. The speech recognition functionality provides various advantages to Pepper, such as voice recognition [7], multilingual support [8], context awareness [9], and speech recording [10]. However, like other speech recognition systems, Pepper's system also struggles to capture speech accurately due to various factors. Furthermore, Pepper is unable to transcribe spoken speech into text format on its own. Thus, it is crucial to incorporate speech-to-text conversion tools for the clarity and understanding of spoken statements.

Speech-to-text technology is very popular and is utilized in many fields, such as healthcare [11], schools [12], and the service sector [13]. Speech recognition can be beneficial for documentation purposes in healthcare settings [14]. A combination of audio-visual content and speech recognition is an effective way to deliver educational content [15]. The speech recognition system of the Pepper robot cannot perform speech-to-text conversion. As per our previous research [10], the potential solution to this issue is to incorporate the speech-to-text conversion tool Whisper into the Pepper robot. It could be advantageous in educational environments to display the text generated by Whisper on Pepper's screen. Continuing our previous research, we aim to explore the complexities of Pepper's speech recognition system more deeply by analyzing audio features. Further, the goal is to create clusters based on these features and select the best cluster by evaluating the Whisper-generated text accuracy. Previous studies showed that a combination of audio features along with machine learning applications provides better performance of speech recognition [16–18]. To the best of our knowledge, machine learning incorporated with Pepper's speech recognition system has not yet been explored. The research questions formulated for the present investigation are as follows:

RQ1: What could be the possible solution to evaluate the performance of Pepper's speech recognition system when both audio and spatial features are considered?

RQ2: Which spatial features, such as age, gender, and distance from Pepper, affect the performance of the speech recognition system?

In the present study, the objective is to combine various audio and spatial features of the recorded speech to better understand the efficacy of the speech recognition system. We selected audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, spectral centroid, spectral flatness, energy, and the Zero-Crossing Rate (ZCR) to analyze the recorded speech by the Pepper robot. These features are selected in our study because they play crucial roles in different tasks related to audio processing and may contribute to providing meaningful information, such as spectral patterns, frequency, and loudness, as well as being helpful in forming clusters. MFCCs are commonly utilized as an efficient feature for audio processing as they help in various tasks such as speech recognition [19,20] and speaker identification [21,22]. Furthermore, they are also a useful indicator of the human auditory system [23]. The spectral centroid is another popular method for speech processing [24] and is mostly used for classification purposes such as genre classification [25] and mood classification [26]. The spectral flatness feature offers digital signal

processing [27] and primarily determines noise-like characteristics [28]. The Zero-Crossing Rate is one of the important indicators of sound signals, and it is useful for determining the dominant frequency [29]. Furthermore, it is widely applied to various audio-processing domains, including speech/music discrimination [30,31], speech analysis for emotion recognition [32], classification of music genre, and singer identification [33]. Audio feature pitch is a valuable tool for the differentiation of various tones present in tonal languages [34]. Moreover, it can also facilitate the detection of speakers [35] and emotions [36]. Energy is an essential parameter of audio processing, which helps in distinguishing emotions and plays an important role in influencing the efficacy of acoustic models for speech recognition purposes [37]. The effectiveness of these features made them suitable for our study. MFCCs can distinguish various spectral patterns, the spectral centroid can differentiate various types of speech, and the spectral flatness helps in separating background noise. With the help of the ZCR, high-frequency components or noisiness can be identified. Pitch and energy are helpful in detecting variations in sound.

Furthermore, we aim to investigate how the above-mentioned audio features are related to spatial features such as distance and demographics such as age and gender. Moreover, the goal is to assess Pepper's speech recognition system performance with the help of clustering audio features and evaluation measures of speech. In order to achieve these goals, the pipeline shown in Figure 1 was followed. In Figure 1, the overview of the experiment and analysis leading to the result has been demonstrated. Firstly, the experiment was conducted to collect the audio data in the form of recordings using Peppers' sensor. Due to the absence of advanced AI capabilities, Pepper is unable to perform tasks such as speech-to-text conversion and the application of machine learning (ML) algorithms on its own. Hence, these recordings were transferred to the local system, and with the help of Whisper, the corresponding text files for these recordings were generated. In addition to that, audio features were extracted from collected data recordings in the local system. In the next step, K-means clustering machine learning was opted to group features of speech recordings based on their inherent properties. Furthermore, the best cluster can be selected based on the higher accuracy of predicted speech. The structure of the paper is as follows: The literature review for the present study is presented in Section 2. The detailed methodologies are described in Section 3. Section 4 contains the results obtained from this study. Discussions are presented in Section 5. Finally, conclusions are outlined in Section 6.

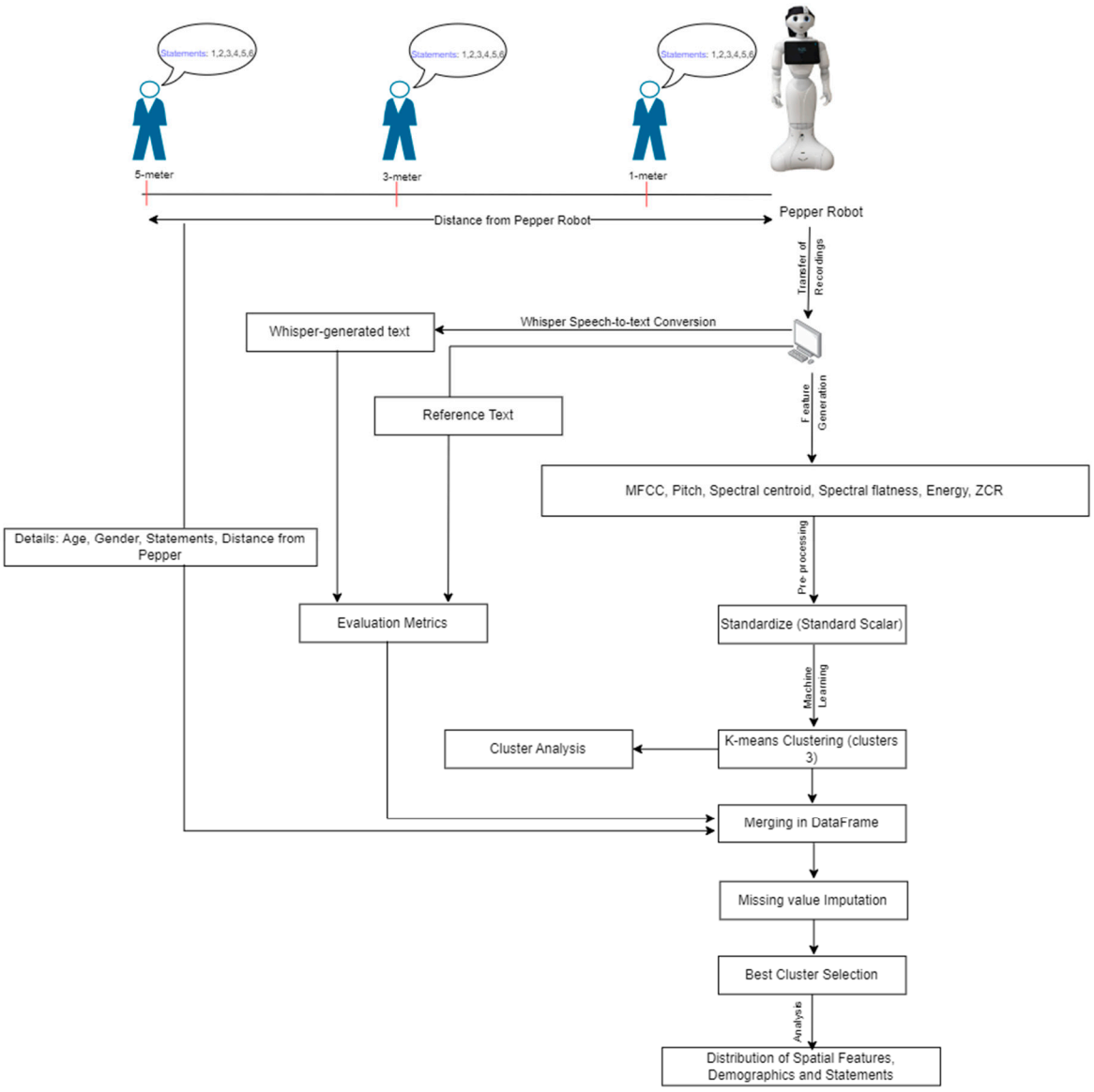


Figure 1. A pipeline of overall work to evaluate the efficiency of Pepper's speech recognition system.

2. Literature Review

One of the main problems usually faced by speech recognition systems is variation in human speech. Accents [38], dialects [39], and speech disorders [40] can influence speech and differences in pronunciation. Different people pronounce the same word in various ways. Feng et al. [41] mentioned that Automatic Speech Recognition (ASR) performance can be influenced by several aspects, such as regional and native accents, age, and gender. Different regions of speakers might have varying speech styles for the same sentence [42]. A person's age could be another factor in producing differences in speech. Karlsson et al. [43] demonstrated that articulation and phonation can be influenced by age. Bóna [44] showed that age can impact speech style. Das et al. [45] conducted empirical studies examining how various speech characteristics and cepstral features change with aging in the context of Bengali vowels. Kennedy et al. [46] pointed out the limited research on children's voice recognition and emphasized the need for further investigation in the human–robot interaction domain. Arslan et al. [47] mentioned that speech disfluency increases with aging. However, their findings suggested that gesture along with speech disfluency was comparable for younger and older adults. Previous studies indicated the incapability of Nao and Pepper robots to understand the voices of children [46,48] and elders [49,50]. Another

aspect influencing speech is a person's gender, as there are differences in pitch, frequency, and vocal tract length between males and females. The findings of Mendoza et al. [51] indicated notable distinctions between genders' breathier quality often observed in female voices. However, Pande et al.'s [52] findings suggest that Pepper's speech recognition system does not detect much difference between male and female voices. Distance and background noise can also inhibit the precision of speech signals. Increasing the distance from the speech recognition system may impact the recognition accuracy and noise [46]. In noisy surroundings, it is difficult to recognize spoken words correctly [53,54]. This variety creates a considerable barrier for any speech recognition system, including Pepper.

The application of machine learning algorithms on speech data, facilitated by the extraction of meaningful features, carries great importance in various fields, such as emotion recognition [55,56], speaker accent identification [57], and speech disorders [58]. Machine learning algorithms such as K-nearest neighbors [59], Convolutional Neural Networks (CNNs) [60,61], and Deep Neural Networks (DNNs) [62] improve the accuracy of speech recognition. Acoustic features provide insights into sound data. Mel-Frequency Cepstral Coefficients (MFCCs), pitch, spectral centroid, spectral flatness, energy, and the Zero-Crossing Rate (ZCR) are essential features among these features. According to Sandhya et al. [63], spectral features include MFCCs, spectral centroid, spectral flatness, Root Mean Square (RMS), and ZCR. Additionally, the research conducted by Micheyl et al. [64] highlights the connection between pitch and fundamental frequency.

The MFCC feature plays an essential role in speech recognition [21,22,65]. Gourisaria et al. [66] applied MFCC and Short-Time Fourier Transform (STFT) to extract audio features and achieved higher accuracy with the Artificial Neural Network (ANN) model. Ittichaichareon et al. [20] showed MFCC feature extraction and the utilization of MFCCs to apply a Support Vector Machine (SVM). Hamza et al. [23] demonstrated that with the help of MFCC features and machine learning techniques such as Random Forest, Decision Tree, and SVM, deepfake audio can be identified. Pitch analysis is essential in identifying speakers [35] and emotion detection [36]. Shagi et al. [67] proposed a machine learning approach for gender identification using pitch features of speech. For this purpose, they used ML methods such as a CNN, the Multilayer Perceptron (MLP), an SVM, and Logistic Regression (LR). The spectral centroid is often employed as a measure of sound signal brightness and timbre in music analysis [68], genre classification [25], and mood classification [26]. Ferdoushi et al. [69] demonstrated that the spectral centroid can effectively distinguish heart sounds with a murmur. The spectral flatness is a metric applied to estimate the level of noise, uniformity, and width [70]. Lazaro et al. [71] used an SVM to classify features like the spectral centroid and spectral flatness for music tempo classification. The Zero-Crossing Rate (ZCR) can be applied to classify percussive sounds and separate voice and unvoice [72]. Panda et al. [73] proposed that a multimodal fusion of features such as MFCCs, the Zero-Crossing Rate (ZCR), and Root Mean Square (RMS) with machine learning approaches XGBoost, SVM, Random Forest, Decision Tree, and KNN to increase the accuracy of emotion detection systems. Paul et al. [74] used feature fusions of MFCCs, the ZCR, pitch, and Linear Predictive Coefficients (LPCs) and applied an SVM, the Decision Tree, and Linear Discriminative Analysis (LDA) for emotion detection. Hammoud et al. [75] utilized ZCR, RMS, and MFCC features with various machine learning approaches, including decision tree, random forest, and KNN, and demonstrated that MFCCs with a Random Forest attained maximum accuracy for infant crying interpretation.

In some studies, unsupervised machine learning was applied to speech data [76,77]. Aldarmaki et al. [78] addressed that ASR with labeled data can be challenging, and they reviewed unsupervised ASR along with its limitations. Esfandian et al. [79] employed Gaussian Mixture Models (GMMs) and Weighted K-means (WKM) clustering techniques to effectively reduce the feature space dimensions for speech recognition in the spectro-temporal domain. Hajarolasvadi et al. [80] used K-mean clustering and spectrograms for 3D CNN-based speech recognition. Vyas et al. [81] showed that MFCCs and the K-means algorithm were helpful in detecting mood from Indian music. Bansal et al. [82] applied

K-means clustering on MFCCs and Linear Predictive Cepstral Coefficients (LPCCs) to classify emotional Hindi speech. Marupaka et al. [83] utilized MFCCs, ZCR, the Dynamic Energy Ratio (DER), and Cyclostationary features for the classification of sound signals using K-means clustering. Poorna et al. [84] used Hybrid Rule-based K-means clustering and multiclass SVM for emotion recognition and demonstrated that Hybrid-Rule-based K-means clustering performed better than the multiclass SVM. However, no known studies have investigated the use of machine learning algorithms in conjunction with the speech recognition system of the Pepper robot.

Apart from applying machine learning, the other characteristic that can provide ideas about the correctness of recorded speech is speech-to-text technology. Speech-to-text technology has been used in various domains. Some reports focus on speech-to-text utilization in education and learning [85]. Debnath et al. [15] demonstrated that audio-visual content combined with automatic speech recognition can efficiently deliver educational content to individuals with disabilities. In another study, Goss et al. [14] revealed, through a survey, widespread belief in the usefulness of speech recognition technology for clinical documentation while also acknowledging the challenges associated with implementing it. OpenAI (San Francisco, CA, USA) has developed a speech-to-text conversion tool called Whisper [45], which efficiently recognizes speech in several languages and executes numerous tasks successfully. Macháček et al. [86] employed Whisper for the transcription of real-time speech. Vásquez-Correa et al. [87] revealed that with the utilization of synthetic data, Whisper-based Automatic Speech Recognition leads to performance enhancement in certain areas. Spiller et al. [88] noted that incorporating Whisper into audio transcription for mental health research can simplify data analysis.

3. Methodology

Experiments were conducted in a controlled closed-room laboratory environment in the Educational Technology Laboratory, NTNU Gjøvik, to test Pepper's speech recognition system. A laptop with an 11th Gen Intel(R) Core (TM) i5-1145G7 @ 2.60 GHz processor, 16 GB RAM, and running the Windows 10 operating system was used. The experiment was carried out using Python version 2.7.16 and Python version 3.9.13.

3.1. Experimental Setup

Eight participants, including three males and five females aged 15–55, were recruited to record six randomly selected statements from Pepper at three distances (1 m, 3 m, and 5 m). The sample included diverse educational backgrounds, including Bachelor's, Master's, and Ph.D. degrees. It should be noted that none of the participants spoke English as their first language, and they belong to Asian and European regions. Pepper was used to record statements. There were 18 recordings available for each person, making 144 recordings for eight people.

3.2. Pepper Robot Function

In this study, the Pepper robot was used to record and save participants' speech within its system using the ALAudioRecorder service (http://doc.aldebaran.com/2-5/naoqi/audio/alaudiorecorder.html (accessed on 20 August 2023)). This service allowed the robot to use its microphones to record audio, and the recording process was initiated and terminated using the "startMicrophoneRecording" and "stopMicrophoneRecording" methods, respectively. The ".wav" extension was used to save the audio files.

3.3. Transfer of Recordings to the Local System

Paramiko is a Python library that allows secure communication with remote servers through the SSH protocol (https://www.paramiko.org/ (accessed on 20 August 2023)). With Paramiko, an encrypted connection can be established to a remote server and authenticated using a private key or password. It also makes it simple to perform secure file transfers between the Pepper robot and the computer.

3.4. Feature Generation

Feature generation plays an important role in analyzing audio recordings. The audio characteristics can be gained through these features, which can further help to extract meaningful information and patterns from the data. In this study, the features were extracted using the Python library 'Librosa'. The features extracted are outlined in the following subsections.

3.4.1. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are the coefficients that create the Mel-frequency cepstrum (MFC) [65]. The MFC demonstrates the short-term power spectrum of sound. MFCCs can be generated by employing various mathematical operations on the power spectrum of a signal. MFCCs are mainly used for audio and speech processing work, such as speech recognition, the classification of music genres, etc. The set of MFCCs was computed using the 'librosa.feature.mfcc' function.

This function takes two input parameters: audio waveform data and the sample rate. The number of MFCCs selected for audio analysis was 13, which shows the number of rows. The number of columns corresponds to the time frames, which were 419. Thus, the MFCC array was set to have a specific shape with 13 rows and 419 columns. Other input parameters, such as frame length and frame shifts, were selected by the function by default.

3.4.2. Pitch

Pitch is used to measure the fundamental frequency of sound [89]. It is mostly used for the recognition of normal speech. The 'librosa.yin' function was employed to calculate the pitch of the recorded audio. YIN [90] is a valuable tool for accurately estimating the fundamental frequency of both speech and music signals.

3.4.3. Spectral Centroid

The spectral centroid [91], as the name suggests, is related to the center of mass of the spectrum of sound. Its analysis is related to the texture of a sound. The output of the 'librosa.feature. spectral_centroid' function is a set of values that denote the spectral centroid of audio.

3.4.4. Spectral Flatness

The spectral flatness [27] can be determined by calculating the ratio of the geometric mean to the arithmetic mean of the values within a particular frequency range. It is the feature of noise [92]. A high spectral flatness indicates a uniform power spread, while a low spectral flatness indicates an inconsistent power distribution [70].

3.4.5. Time Domain Features

The Root Mean Square (RMS) and Zero-Crossing Rate (ZCR) are features of the time domain [93]. To calculate the energy present in speech, the Root Mean Square (RMS) value was considered using 'librosa.feature.rms'. The Zero-Crossing Rate (ZCR) measures how many times a waveform crosses the zero axis within a specific amount of time [94]. Energy and the ZCR are both important components for separating voice–unvoice from sound [95].

3.5. Pre-Processing of Dataset

Pre-processing of the dataset is an important step so that the data are clean enough before applying the machine learning algorithm. All features may be present in different scales, and they should be on some common scale. For this purpose, standardization is used. Standardization is a process used on datasets to convert the features into a distribution that resembles a normal curve containing a mean of 0 and a standard deviation of 1. Scikit-learn [96] comes equipped with a module called StandardScaler that carries out data standardization.

3.6. Application of Machine Learning

There exist many unsupervised algorithms; however, we selected the K-means algorithm for data analysis because its computational complexity is low and it is efficient as well as simple to implement [97]. Abdalla et al. [98] compared the K-mean algorithm with agglomerative hierarchical clustering (AHC) for small datasets and found that K-means clustering had better performance than AHC for purity and entropy when the cosine similarity measure was applied. However, when the Euclidean distance was employed, AHC performed better than K-means. Rathore et al. [99] showed that in the context of performance, K-means is better suited than hierarchical methods for document clustering. Similarly, Abbas et al. [100] also demonstrated that K-means excels over hierarchical clustering and provides good results for large datasets.

K-means [101] is an unsupervised algorithm that is data-driven as it identifies clusters based on information present in the dataset. K-means clustering [102] is a well-accepted and widely used [97] unsupervised machine learning algorithm. It is an appropriate selection for acquiring knowledge of the intrinsic properties within data and discovering hidden trends and associations. The K-means clustering algorithm is mostly utilized to organize large datasets [103]. The essential aim of K-means is to divide a dataset into unique groups, or clusters, based on the similar properties of data points within each cluster. By utilizing K-means clustering, it is possible to determine which spatial features belong to the same group. We used the elbow method to select the optimal number of clusters. The elbow method helps in identifying the elbow by plotting the sum of squares (SSEs) against different cluster numbers (k). Further, the K-mean algorithm is fitted on an optimal number of clusters, and the output assigns a cluster label to each data point. The clustering will help in identifying patterns of audio features present in different groups.

3.7. Integration with Whisper: Performance Evaluation Using WER, MER, WIL, and CER

Whisper [104] is a versatile open-source speech recognition model that has been trained on a comprehensive audio dataset. Whisper can be used to execute several tasks, including multilingual speech recognition and translation. The audio recordings, which are saved in the Pepper robot, can be transferred to the computer system with the help of the Paramiko library, where Whisper can be used to convert them into text. A Python version, Python 3.9.13, was employed for audio-to-text conversion by Whisper. In Python code, the Whisper library was imported. A pre-trained model called 'base' was loaded, which had undergone extensive training on a substantial audio dataset, guaranteeing efficient speech-to-text conversion. Furthermore, the performance of '.en' models is better for 'base.en' models [105]. Therefore, we selected the 'base' model. This model was applied to convert the recorded audio data into text format using the transcribe(audio_file(in .wav format)) method. The performance of speech recognition tools can be evaluated by evaluation metrics such as WER [106], MER [107], WIL [108], and CER [109]. These evaluation metrics help in comparing the original text and Whisper-generated text. Lower values of WER, MER, WIL, and CER correspond to higher accuracy [110].

3.8. Missing Value Imputation

The audio features, evaluation metrics, persons' demographics (age and gender), statements, and distances (1 m, 3 m, and 5 m) were merged in a pandas dataframe. Further, the missing values for data points were searched using the isnull(audio_file(in .wav format)) function. The missing values were filled with the mean value of the column.

3.9. Selection of Best Cluster

The evaluation metrics provide insights into the quality of speech recognition for individual data points. A threshold for Word Error Rate (WER) values was set at less than 0.3 to identify the most suitable data points within a cluster. The cluster with a higher proportion of WER values below 0.3 can be considered optimal.

4. Results

4.1. Feature Generation from Audio Recordings, Pre-Processing and Machine Learning

Extracting various features from audio recordings is necessary for feature generation to better understand and analyze the recorded speech. In addition to audio features such as the MFCC, pitch, spectral centroid, spectral flatness, energy, and ZCR, other features such as the distance from the robot, spoken statements, age, and gender may also be required for a comprehensive analysis. Figure 2 shows an overview of extracted audio features from the signal. The figure includes the top five rows and six columns—MFCC, Pitch, spectral_centoid, spectral_flatness, energy, and zcr. There were eighteen recordings for each person, as each person was placed at three different distances from Pepper, and at each position, they spoke six statements. Furthermore, the vector size of each frame was 419. Hence, there were 7542 records generated for a single person. This study recruited eight participants, resulting in a grand total of 60,336 recordings.

	MFCC	Pitch	spectral_centroid	spectral_flatness	energy	zcr
0	[[-507.6966, -471.8882, -469.5984, -463.8056, ...	[65.35, 209.0854, 51.2889, 219.8571, 86.5871, ...	[[1369.119, 1290.6343, 1338.5117, 1352.5655, 1...	[[0.0006, 0.0004, 0.0004, 0.0004, 0.0004, 0.00...	[[0.0019, 0.0023, 0.0028, 0.0029, 0.003, 0.003...	[[0.0356, 0.0566, 0.0747, 0.0757, 0.0713, 0.06...
1	[[-490.5626, -472.8516, -491.7336, -492.8414, ...	[59.5123, 83.616, 103.6455, 105.4294, 211.135,...	[[1698.1665, 1503.8689, 1131.6629, 1068.3465, ...	[[0.0012, 0.0007, 0.0003, 0.0002, 0.0002, 1e-0...	[[0.0017, 0.0022, 0.0028, 0.003, 0.0034, 0.003...	[[0.0356, 0.0498, 0.0645, 0.0566, 0.0547, 0.05...
2	[[-486.0734, -463.5099, -476.4918, -466.5966, ...	[98.1087, 104.5215, 140.3482, 53.1877, 53.0249...	[[1632.8016, 1519.5047, 1436.3318, 1618.867, 1...	[[0.0006, 0.0004, 0.0004, 0.0005, 0.0006, 0.00...	[[0.0024, 0.0029, 0.0032, 0.0032, 0.0032, 0.00...	[[0.0278, 0.043, 0.0635, 0.0723, 0.082, 0.0898...
3	[[-380.044, -373.7766, -387.212, -395.8484, -4...	[59.0734, 169.3021, 169.026, 168.4382, 167.462...	[[1047.0498, 1149.1074, 1225.1155, 1222.2725, ...	[[1e-04, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0....	[[0.017, 0.0201, 0.0225, 0.021, 0.0201, 0.0196...	[[0.0186, 0.0259, 0.0347, 0.0322, 0.0391, 0.04...
4	[[-488.689, -465.3022, -487.2518, -498.2076, -...	[76.899, 85.8926, 65.6659, 68.9147, 69.7119, 5...	[[1480.2912, 1482.3726, 1173.0456, 1079.8927, ...	[[0.0009, 0.0008, 0.0003, 0.0002, 0.0003, 0.00...	[[0.002, 0.0025, 0.0028, 0.0028, 0.0028, 0.002...	[[0.0312, 0.04, 0.0532, 0.0479, 0.0489, 0.0522...

Figure 2. Screenshot of top five rows (out of 18 records) of extracted audio features of a person.

The features were on different scales, and it is important that they should be on a common scale before applying any machine learning algorithms. StandardScaler was used to obtain the standardized features. Figure 3 contains the top five rows and eighteen columns. The column names from MFCC1 to MFCC13 correspond to thirteen MFCCs. The rest of the columns are for Pitch, spectral_centoid, spectral_flatness, energy, and zcr. It is important to apply machine learning algorithms to determine insights into the audio features.

MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13	Pitch	spectral_c	spectral_fl	energy	zcr
−1.1326	−0.2834	0.4296	−1.1069	0.3728	−0.1615	1.1191	0.3596	0.695	−0.9532	0.8007	0.5663	−1.1091	−0.3263	−0.103	0.6163	−0.2981	−1.2121
−0.3393	−0.2111	0.1175	−0.5437	0.3932	0.0265	0.5071	0.3453	0.6758	−0.8465	0.8839	0.6915	−0.9432	−0.5536	0.0362	0.4737	−0.2778	−0.5979
−0.4794	−0.251	0.1916	−0.4658	0.4606	−0.2276	0.6369	0.5671	1.2307	−0.3623	−0.1466	0.2253	−0.6016	−0.5537	0.0605	0.652	−0.2513	−0.0247
−0.5096	−0.2026	0.2346	−0.6467	0.2386	−0.381	0.428	−0.1082	0.5661	0.055	−0.7775	−0.8249	0.0937	−0.9553	−0.0257	0.3895	−0.2473	−0.3727
−0.5524	−0.2371	0.2088	−0.5555	0.4223	−0.3081	0.4382	−0.408	−0.1006	−0.6036	−0.6912	−0.3572	0.9417	0.3535	−0.066	0.2037	−0.2449	−0.3523

Figure 3. Screenshot of top 5 rows (out of 60,336 records) with standardized 18 audio features.

K-means clustering was applied here to group the data points based on their feature similarities. Three clusters were formed (Figure 4). As per Figure 4, Cluster 1 is the biggest cluster (indicated by blue color), followed by Cluster 0 (indicated by orange color), and Cluster 2 (indicated by green color) has the minimum number of data points.

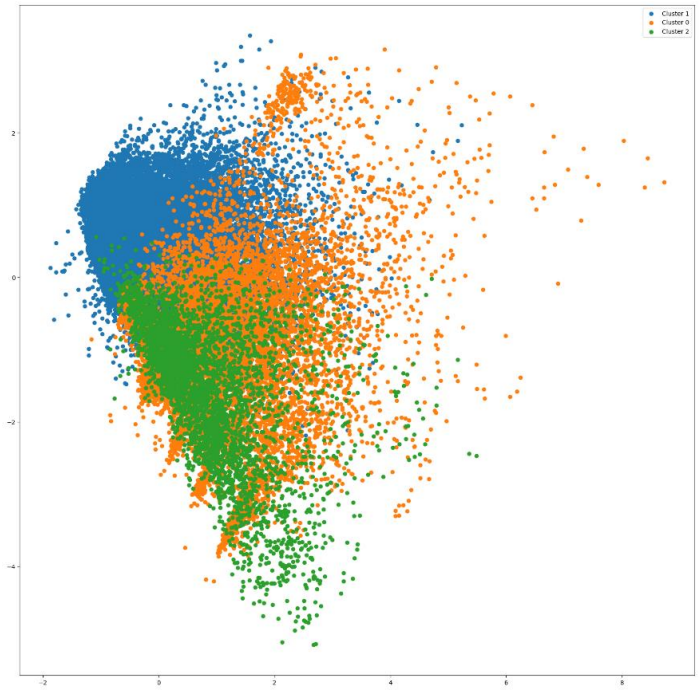


Figure 4. Three clusters generated by K-mean clustering. The color indicators for cluster labels are shown in the box.

4.2. Visualization of Clusters with Features

The distribution of clusters is shown in Figure 5, which indicates that Cluster 1 has the highest number of records. Cluster 0 is the second-largest cluster, but there is a significant difference in the number of elements compared to Cluster 1. Cluster 2 contains the minimum number of records. It is crucial to visualize the individual features present in each cluster.

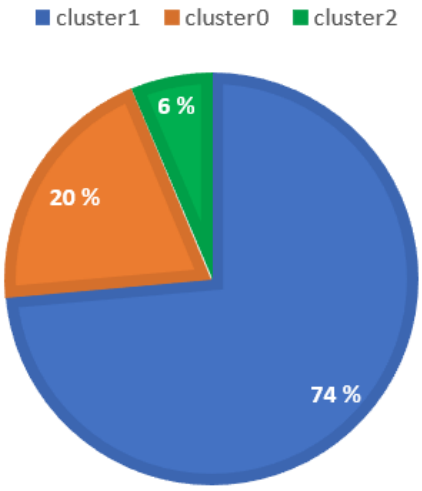


Figure 5. Distribution of data records in each cluster. Cluster 0, Cluster 1, and Cluster 2 are indicated by orange, blue, and green colors, respectively.

This study utilized a scatter matrix to effectively visualize and analyze a range of audio features and their respective clusters. Figure 6 provides a valuable overview of this matrix, which is particularly helpful in identifying correlations and relationships between feature pairs. Notably, the diagonal of the matrix provides insight into the distribution of individual features within the dataset.

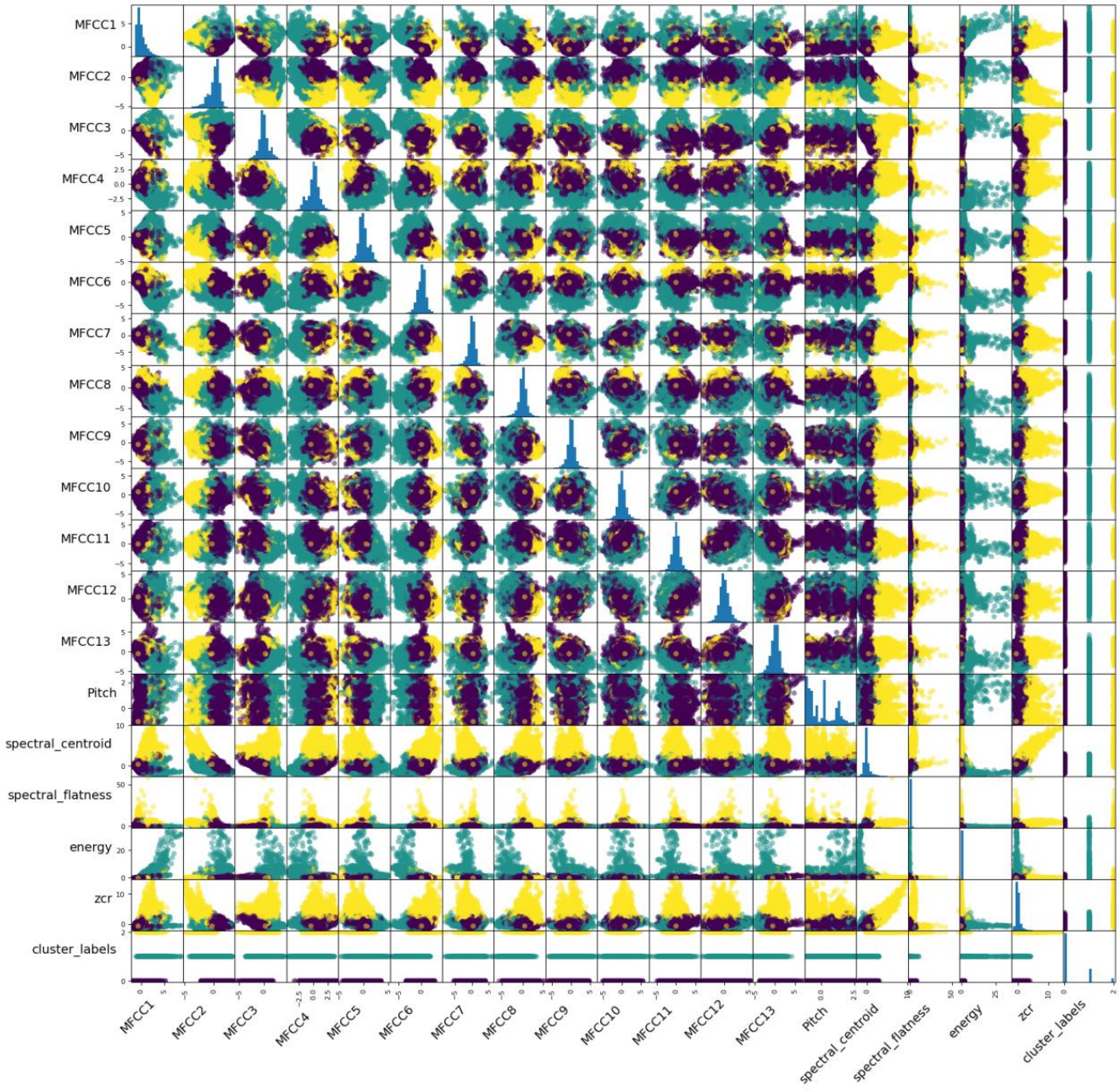


Figure 6. Scatter matrix of all eighteen features in each cluster. Purple color indicates Cluster 0, green indicates Cluster 1, and yellow indicates Cluster 2. Blue color plots in the diagonal represent the data distribution for each feature.

In addition, Figure 6 offers an illuminating breakdown of features across different clusters, with purple denoting Cluster 0, green denoting Cluster 1, and yellow denoting Cluster 2. It basically provides information about the pairwise relationships of any of the two features in all three clusters. To further enhance our understanding of these patterns, we have included three separate figures: Figure 7 for Cluster 0, Figure 8 for Cluster 1, and Figure 9 for Cluster 2.

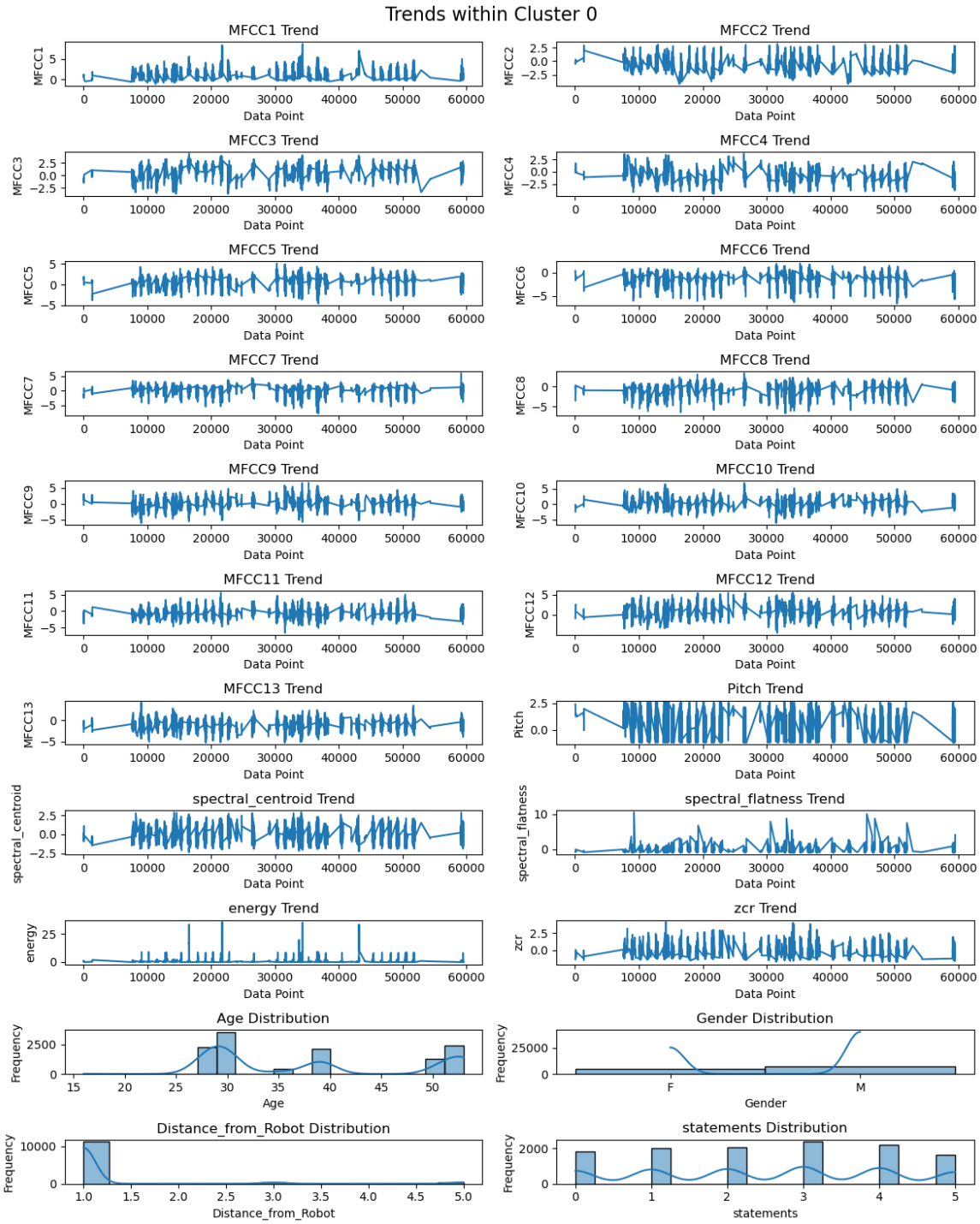


Figure 7. Visualization of trends of twenty-two features in Cluster 0.

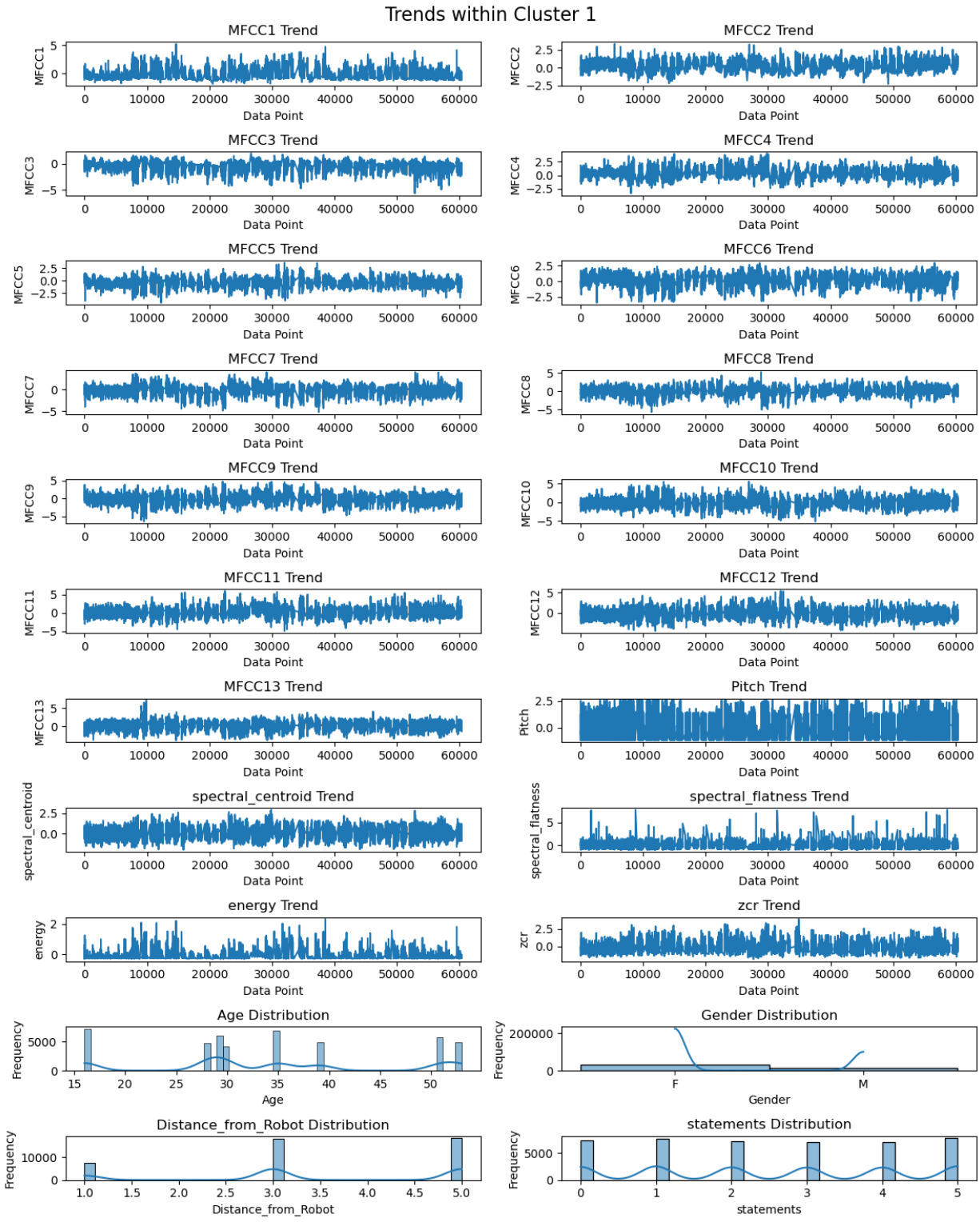


Figure 8. Visualization of trends of twenty-two features in Cluster 1.

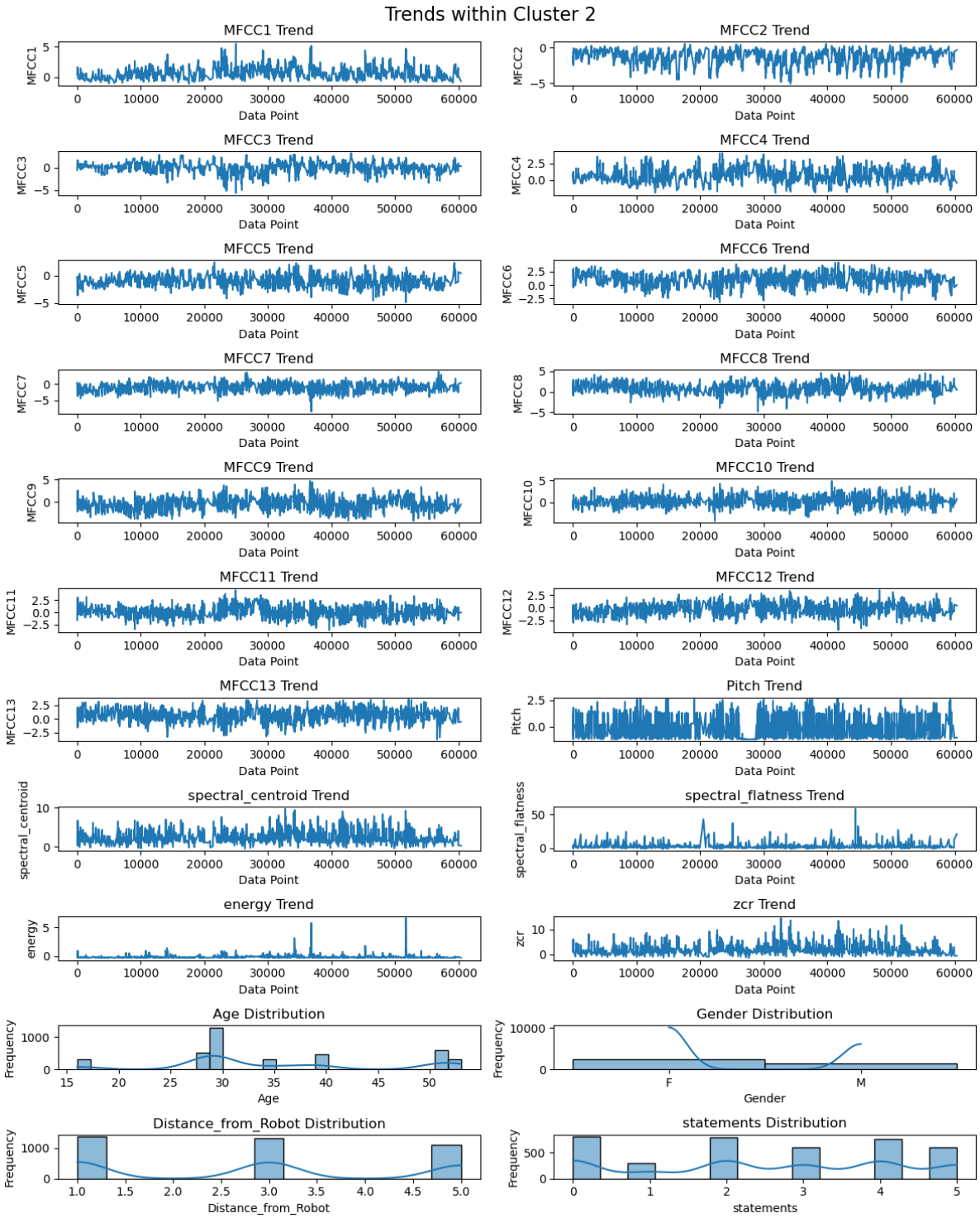


Figure 9. Visualization of trends of twenty-two features in Cluster 2.

The descriptive statistics of each feature of Cluster 0 are illustrated in Table 1. These descriptive statistics provide a complete view of audio features and contain values for the count, mean, standard deviation (std), min, 25%, 50%, 75%, and max. The central tendency and variability of these features can be computed by observing the mean and standard deviation; however, the range of audio features can be found by observing the corresponding min and max values. As per Table 1, the number of records for each feature in Cluster 0 is 44406. The mean values of MFCCs lie between −0.34 and 0.42, whereas the standard deviation (s.d.) varies from 0.5 to 0.9. The mean values of MFCCs close

to 0 indicate that the distribution is balanced. On the other hand, a higher standard deviation denotes the spread of diverse audio features. The pitch has a mean value of −0.06 with a standard deviation of 0.98. The standard deviation confirms the diversity in pitch similar to the standard deviations of the MFC coefficients. The mean values for spectral centroid, spectral flatness, energy, and ZCR are −0.18 with s.d. of 0.5, −0.12 with s.d. of 0.47, −0.12 with s.d. of 0.2, and −0.18 with s.d. of 0.53, respectively. Further, the data presented in Figure 7 outline the observed patterns across twenty-two features within Cluster 0. These twenty-two features contain audio features, including thirteen MFCC features (MFCC1–MFCC13), pitch, spectral centroid, spectral flatness, energy, ZCR, and spectral features such as age, gender, distance, and statements. The chart highlights that MFCC1 is the coefficient that contains mostly positive peaks, while MFCC2 to MFCC13 have both negative and positive peaks. The pitch chart contains both positive and negative peak values. The spectral centroid fluctuates between negative and positive values, while spectral flatness consistently remains positive, with a maximum peak value higher than 10. The energy pattern is predominantly positive, with a maximum peak value exceeding 25, and the Zero-Crossing Rate (ZCR) varies between positive and negative values. Regarding demographics, people aged between 25 and 30 are the highest in number in Cluster 0, and the gender distribution appears to be roughly equal, with slightly more males. The data indicate that most data points in this cluster are associated with a distance of 1 m from the robot. Furthermore, Cluster 0 shows a higher occurrence of the fourth and fifth statements compared to other statements.

Table 1. Descriptive statistics of each feature of Cluster 0.

	MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13	Pitch	Spectral _Centroid	Spectral _Flatness	Energy	zcr
count	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0	44,406.0
mean	−0.34	0.42	−0.32	0.27	−0.24	0.23	−0.03	0.1	−0.0	−0.12	0.2	−0.17	0.22	−0.06	−0.18	−0.12	−0.12	−0.18
std	0.7	0.53	0.66	0.68	0.65	0.67	0.79	0.83	0.9	0.88	0.87	0.87	0.81	0.98	0.5	0.47	0.2	0.53
min	−1.87	−2.28	−5.62	−3.34	−4.41	−3.42	−5.27	−5.68	−6.41	−5.11	−4.91	−4.48	−3.81	−1.21	−1.98	−0.95	−0.32	−1.83
25%	−0.83	0.08	−0.61	−0.16	−0.64	−0.15	−0.42	−0.33	−0.49	−0.61	−0.32	−0.7	−0.23	−0.89	−0.45	−0.45	−0.23	−0.45
50%	−0.57	0.45	−0.24	0.22	−0.26	0.29	0.04	0.13	0.02	−0.16	0.15	−0.2	0.28	−0.37	−0.25	−0.13	−0.2	−0.23
75%	−0.07	0.77	0.11	0.62	0.17	0.68	0.46	0.58	0.51	0.32	0.65	0.32	0.74	0.59	0.01	0.14	−0.08	0.04
max	5.23	3.34	1.96	4.06	3.60	2.96	4.00	5.26	4.77	5.52	6.02	5.40	7.07	2.57	2.95	7.76	2.32	3.87

The descriptive statistics of each feature of Cluster 1 are illustrated in Table 2. As per Table 2, the number of records for each feature in Cluster 1 is 12122. The mean values of MFCCs lie between −1.25 and 1.23, whereas the standard deviation (s.d.) varies from 0.96 to 1.27. The pitch has a mean value of 0.24 with a standard deviation of 1.05. The mean values for spectral centroid, spectral flatness, energy, and ZCR are −0.24 with s.d. of 0.73, −0.28 with s.d. of 0.60, 0.46 with s.d. of 2.13, and −0.17 with s.d. of 0.64, respectively. Further, Figure 8 in Cluster 1 showcases the distribution of twenty-two distinct features. These features contain audio features, including thirteen MFCC features (MFCC1–MFCC13), pitch, spectral centroid, spectral flatness, energy, and ZCR, as well as spectral features such as age, gender, distance, and statements. The graph depicts that MFCC1 has predominantly positive peaks, with the highest peak reaching greater than +5. On the other hand, other coefficients, such as MFCC2 to MFCC13, exhibit a mix of positive and negative peaks. Pitch's highest peak value is more than 2.5, though it dips into negative values for some data records. The trend of the spectral centroid and spectral flatness varies between positive and negative peak values. Energy levels also remain mostly positive, with the highest peak exceeding 2. ZCR fluctuates between positive and negative values. Within Cluster 1, the most frequent age is 16 years, followed by individuals aged 35. Females are in a higher proportion of the cluster than males. Upon analyzing the distance from the robot, it appears that there are significantly more data records at distances of 3 and 5 m within Cluster 1.

Table 2. Descriptive statistics of each feature of Cluster 1.

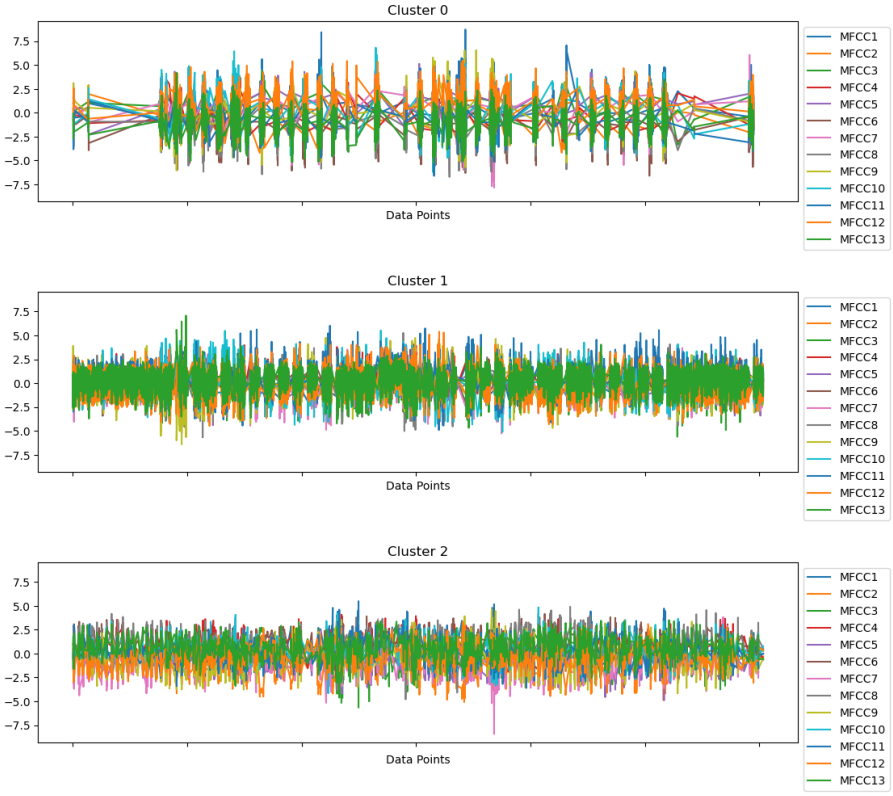
	MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13	Pitch	Spectral _Centroid	Spectral _Flatness	Energy	zcr
count	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00	12,122.00
mean	0.98	−1.06	1.17	−1.25	1.23	−1.19	0.46	−0.68	0.15	0.34	−0.74	0.75	−1.02	0.24	−0.24	−0.28	0.46	−0.17
std	1.12	1.06	1.05	0.96	1.01	0.97	1.27	1.08	1.18	1.24	1.08	1.08	0.94	1.05	0.73	0.60	2.13	0.64
min	−1.16	−4.21	−3.70	−4.45	−4.62	−6.60	−7.85	−6.72	−6.05	−6.21	−6.61	−4.50	−5.23	−1.21	−2.38	−0.96	−0.30	−1.76
25%	0.07	−1.50	0.66	−1.80	0.96	−1.44	0.12	−1.11	−0.42	−0.29	−1.39	0.15	−1.53	−0.73	−0.64	−0.76	−0.21	−0.54
50%	0.69	−1.17	1.43	−1.45	1.40	−1.03	0.81	−0.48	0.14	0.38	−0.83	0.73	−0.94	0.17	−0.30	−0.35	−0.09	−0.23
75%	1.61	−0.59	1.79	−0.84	1.75	−0.71	1.19	−0.03	0.74	0.98	−0.21	1.33	−0.45	1.21	0.05	0.05	0.33	0.08
max	8.72	3.15	4.38	3.65	5.13	1.95	6.05	3.36	6.56	6.82	5.56	5.46	4.20	2.57	2.91	10.58	35.54	4.11

The descriptive statistics of each feature of Cluster 2 are illustrated in Table 3. As per Table 3, the number of records for each feature in Cluster 2 is 3808. The mean values of MFCCs lie between −1.55 and 1.10, whereas the standard deviation (s.d.) varies from 0.93 to 1.30. The pitch has a mean value of −0.03 with a standard deviation of 0.98. The mean values for spectral centroid, spectral flatness, energy, and ZCR are 2.88 with s.d. of 1.51, 2.24 with s.d. of 2.59, −0.10 with s.d. of 0.30, and 2.60 with s.d. of 2.02, respectively. Moreover, Figure 9 depicts the patterns of twenty-two features within Cluster 2. The features are related to both audio features, including thirteen MFCC features (MFCC1–MFCC13), pitch, spectral centroid, spectral flatness, energy, and ZCR, as well as spectral features such as age, gender, distance, and statements. It shows that MFCC1 has mostly positive peaks, whereas other MFCCs (MFCC2 to MFCC13) display both positive and negative peaks. The trend of the spectral centroid varies, while spectral flatness remains consistently positive, with a maximum peak value of more than 50. The energy pattern is also consistently positive, with a maximum peak value exceeding 5. ZCR peaks fall between positive and negative values. Regarding the age distribution in Cluster 0, the majority falls within the 25–30 age group. In Cluster 2, there are more females than males. Analysis of the distance from the robot reveals that there are more occurrences of distances around 1 m, followed by distances of approximately 3 m. Furthermore, this cluster exhibits a higher frequency of the first and third statements.

Table 3. Descriptive statistics of each feature of Cluster 2.

	MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13	Pitch	Spectral _Centroid	Spectral _Flatness	Energy	zcr
count	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00	3808.00
mean	0.89	−1.55	−0.00	0.84	−1.10	1.10	−1.14	0.94	−0.48	0.37	0.05	−0.37	0.64	−0.03	2.88	2.24	−0.10	2.60
std	0.93	1.00	1.30	1.05	0.97	1.14	1.25	1.27	1.30	1.07	1.04	1.02	1.04	0.98	1.51	2.59	0.30	2.02
min	−1.08	−5.09	−5.67	−2.00	−4.90	−3.19	−8.44	−4.79	−4.18	−4.17	−3.60	−4.44	−3.77	−1.21	−0.45	−0.94	−0.31	−1.54
25%	0.20	−2.17	−0.70	0.11	−1.75	0.42	−2.04	0.09	−1.40	−0.35	−0.64	−1.03	−0.01	−0.89	1.81	1.11	−0.21	1.24
50%	0.74	−1.34	0.17	0.77	−1.20	1.21	−1.18	0.92	−0.48	0.37	0.02	−0.40	0.71	−0.36	2.50	1.83	−0.18	2.12
75%	1.40	−0.82	0.82	1.56	−0.47	1.93	−0.26	1.81	0.39	1.08	0.70	0.27	1.34	0.82	3.62	2.83	−0.09	3.44
max	5.49	0.55	3.20	4.07	2.48	4.21	3.82	4.90	4.68	4.84	4.48	3.44	3.50	2.57	9.77	58.17	6.64	14.53

Furthermore, a comprehensive examination of the MFCC features is performed, and the outcomes are illustrated in Figure 10, where MFCC features are denoted in different colors. The findings indicate that within Cluster 0, the most prominent coefficient is MFCC10, while the least significant is MFCC7. For Cluster 1, the highest coefficient is MFCC13 and the smallest is MFCC9. In contrast, within Cluster 2, MFCC8 has the highest coefficient and MFCC7 has the lowest.




```
Cluster 0: Highest Value MFCCs in Descending Order: ['MFCC10', 'MFCC9', 'MFCC7', 'MFCC11', 'MFCC12', 'MFCC5', 'MFCC3', 'MFCC
13', 'MFCC4', 'MFCC8', 'MFCC2', 'MFCC6']
Cluster 0: Lowest Value MFCCs in Ascending Order: ['MFCC7', 'MFCC8', 'MFCC11', 'MFCC6', 'MFCC10', 'MFCC9', 'MFCC13', 'MFCC
5', 'MFCC12', 'MFCC4', 'MFCC2', 'MFCC3']
Cluster 1: Highest Value MFCCs in Descending Order: ['MFCC13', 'MFCC11', 'MFCC10', 'MFCC12', 'MFCC8', 'MFCC9', 'MFCC4', 'MFC
C7', 'MFCC5', 'MFCC2', 'MFCC6', 'MFCC3']
Cluster 1: Lowest Value MFCCs in Ascending Order: ['MFCC9', 'MFCC8', 'MFCC3', 'MFCC7', 'MFCC10', 'MFCC11', 'MFCC12', 'MFCC
5', 'MFCC13', 'MFCC6', 'MFCC4', 'MFCC2']
Cluster 2: Highest Value MFCCs in Descending Order: ['MFCC8', 'MFCC10', 'MFCC9', 'MFCC11', 'MFCC6', 'MFCC4', 'MFCC7', 'MFCC1
3', 'MFCC12', 'MFCC3', 'MFCC5', 'cluster_labels']
Cluster 2: Lowest Value MFCCs in Ascending Order: ['MFCC7', 'MFCC3', 'MFCC2', 'MFCC5', 'MFCC8', 'MFCC12', 'MFCC9', 'MFCC10',
'MFCC13', 'MFCC11', 'MFCC6', 'MFCC4']
```

Figure 10. Visualization of MFCCs in each of the clusters.

Furthermore, the 10 most important features within each cluster were selected using the KMeansInterp function, as suggested by Yousef et al. [111]. The output obtained for the three clusters is shown in Figure 11.

```
kms.feature_importances_[0][:10]
```

```
[('spectral_centroid', 2.8846697934023986),
 ('zcr', 2.598194546096255),
 ('spectral_flatness', 2.236611201798363),
 ('MFCC2', 1.5508817303249367),
 ('MFCC7', 1.1362999323988712),
 ('MFCC5', 1.098433010767682),
 ('MFCC6', 1.0969346586455049),
 ('MFCC8', 0.9398093397662787),
 ('MFCC1', 0.8858775409806607),
 ('MFCC4', 0.8396091417530148)]
```

```
kms.feature_importances_[1][:10]
```

```
[('MFCC2', 0.42253792057540107),
 ('MFCC1', 0.344027897555201),
 ('MFCC3', 0.3183882384150074),
 ('MFCC4', 0.2681007255009142),
 ('MFCC5', 0.24301260696088056),
 ('MFCC6', 0.23027944052940505),
 ('MFCC13', 0.223856418795945),
 ('MFCC11', 0.19841275970438094),
 ('spectral_centroid', 0.18064760172758818),
 ('zcr', 0.17564881486171846)]
```

```
kms.feature_importances_[2][:10]
```

```
[('MFCC4', 1.2448306771779756),
 ('MFCC5', 1.234233405478045),
 ('MFCC6', 1.1871563994664858),
 ('MFCC3', 1.1659055175466617),
 ('MFCC2', 1.0598560913668689),
 ('MFCC13', 1.0206231092507496),
 ('MFCC1', 0.9812002328969539),
 ('MFCC12', 0.745391097178917),
 ('MFCC11', 0.7406951733968848),
 ('MFCC8', 0.6799917685714225)]
```

Figure 11. Ten most important features in three clusters.

4.3. Integration of Whisper to Evaluate Pepper's Speech Recognition System

The recorded audio can be converted into text with the help of the Whisper speech-to-text recognition tool. The evaluation metrics suggest the extent of similarity between the Whisper-generated text and the original text. The measures WER, MER, WIL, and CER were calculated for all recordings. Then, we combined these values with the above-mentioned audio features to evaluate which cluster has the best values.

However, before the best cluster selection, imputing the missing values corresponding to WER, MER, WIL, and CER was necessary. These values were filled by taking the mean of the columns where these values were present. Figure 12 shows the 28 combined features, which include the person's demographics, statements, position, audio features, and evaluation measures. This figure provides a glimpse of features included for analysis purposes and not the complete details of all the participants.

	user	Age	Gender	statement	Distance_	MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	...	Pitch	spectral_c	spectral_f	energy	zcr	cluster_la	WER	MER	WIL	CER
0	user1	16	F	2	1	-1.1326	-0.2834	0.4296	-1.1069	0.3728	...	-0.3263	-0.103	0.6163	-0.2981	-1.2121	0	0.07	0.07	0.14	0.08
1	user1	16	F	2	1	-0.3393	-0.2111	0.1175	-0.5437	0.3932	...	-0.5536	0.0362	0.4737	-0.2778	-0.5979	0	0.07	0.07	0.14	0.08
2	user1	16	F	2	1	-0.4794	-0.251	0.1916	-0.4658	0.4606	...	-0.5537	0.0605	0.652	-0.2513	-0.0247	0	0.07	0.07	0.14	0.08
3	user1	16	F	2	1	-0.5096	-0.2026	0.2346	-0.6467	0.2386	...	-0.9553	-0.0257	0.3895	-0.2473	-0.3727	0	0.07	0.07	0.14	0.08
4	user1	16	F	2	1	-0.5524	-0.2371	0.2088	-0.5555	0.4223	...	0.3535	-0.066	0.2037	-0.2449	-0.3523	0	0.07	0.07	0.14	0.08

Figure 12. Screenshot of top 5 rows (out of 60,336 records) and 28 features, including person's demographics, position, audio features, and evaluation metrics.

4.4. Selection of Best Cluster Depending on WER Threshold

The lower value of WER indicates a higher accuracy. The best cluster should contain the maximum number of WER values rather than a larger number of data points. A threshold of less than or equal to 0.3 was implemented for WER and all records were filtered. The quality of the cluster was checked, as shown in Figure 13, in which a higher value indicates the best-quality cluster. This suggests that Cluster 1 is the best cluster as it contains records with higher accuracies in speech recognition.

```
q1=len(filtered_WER_cluster_0)/len(cluster_0)
q2=len(filtered_WER_cluster_1)/len(cluster_1)
q3=len(filtered_WER_cluster_2)/len(cluster_2)
```

```
print(q1)
print(q2)
print(q3)
```

```
0.27750754402558214
0.5131166474179178
0.3019957983193277
```

Figure 13. Screenshot of code snippet to select the best cluster.

4.5. Visualization of Best Records in Cluster 1

It is important to analyze the pattern of data records that have good accuracies in Cluster 1. For this purpose, different pie charts (Figure 14) have been created. Figure 14a shows the distribution of gender in filtered Cluster 1 and suggests that there is no great effect of male and female voices as they are almost equal in number. Figure 14b illustrates that speech recognition was best when the distance between the person and the robot is lowest, i.e., 1 m. At 3 m and 5 m, there is not much difference. As per Figure 14c, the most accurately recognized statement was statement3 followed by statement1. It should be noted that the statement0 was not present in this chart, which shows that that statement was not recognized. Figure 14d shows that speech recognition was best for those who are nearly 30 years of age.

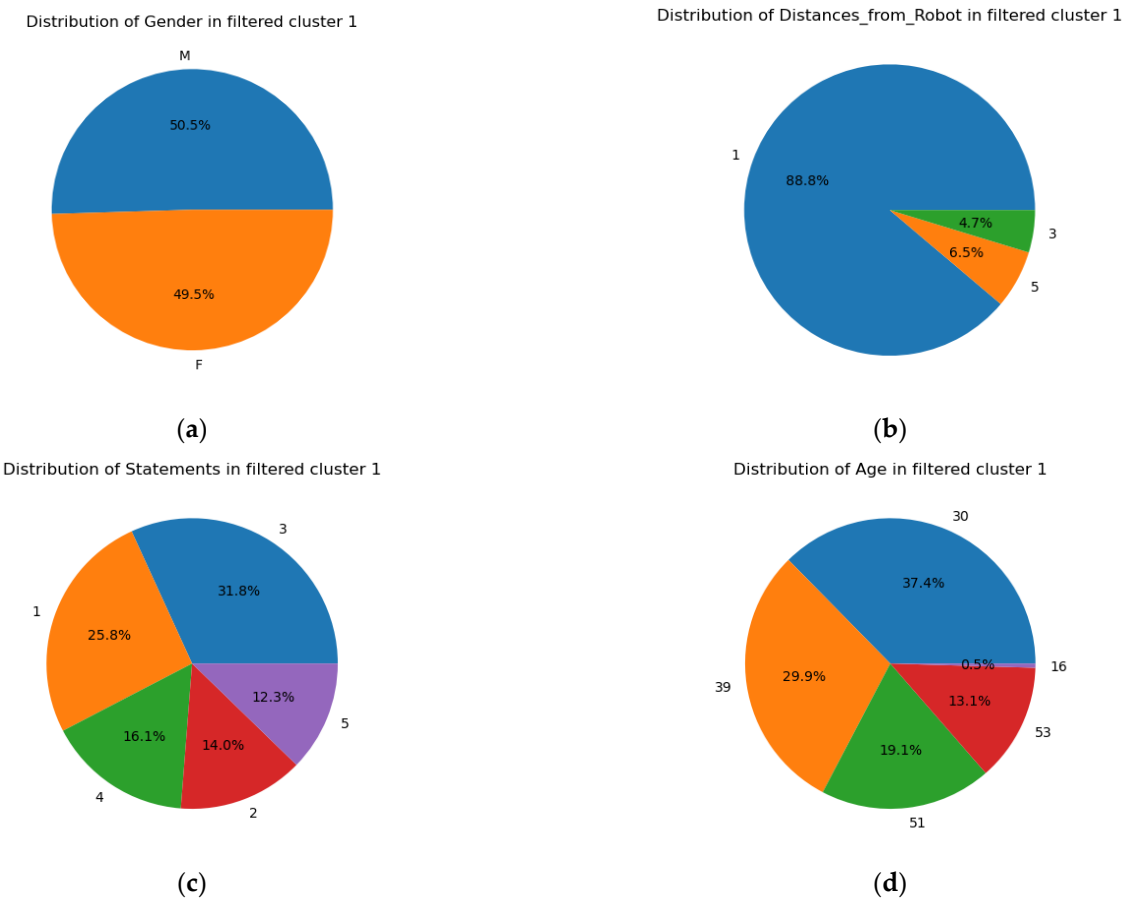


Figure 14. Analysis of the distribution of (**a**) gender, (**b**) distances from the robot, (**c**) statements spoken, and (**d**) age for best records in Cluster 1.

5. Discussions

The present study employs an amalgamation of features—MFCCs, pitch, the spectral centroid, spectral flatness, energy, and the ZCR—to advance the understanding of speech recordings. These speech recordings were captured using Pepper's speech recognition system. The objective is to understand the impact of spatial features such as age, distance from Pepper, statements, and gender on Pepper's speech recognition system. To achieve this, the recordings were transferred to a local system for further analysis, where the unsupervised machine learning K-means technique was applied to audio features to group them based on their properties. Further, Whisper was applied for speech-to-text conversion and evaluation metrics were used to calculate the accuracy of captured audio. The cluster that contains higher accuracies was selected as the best. In the current study, more than one audio feature is selected. Previous studies [16–18,112] also demonstrated that the combination of features can provide comprehensive insights into speech data. For instance, Kerkeni et al. [17] proposed that combining two features can enhance the performance of speech emotion recognition (SER) performance. Similarly, Dash et al. [16] employed gradient-boosting machines to accomplish consistent performance across different datasets via the fusion of spectral, cepstral, and periodicity features. Furthermore, Jiang et al. [112] utilized a combination of MFCCs, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), and high-level acoustic features from speech recognition networks to extract relevant speech-related information. Moreover, Wang et al. [18] incorporated MFCC and spectrogram features for speech emotion recognition, achieving 29.4% and 23.8% higher classification accuracy than MFCC and spectrogram features, respectively. It has been found that incorporating features such as MFCCs, the ZCR, and LPC into the speech recognition system resulted in consistent performance, even with larger numbers of

participants [113]. In past studies, unsupervised learning was also applied for different purposes related to speech recognition, such as the reduction in feature space dimension [79], mood detection [81], emotion classification [82], and the classification of sound signals [83]. Similarly, in the present research, K-mean clustering is used to better understand the impact of spatial features on the speech recognition system of the Pepper robot.

In this study, relevant features were selected based on the existing literature. MFCCs, the spectral centroid, spectral flatness, and the ZCR are generally employed in speech processing to extract relevant information. Pitch and RMS are essential features for analyzing speech recordings. Davis et al. [114] suggested MFCCs as an approach to demonstrating speech's spectral features, laying the basis for applying MFCCs as the fundamental feature set for speech analysis. Lehner et al. [115] achieved significant recognition results with a minimal and optimized feature set mostly consisting of selected MFCCs. Reports indicate that spectral centroid features effectively capture more information from subband spectral distributions, especially in speech recognition systems [116,117]. Further, spectral centroid features can also help in speech emotion detection [63,118]. Qadri et al. [119] differentiated gender-specific speech using spectral features such as the spectral centroid and spectral flatness. Another study [120] showed that spectral flatness denotes the noise-like behavior of a waveform where the high peaks indicate lower spectral flatness. The ZCR can be helpful in speech processing, including capturing information about loud and short sounds [121]. Chauhan et al. [122] proposed that integrating prosodic information with spectral features can be beneficial for advancement in speech recognition systems. Furthermore, they also suggested that pitch and RMS features are less affected by channel and noise.

There have been few studies focusing on the analysis of audio features obtained from social robots. Bird et al. [123] mentioned that Pepper and Nao robots are suitable for general speech recognition; however, they do not have the advanced capability of classifying speakers. Shen et al. [124] proposed a framework for personality trait identification through HRI using various features, including vocal features such as MFCCs, pitch, and energy. In another study [125], audio data were collected during human–robot interaction using Pepper's built-in sensor and lapel microphone; further, MFCCs and pitch features were utilized to analyze the data. In the past, studies were conducted on the analysis of audio captured by robots other than social robots. Telembici et al. [126] extracted MFCCs to analyze the accuracy of audio captured by the service robot TIAGo. Wu et al. [127] employed a surveillance robot to identify abnormal audio sounds using MFCC feature extraction. Drawing inspiration from the above-mentioned studies, we adopted the Pepper robot for the present research.

The performance of speech recognition systems can be tested using evaluation metrics. Pande et al.'s [110] findings suggest that Whisper is the best speech recognition tool among other speech recognition tools. Furthermore, they tested the accuracy based on the evaluation measures WER, MER, WIL, and CER and found that these measures were low in the case of Whisper, which is an indicator of good performance. The continuous monitoring of these measures in the many real-world applications of Pepper can be helpful in understanding how the speech recognition system performs. This, in turn, can be advantageous for user interaction, satisfaction, and reliability. Similar to our approach in the present study, Abdelhadi et al. [128], in their research, also employed Whisper to convert audio recordings captured by the Pepper robot.

The impact of gendered voices on speech recognition systems has been extensively explored in the field of ASR. Research studies highlight the importance of understanding and addressing gender-related differences in speech characteristics to enhance the accuracy of speech recognition technologies. These differences include variations in vocal tract length [129] and pitch [130]. Garnerin et al. [131] discovered significant differences in the performance of ASR systems, measured by the Word Error Rate (WER), between males and females. Adda-Decker et al. [132] found that females tended to achieve better speech recognition results on average for both English and French speech than males. However, a study conducted by Tatman et al. [133] suggested the opposite, finding the suitability of

the male voice for speech recognition. Doddington et al. [134] discussed that both male and female voices exhibit distinct physical attributes such as pitch and vocal tract length. Furthermore, their findings suggested a minor performance difference between genders, with a slight tendency for males to perform better. The present study also demonstrated that the difference in the voice ratio between male and female participants is very small, with slightly a higher value for male voices.

The speaker's distance from the system can affect the quality and accuracy of speech recognition. Studies by [135,136] have shown that as the speaker moves further away from the microphone, the accuracy of speech recognition decreases. Similarly, the present study shows that in best-selected Cluster 1, a shorter distance between the speaker and the microphone results in more accurate speech recognition. The existing literature also suggested an impact on vocal intensity with increasing distance [137,138]. Meyer et al. [54] examined how environmental noise and varying distances affect word recognition. Attawibulkul et al. [53] proposed that environmental noise presents a significant challenge for speech recognition in robots.

Previous research studies have demonstrated that speech recognition accuracy could be influenced by factors such as statement complexity and age. Studies [135,136] demonstrated that the minimum distance from speech recognition devices provides better and more accurate results. Our previous research [10] also indicated that the accuracy of speech recognition systems is affected by the complexity of statements, whereby simple statements made at 1 m from the Pepper robot were the most optimal. Similarly, we found in the present study that the best cluster mostly contains 1 m distance data (88.8%), suggesting speakers should maintain the minimum distance from the robot. Further, participants from different age groups were included in the experiment, and it was found that in the best-performing cluster, the majority of the participants were from the 25–30 age group. However, this could be due to the fact that our sample is skewed towards participants of that age group. The speech characteristics in older people, such as variations in the speech rate, pauses, and reduced articulation, have been identified as factors contributing to increased WERs in ASR systems [139]. Furthermore, children may also struggle with pronunciation when articulating a word. Li et al. [140] illustrated that it is comparatively more difficult for ASR to perceive children's voices than adults. There are studies that showed problems in perceiving spoken words by social robots, including Nao and Pepper, for both children [46,48] and elders [49,50]. Educational qualification might not directly influence the speech recognition system, but it can impact various factors such as pronunciation, speaking style, and rich vocabulary, which in turn can influence the performance of the speech recognition system.

6. Conclusions, Implications, Limitations, and Future Work

The present study employs a humanoid robot, Pepper, to record spoken statements to evaluate its speech recognition system. Different audio features containing accurate information about the speech recordings were selected. The capability of a speech recognition system with the application of K-mean clustering on audio features can provide an effectual assessment. With reference to RQ1, the integration of the speech-to-text converter Whisper is a potential method to assess the precision of recordings performed by Pepper's speech recognition system. With reference to RQ2, the cluster analysis of audio features suggests that the best accuracy of the speech recognition system can be found at a 1 m distance from the Pepper robot in comparison to other distances. On the other hand, age and gender do not seem to have much influence on the performance of speech recognition systems. Within a 1 m distance, the speech recognition system of Pepper performs the best to perceive the audio. Thus, it is recommended that the Pepper robot should be at a maximum distance of one meter from the speaker in diverse settings where the display of subtitles is needed along with speech.

The insights gained from the present study have a significant impact on different domains of the real world. One such domain is hospitality settings, where Pepper can be

utilized as a receptionist with the facility to provide information and handle customers' queries. Another area is healthcare, which, according to the need, Pepper, at the closest distance, can provide companionship, advice, and entertainment to different age groups, including elders and children. The proposed setup can also be advantageous for educational and meeting purposes, including online and physical classroom teaching and meetings. Furthermore, this setup could also be applied to crowded places to comprehend individuals' speech in a better way. It not only increases the understanding of spoken words but also reduces a person's cognitive burden. However, there are challenges in deploying Pepper in crowded places due to the presence of noise. Background noise makes it difficult for Pepper's speech recognition system to perceive its environment efficiently. Moreover, language diversity can also pose challenges for Pepper since to support multilingual environments, switching between languages is needed, which can further impact the accuracy of the speech recognition system. Additionally, the varying accents of people and the limited vocabulary of Pepper can also create another layer of complexity for the proper comprehension of the spoken statements.

The implications of the present study have potential possibilities for advancing human–robot interactions in various contexts. Speech can play a role in the identification of a person's emotional state, including stress and loneliness. The robot can be further programmed to provide appropriate responses after precise capturing of spoken words. The aim of these responses is to help individuals to deal with problems. The responses could be related to effective dialogue establishment. In healthcare scenarios, robots can be programmed to provide remedies and suggestions, including games and exercises, which will help to engage individuals. Meanwhile, in educational settings, the robot can be prepared to help students by providing solutions to their queries. This kind of interaction can increase their engagement and learning experience. A user-friendly robotic system can be designed in which users will have the option to select their choice for engagement; for instance, upon the identification of stress in a person's voice by a robot, the person will have many options such as music, games, and exercises to select from the robot's interface. Similarly, upon detection of not understanding the lecture, there will be options for students to listen to the recordings of the lecture or select another tutorial of their choice.

We experienced some difficulties with Pepper's speech recognition system. One of the biggest problems that Pepper has is the lack of speech-to-text transcription. Therefore, the audio files need to be transferred into the local system using the Paramiko library to convert speech into text, which requires another program to run in addition to the program that is compiled for speech recording. For speech recording, the programmer had to run the same program 18 times for each participant (covering six statements at three different distances). There were 8 participants, and therefore this required running the same program 144 times, demonstrating the workload on the person collecting the data. Sometimes, during the experiment, Pepper's in-built capabilities led it to respond to certain recognized keywords in the participant's speech. In that case, a recording was repeated, which increased the experiment duration. Other challenges of the speech recognition system of Pepper include difficulty in understanding accents, pronunciation, and a restricted vocabulary. Background noise also poses a problem for Pepper to recognize speech accurately. The result of this study is based on a limited sample size. Further, the set of participants selected was random; however, it was skewed towards a certain group. Therefore, the findings should be further strengthened by including a bigger sample size in the future. The present study serves as a prototype for conducting the experiment in real settings. The diversity in sample collection can be maintained by incorporating participants from different countries and different age groups. It is expected that it will help in collecting diverse speech samples. The present study does not focus on emotion detection through speeches. In the future, we will also attempt to collect the emotions of participants, which are expressed through speeches, by using certain stimuli such as watching video clips or playing games.

This study is a crucial initial step in comprehending the system's limitations in a controlled laboratory environment before implementing it in a real setting. Our findings

indicate that the distance between the robot and the speaker has an impact on Pepper's speech recognition capabilities. In the present study, we did not incorporate noise reduction techniques, which might be helpful to increase the performance of speech recognition. It should also be noted that microphones were not used for the study, and the use of microphones can further enhance the accuracy of capturing speech, even from larger distances. In future studies, we will also investigate whether the placement of the microphone with the speaker influences Pepper's speech recognition. We plan to improve human–robot interaction and engagement by displaying the robot's responses on its screen alongside the text.

Author Contributions: Conceptualization, A.P. and D.M.; methodology, A.P. and D.M.; software, A.P.; validation, A.P.; formal analysis, A.P.; investigation, D.M.; resources, A.P. and D.M.; data curation, A.P.; writing—original draft preparation, A.P. and D.M.; writing—review and editing, A.P. and D.M.; visualization, A.P.; supervision, D.M.; project administration, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Norwegian Agency for Shared Services in education and Research (SIKT).

Data Availability Statement: Data will be provided upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abdollahi, H.; Mahoor, M.H.; Zandie, R.; Siewierski, J.; Qualls, S.H. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2020–2032. [CrossRef]
2. Cabibihan, J.-J.; Javed, H.; Ang, M.; Aljunied, S.M. Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *Int. J. Soc. Robot.* **2013**, *5*, 593–618. [CrossRef]
3. Donnermann, M.; Schaper, P.; Lugrin, B. Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. *Front. Robot. AI* **2022**, *9*, 831633. [CrossRef]
4. Lanzilotti, R.; Piccinno, A.; Rossano, V.; Roselli, T. Social Robot to teach coding in primary school. In Proceedings of the 2021 International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 12–15 July 2021; pp. 102–104.
5. Nakanishi, J.; Kuramoto, I.; Baba, J.; Ogawa, K.; Yoshikawa, Y.; Ishiguro, H. Continuous hospitality with social robots at a hotel. *SN Appl. Sci.* **2020**, *2*, 452. [CrossRef]
6. Youssef, K.; Said, S.; Beyrouthy, T.; Alkork, S. A social robot with conversational capabilities for visitor reception: Design and framework. In Proceedings of the 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), Paris, France, 8–10 December 2021; pp. 1–4.
7. Mishra, D.; Romero, G.A.; Pande, A.; Nachenahalli Bhuthegowda, B.; Chaskopoulos, D.; Shrestha, B. An Exploration of the Pepper Robot's Capabilities: Unveiling Its Potential. *Appl. Sci.* **2023**, *14*, 110. [CrossRef]
8. Ghiță, A.; Gavril, A.F.; Nan, M.; Hoteit, B.; Awada, I.A.; Sorici, A.; Mocanu, I.G.; Florea, A.M. The AMIRO Social Robotics Framework: Deployment and Evaluation on the Pepper Robot. *Sensors* **2020**, *20*, 7271. [CrossRef]
9. Pandey, A.K.; Gelin, R.; Robot, A. Pepper: The first machine of its kind. *IEEE Robot. Autom. Mag.* **2018**, *25*, 40–48. [CrossRef]
10. Pande, A.; Mishra, D. The Synergy between a Humanoid Robot and Whisper: Bridging a Gap in Education. *Electronics* **2023**, *12*, 3995. [CrossRef]
11. Ganesh, B.B.; Damodar, B.; Dharmesh, R.; Karthik, K.T.; Vasudevan, S.K. An innovative hearing-impaired assistant with sound-localisation and speech-to-text application. *Int. J. Med. Eng. Inform.* **2022**, *14*, 63–73. [CrossRef]
12. Matre, M.E.; Cameron, D.L. A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education. *Disabil. Rehabil. Assist. Technol.* **2022**, *19*, 1103–1116. [CrossRef]
13. Athikkal, S.; Jenq, J. Voice Chatbot for Hospitality. *arXiv* **2022**, arXiv:2208.10926.
14. Goss, F.R.; Blackley, S.V.; Ortega, C.A.; Kowalski, L.T.; Landman, A.B.; Lin, C.-T.; Meteer, M.; Bakes, S.; Gradwohl, S.C.; Bates, D.W. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *Int. J. Med. Inform.* **2019**, *130*, 103938. [CrossRef]
15. Debnath, S.; Roy, P.; Namasudra, S.; Crespo, R.G. Audio-Visual Automatic Speech Recognition Towards Education for Disabilities. *J. Autism Dev. Disord.* **2023**, *53*, 3581–3594. [CrossRef]
16. Dash, T.K.; Chakraborty, C.; Mahapatra, S.; Panda, G. Gradient boosting machine and efficient combination of features for speech-based detection of COVID-19. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5364–5371. [CrossRef]
17. Kerkeni, L.; Serrestou, Y.; Raoof, K.; Mbarki, M.; Mahjoub, M.A.; Cleder, C. Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Commun.* **2019**, *114*, 22–35. [CrossRef]

18. Wang, Z.; Liu, G.; Song, H. Speech emotion recognition method based on multiple kernel learning feature fusion. *Comput. Eng.* **2019**, *45*, 248–254.
19. Gupta, M.; Chandra, S. Speech emotion recognition using MFCC and wide residual network. In Proceedings of the 2021 Thirteenth International Conference on Contemporary Computing, Noida, India, 5–7 August 2021; pp. 320–327.
20. Ittichaichareon, C.; Suksri, S.; Yingthawornsuk, T. Speech recognition using MFCC. In Proceedings of the International Conference on Computer Graphics, Simulation and Modeling, Pattaya, Thailand, 28–29 July 2012.
21. Ganchev, T.; Fakotakis, N.; Kokkinakis, G. Comparative evaluation of various MFCC implementations on the speaker verification task. In Proceedings of the SPECOM, Patras, Greece, 17–19 October 2005; pp. 191–194.
22. Zhen, B.; Wu, X.; Liu, Z.; Chi, H. On the Importance of Components of the MFCC in Speech and Speaker Recognition. In Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, 16–20 October 2000.
23. Hamza, A.; Javed, A.R.R.; Iqbal, F.; Kryvinska, N.; Almadhor, A.S.; Jalil, Z.; Borghol, R. Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access* **2022**, *10*, 134018–134028. [CrossRef]
24. Massar, M.L.; Fickus, M.; Bryan, E.; Petkie, D.T.; Terzuoli, A.J. Fast computation of spectral centroids. *Adv. Comput. Math.* **2011**, *35*, 83–97. [CrossRef]
25. Li, T.; Ogihara, M.; Li, Q. A comparative study on content-based music genre classification. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Toronto, ON, Canada, 28 July–1 August 2003; pp. 282–289.
26. Lu, L.; Liu, D.; Zhang, H.-J. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio Speech Lang. Process.* **2005**, *14*, 5–18. [CrossRef]
27. Madhu, N. Note on measures for spectral flatness. *Electron. Lett.* **2009**, *45*, 1195–1196. [CrossRef]
28. Uddin, M.A.; Pathan, R.K.; Hossain, M.S.; Biswas, M. Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN. *J. Inf. Telecommun.* **2022**, *6*, 27–42. [CrossRef]
29. Kedem, B. Spectral analysis and discrimination by zero-crossings. *Proc. IEEE* **1986**, *74*, 1477–1493. [CrossRef]
30. Panagiotakis, C.; Tziritas, G. A speech/music discriminator based on RMS and zero-crossings. *IEEE Trans. Multimed.* **2005**, *7*, 155–166. [CrossRef]
31. Saunders, J. Real-time discrimination of broadcast speech/music. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; pp. 993–996.
32. Chuang, Z.-J.; Wu, C.-H. Emotion recognition using acoustic features and textual content. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No. 04TH8763), Taipei, Taiwan, 27–30 June 2004; pp. 53–56.
33. Mitrović, D.; Zeppelzauer, M.; Breiteneder, C. Features for content-based audio retrieval. In *Advances in Computers*; Elsevier: Amsterdam, The Netherlands, 2010; Volume 78, pp. 71–150.
34. Guglani, J.; Mishra, A. Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit. *Appl. Acoust.* **2020**, *167*, 107386. [CrossRef]
35. Ramakrishnan, A.; Abhiram, B.; Mahadeva Prasanna, S. Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *J. Acoust. Soc. Am.* **2015**, *137*, EL469–EL475. [CrossRef]
36. Sudhakar, R.S.; Anil, M.C. Analysis of Speech Features for Emotion Detection: A Review. In Proceedings of the 2015 International Conference on Computing Communication Control and Automation, Pune, India, 26–27 February 2015; pp. 661–664.
37. Abhang, P.A.; Gawali, B.W.; Mehrotra, S.C. Chapter 5—Emotion Recognition. In *Introduction to EEG-and Speech-Based Emotion Recognition*; Abhang, P.A., Gawali, B.W., Mehrotra, S.C., Eds.; Academic Press: Cambridge, MA, USA, 2016; pp. 97–112.
38. Johnson, E.K.; van Heugten, M.; Buckler, H. Navigating accent variation: A developmental perspective. *Annu. Rev. Linguist.* **2022**, *8*, 365–387. [CrossRef]
39. Wu, J.; Zhao, J. Systematic correspondence in co-evolving languages. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 469. [CrossRef]
40. Cassar, C.; McCabe, P.; Cumming, S. "I still have issues with pronunciation of words": A mixed methods investigation of the psychosocial and speech effects of childhood apraxia of speech in adults. *Int. J. Speech-Lang. Pathol.* **2023**, *25*, 193–205. [CrossRef]
41. Feng, S.; Halpern, B.M.; Kudina, O.; Scharenborg, O. Towards inclusive automatic speech recognition. *Comput. Speech Lang.* **2024**, *84*, 101567. [CrossRef]
42. Kaur, J.; Singh, A.; Kadyan, V. Automatic speech recognition system for tonal languages: State-of-the-art survey. *Arch. Comput. Methods Eng.* **2021**, *28*, 1039–1068. [CrossRef]
43. Karlsson, F.; Hartelius, L. On the primary influences of age on articulation and phonation in maximum performance tasks. *Languages* **2021**, *6*, 174. [CrossRef]
44. Bóna, J. Temporal characteristics of speech: The effect of age and speech style. *J. Acoust. Soc. Am.* **2014**, *136*, EL116–EL121. [CrossRef]
45. Das, B.; Mandal, S.; Mitra, P.; Basu, A. Effect of aging on speech features and phoneme recognition: A study on Bengali voicing vowels. *Int. J. Speech Technol.* **2013**, *16*, 19–31. [CrossRef]
46. Kennedy, J.; Lemaignan, S.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F.; Senft, E.; Belpaeme, T. Child speech recognition in human-robot interaction: Evaluations and recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; pp. 82–90.
47. Arslan, B.; Göksun, T. Aging, gesture production, and disfluency in speech: A comparison of younger and older adults. *Cogn. Sci.* **2022**, *46*, e13098. [CrossRef]

48. Ekström, S.; Pareto, L. The dual role of humanoid robots in education: As didactic tools and social actors. *Educ. Inf. Technol.* **2022**, *27*, 12609–12644. [CrossRef]
49. Carros, F.; Meurer, J.; Löffler, D.; Unbehaun, D.; Matthies, S.; Koch, I.; Wieching, R.; Randall, D.; Hassenzahl, M.; Wulf, V. Exploring human-robot interaction with the elderly: Results from a ten-week case study in a care home. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
50. Olde Keizer, R.A.; van Velsen, L.; Moncharmont, M.; Riche, B.; Ammour, N.; Del Signore, S.; Zia, G.; Hermens, H.; N'Dja, A. Using socially assistive robots for monitoring and preventing frailty among older adults: A study on usability and user experience challenges. *Health Technol.* **2019**, *9*, 595–605. [CrossRef]
51. Mendoza, E.; Valencia, N.; Muñoz, J.; Trujillo, H. Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *J. Voice* **1996**, *10*, 59–66. [CrossRef]
52. Pande, A.; Mishra, D. Humanoid robot as an educational assistant–insights of speech recognition for online and offline mode of teaching. *Behav. Inf. Technol.* **2024**, 1–18. [CrossRef]
53. Attawibulkul, S.; Kaewkamnerdpong, B.; Miyanaga, Y. Noisy speech training in MFCC-based speech recognition with noise suppression toward robot assisted autism therapy. In Proceedings of the 2017 10th Biomedical Engineering International Conference (BMEiCON), Hokkaido, Japan, 31 August–2 September 2017; pp. 1–5.
54. Meyer, J.; Dentel, L.; Meunier, F. Speech recognition in natural background noise. *PLoS ONE* **2013**, *8*, e79279. [CrossRef]
55. Agarwal, G.; Om, H. Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition. *Multimed. Tools Appl.* **2021**, *80*, 9961–9992. [CrossRef]
56. Doğdu, C.; Kessler, T.; Schneider, D.; Shadaydeh, M.; Schweinberger, S.R. A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech. *Sensors* **2022**, *22*, 7561. [CrossRef]
57. Ayrancı, A.A.; Atay, S.; Yıldırım, T. Speaker Accent Recognition Using Machine Learning Algorithms. In Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–6.
58. Mulfari, D.; Meoni, G.; Marini, M.; Fanucci, L. Machine learning assistive application for users with speech disorders. *Appl. Soft Comput.* **2021**, *103*, 107147. [CrossRef]
59. Abdusalomov, A.B.; Safarov, F.; Rakhimov, M.; Turaev, B.; Whangbo, T.K. Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm. *Sensors* **2022**, *22*, 8122. [CrossRef]
60. Li, F.; Liu, M.; Zhao, Y.; Kong, L.; Dong, L.; Liu, X.; Hui, M. Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J. Adv. Signal Process.* **2019**, *2019*, 59. [CrossRef]
61. Singh, V.; Prasad, S. Speech emotion recognition system using gender dependent convolution neural network. *Procedia Comput. Sci.* **2023**, *218*, 2533–2540. [CrossRef]
62. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
63. Sandhya, P.; Spoorthy, V.; Koolagudi, S.G.; Sobhana, N.V. Spectral Features for Emotional Speaker Recognition. In Proceedings of the 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India, 11–12 December 2020; pp. 1–6.
64. Micheyl, C.; Ryan, C.M.; Oxenham, A.J. Further evidence that fundamental-frequency difference limens measure pitch discrimination. *J. Acoust. Soc. Am.* **2012**, *131*, 3989–4001. [CrossRef]
65. Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and its applications: A Review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
66. Gourisaria, M.K.; Agrawal, R.; Sahni, M.; Singh, P.K. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discov. Internet Things* **2024**, *4*, 1. [CrossRef]
67. Shagi, G.U.; Aji, S. A machine learning approach for gender identification using statistical features of pitch in speeches. *Appl. Acoust.* **2022**, *185*, 108392. [CrossRef]
68. Agostini, G.; Longari, M.; Pollastri, E. Musical instrument timbres classification with spectral features. *EURASIP J. Adv. Signal Process.* **2003**, *2003*, 943279. [CrossRef]
69. Ferdoushi, M.; Paul, M.; Fattah, S.A. A Spectral Centroid Based Analysis of Heart sounds for Disease Detection Using Machine Learning. In Proceedings of the 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Banglore, India, 15–16 November 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
70. Ma, Y.; Nishihara, A. Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP J. Audio Speech Music. Process.* **2013**, *2013*, 87. [CrossRef]
71. Lazaro, A.; Sarno, R.; Andre, R.J.; Mahardika, M.N. Music tempo classification using audio spectrum centroid, audio spectrum flatness, and audio spectrum spread based on MPEG-7 audio features. In Proceedings of the 2017 3rd International Conference on Science in Information Technology (ICSITech), Bandung, Indonesia, 25–26 October 2017; pp. 41–46.
72. Gouyon, F.; Pachet, F.; Delerue, O. On the use of zero-crossing rate for an application of classification of percussive sounds. In Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy, 7–9 December 2000; p. 16.
73. Panda, S.K.; Jena, A.K.; Panda, M.R.; Panda, S. Speech emotion recognition using multimodal feature fusion with machine learning approach. *Multimed. Tools Appl.* **2023**, *82*, 42763–42781. [CrossRef]

74. Paul, B.; Bera, S.; Dey, T.; Phadikar, S. Machine learning approach of speech emotions recognition using feature fusion technique. *Multimed. Tools Appl.* **2024**, *83*, 8663–8688. [CrossRef]
75. Hammoud, M.; Getahun, M.N.; Baldycheva, A.; Somov, A. Machine learning-based infant crying interpretation. *Front. Artif. Intell.* **2024**, *7*, 1337356. [CrossRef]
76. Li, M.; Yang, B.; Levy, J.; Stolcke, A.; Rozgic, V.; Matsoukas, S.; Papayiannis, C.; Bone, D.; Wang, C. Contrastive Unsupervised Learning for Speech Emotion Recognition. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 6329–6333.
77. Zhang, Z.; Weninger, F.; Wöllmer, M.; Schuller, B. Unsupervised learning in cross-corpus acoustic emotion recognition. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 523–528.
78. Aldarmaki, H.; Ullah, A.; Ram, S.; Zaki, N. Unsupervised automatic speech recognition: A review. *Speech Commun.* **2022**, *139*, 76–91. [CrossRef]
79. Esfandian, N.; Razzazi, F.; Behrad, A. A clustering based feature selection method in spectro-temporal domain for speech recognition. *Eng. Appl. Artif. Intell.* **2012**, *25*, 1194–1202. [CrossRef]
80. Hajarolasvadi, N.; Demirel, H. 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy* **2019**, *21*, 479. [CrossRef]
81. Vyas, G.; Dutta, M.K. Automatic mood detection of indian music using mfccs and k-means algorithm. In Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 7–9 August 2014; pp. 117–122.
82. Bansal, S.; Dev, A. Emotional Hindi speech: Feature extraction and classification. In Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 11–13 March 2015; pp. 1865–1868.
83. Marupaka, P.T.; Singh, R.K. Comparison of classification results obtained by using cyclostationary features, MFCC, proposed algorithm and development of an environmental sound classification system. In Proceedings of the 2014 International Conference on Advances in Electronics Computers and Communications, Bangalore, India, 10–11 October 2014; pp. 1–6.
84. Poorna, S.S.; Jeevitha, C.Y.; Nair, S.J.; Santhosh, S.; Nair, G.J. Emotion recognition using multi-parameter speech feature classification. In Proceedings of the 2015 International Conference on Computers, Communications, and Systems (ICCCS), Kanyakumari, India, 2–3 November 2015; pp. 217–222.
85. Shadiev, R.; Hwang, W.-Y.; Chen, N.-S.; Huang, Y.-M. Review of speech-to-text recognition technology for enhancing learning. *J. Educ. Technol. Soc.* **2014**, *17*, 65–84.
86. Macháček, D.; Dabre, R.; Bojar, O. Turning Whisper into Real-Time Transcription System. *arXiv* **2023**, arXiv:2307.14743.
87. Vásquez-Correa, J.C.; Arzelus, H.; Martin-Doñas, J.M.; Arellano, J.; Gonzalez-Docasal, A.; Álvarez, A. When Whisper Meets TTS: Domain Adaptation Using only Synthetic Speech Data. In Proceedings of the International Conference on Text, Speech, and Dialogue, Pilsen, Czech Republic, 4–6 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 226–238.
88. Spiller, T.R.; Ben-Zion, Z.; Korem, N.; Harpaz-Rotem, I.; Duek, O. Efficient and Accurate Transcription in Mental Health Research-A Tutorial on Using Whisper AI for Sound File Transcription. *OSF Prepr.* **2023**. [CrossRef]
89. Liu, S.; Hu, S.; Liu, X.; Meng, H. On the Use of Pitch Features for Disordered Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 4130–4134.
90. De Cheveigné, A.; Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930. [CrossRef]
91. Giannakopoulos, T. *Pikrakis, A. Introduction to Audio Analysis: A MATLAB Approach*; Academic Press: Oxford, UK, 2014; p. i.
92. Ijaz, A.; Nabeel, M.; Masood, U.; Mahmood, T.; Hashmi, M.S.; Posokhova, I.; Rizwan, A.; Imran, A. Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey. *Inform. Med. Unlocked* **2022**, *29*, 100832. [CrossRef]
93. Krishnamurthi, R.; Gopinathan, D.; Kumar, A. Chapter 10—Using wavelet transformation for acoustic signal processing in heavy vehicle detection and classification. In *Autonomous and Connected Heavy Vehicle Technology*; Krishnamurthi, R., Kumar, A., Gill, S.S., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 199–209.
94. Torres-García, A.A.; Mendoza-Montoya, O.; Molinas, M.; Antelis, J.M.; Moctezuma, L.A.; Hernández-Del-Toro, T. Chapter 4—Pre-processing and feature extraction. In *Biosignal Processing and Classification Using Computational Learning and Intelligence*; Torres-García, A.A., Reyes-García, C.A., Villaseñor-Pineda, L., Mendoza-Montoya, O., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 59–91.
95. Shete, D.; Patil, S.; Patil, S. Zero crossing rate and Energy of the Speech Signal of Devanagari Script. *IOSR J. VLSI Signal Process. (IOSR-JVSP)* **2014**, *4*, 1–5. [CrossRef]
96. Bisong, E. Introduction to Scikit-learn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Bisong, E., Ed.; Apress: Berkeley, CA, USA, 2019; pp. 215–229.
97. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]
98. Abdalla, H.I. A Brief Comparison of K-means and Agglomerative Hierarchical Clustering Algorithms on Small Datasets. In Proceedings of the 2021 International Conference on Wireless Communications, Networking and Applications, Berlin, Germany, 17–19 December 2021; Springer Nature: Singapore, 2022; pp. 623–632.

99. Rathore, P.; Shukla, D. Analysis and performance improvement of K-means clustering in big data environment. In Proceedings of the 2015 International Conference on Communication Networks (ICCN), Gwalior, India, 19–21 November 2015; IEEE: New York, NY, USA, 2015; pp. 43–46.
100. Abbas, O.A. Comparisons between data clustering algorithms. *Int. Arab. J. Inf. Technol. (IAJIT)* **2008**, *5*, 320–325.
101. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
102. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]
103. Peng, K.; Leung, V.C.M.; Huang, Q. Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System Over Big Data. *IEEE Access* **2018**, *6*, 11897–11906. [CrossRef]
104. OpenAI Whisper. Available online: https://openai.com/research/whisper (accessed on 7 May 2023).
105. Openai-Whisper. Available online: https://pypi.org/project/openai-whisper/ (accessed on 21 April 2024).
106. Klakow, D.; Peters, J. Testing the correlation of word error rate and perplexity. *Speech Commun.* **2002**, *38*, 19–28. [CrossRef]
107. Filippidou, F.; Moussiades, L. A benchmarking of IBM, Google and Wit automatic speech recognition systems. In Proceedings of the Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, 5–7 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 73–82.
108. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Republic of Korea, 4–8 October 2004.
109. Vidal, E.; Toselli, A.H.; Ríos-Vila, A.; Calvo-Zaragoza, J. End-to-End page-Level assessment of handwritten text recognition. *Pattern Recognit.* **2023**, *142*, 109695. [CrossRef]
110. Pande, A.; Shrestha, B.; Rani, A.; Mishra, D. A Comparative Analysis of Real Time Open-Source Speech Recognition Tools for Social Robots. In Proceedings of the Design, User Experience, and Usability, Copenhagen, Denmark, 23–28 July 2023; Springer Nature: Cham, Switzerland, 2023; pp. 355–365.
111. Alghofaili, Y. Kmeans-Feature-Importance. Available online: https://github.com/YousefGh/kmeans-feature-importance (accessed on 20 April 2024).
112. Jiang, W.; Wang, Z.; Jin, J.S.; Han, X.; Li, C. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors* **2019**, *19*, 2730. [CrossRef] [PubMed]
113. Chauhan, N.; Isshiki, T.; Li, D. Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019; pp. 130–133.
114. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
115. Lehner, B.; Sonnleitner, R.; Widmer, G. Towards Light-Weight, Real-Time-Capable Singing Voice Detection. In Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR), Curitiba, Brazil, 4–8 November 2013; pp. 1–6.
116. Gajic, B.; Paliwal, K.K. Robust speech recognition in noisy environments based on subband spectral centroid histograms. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 600–608. [CrossRef]
117. Paliwal, K.K. Spectral subband centroid features for speech recognition. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), Seattle, WA, USA, 15 May 1998; IEEE: New York, NY, USA, 1998; pp. 617–620.
118. Huang, Y.; Ao, W.; Zhang, G. Novel sub-band spectral centroid weighted wavelet packet features with importance-weighted support vector machines for robust speech emotion recognition. *Wirel. Pers. Commun.* **2017**, *95*, 2223–2238. [CrossRef]
119. Qadri, S.A.A.; Gunawan, T.S.; Wani, T.; Alghifari, M.F.; Mansor, H.; Kartiwi, M. Comparative Analysis of Gender Identification using Speech Analysis and Higher Order Statistics. In Proceedings of the 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), Kuala Lumpur, Malaysia, 27–29 August 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
120. Sebastian, J.; Kumar, M.; Murthy, H.A. An analysis of the high resolution property of group delay function with applications to audio signal processing. *Speech Commun.* **2016**, *81*, 42–53. [CrossRef]
121. Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. *Int. J. Speech Technol.* **2020**, *23*, 45–55. [CrossRef]
122. Chauhan, N.; Isshiki, T.; Li, D. Text-Independent Speaker Recognition System Using Feature-Level Fusion for Audio Databases of Various Sizes. *SN Comput. Sci.* **2023**, *4*, 531. [CrossRef]
123. Bird, J.J.; Faria, D.R.; Premebida, C.; Ekárt, A.; Ayrosa, P.P. Overcoming data scarcity in speaker identification: Dataset augmentation with synthetic mfccs via character-level rnn. In Proceedings of the 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Ponta Delgada, Portugal, 15–17 April 2020; IEEE: New York, NY, USA, 2020; pp. 146–151.
124. Shen, Z.; Elibol, A.; Chong, N.Y. Inferring human personality traits in human-robot social interaction. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Republic of Korea, 11–14 March 2019; IEEE: New York, NY, USA, 2019; pp. 578–579.

125. Li, N.; Ross, R. Invoking and identifying task-oriented interlocutor confusion in human-robot interaction. *Front. Robot. AI* **2023**, *10*, 1244381. [CrossRef] [PubMed]
126. Telembici, T.; Grama, L.; Muscar, L.; Rusu, C. Results on the MFCC extraction for improving audio capabilities of TIAGo service robot. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 25–27 October 2021; IEEE: New York, NY, USA, 2021; pp. 57–61.
127. Wu, X.; Gong, H.; Chen, P.; Zhong, Z.; Xu, Y. Surveillance robot utilizing video and audio information. *J. Intell. Robot. Syst.* **2009**, *55*, 403–421. [CrossRef]
128. Hireche, A.; Belkacem, A.N.; Jamil, S.; Chen, C. NewsGPT: ChatGPT Integration for Robot-Reporter. *arXiv* **2023**, arXiv:2311.06640.
129. Pépiot, E. Voice, speech and gender: Male-female acoustic differences and cross-language variation in english and french speakers. In Proceedings of the 15th Rencontres Jeunes Chercheurs (RJC 2012), Paris, France, 15–16 June 2012. [CrossRef]
130. Tsantani, M.S.; Belin, P.; Paterson, H.M.; McAleer, P. Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception* **2016**, *45*, 946–963. [CrossRef] [PubMed]
131. Garnerin, M.; Rossato, S.; Besacier, L. Gender representation in French broadcast corpora and its impact on ASR performance. In Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, Nice, France, 21 October 2019; pp. 3–9.
132. Adda-Decker, M.; Lamel, L. Do speech recognizers prefer female speakers? In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
133. Tatman, R. Gender and dialect bias in YouTube's automatic captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 4 April 2017; pp. 53–59.
134. Doddington, G.R.; Przybocki, M.A.; Martin, A.F.; Reynolds, D.A. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. *Speech Commun.* **2000**, *31*, 225–254. [CrossRef]
135. Rodrigues, A.; Santos, R.; Abreu, J.; Beça, P.; Almeida, P.; Fernandes, S. Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. In Proceedings of the XX International Conference on Human Computer Interaction, Donostia Gipuzkoa, Spain, 25–28 June 2019; pp. 1–8.
136. Nematollahi, M.A.; Al-Haddad, S.A.R. Distant speaker recognition: An overview. *Int. J. Humanoid Robot.* **2016**, *13*, 1550032. [CrossRef]
137. Michael, D.D.; Siegel, G.M.; Pick, H.L., Jr. Effects of distance on vocal intensity. *J. Speech Lang. Hear. Res.* **1995**, *38*, 1176–1183. [CrossRef] [PubMed]
138. Zahorik, P.; Kelly, J.W. Accurate vocal compensation for sound intensity loss with increasing distance in natural environments. *J. Acoust. Soc. Am.* **2007**, *122*, EL143–EL150. [CrossRef]
139. Diaz-Asper, C.; Chandler, C.; Turner, R.S.; Reynolds, B.; Elvevåg, B. Acceptability of collecting speech samples from the elderly via the telephone. *Digit. Health* **2021**, *7*, 20552076211002103. [CrossRef]
140. Li, Q.; Russell, M.J. An analysis of the causes of increased error rates in children's speech recognition. In Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002.