IEEE Signal Processing Cup 2024

Team Wavemasters

Nidula Gunawardana, Lasindu Dilshan, Nilupulee Amarathunga, Erandee Jayathilaka, Dasuni Herath, Nimshi Wanniarachchi, Mavishan Pasira, Chandeepa Janith, Shemal Perera, Mihiraja Kuruppu

Department of Electronic and Telecommunication Engineering
University of Moratuwa, Sri Lanka

1 Abstract

The IEEE Signal Processing Cup 2024 challenges participants to address the problem of text-independent far-field speaker recognition under noise and reverberation conditions for mobile robots. This paper presents the approach and results of Team Wavemasters in tackling this challenge. Leveraging signal processing techniques and machine learning algorithms, our solution aims to robustly identify speakers from speech signals captured by mobile robots amidst challenging acoustic environments. Key components of our approach include feature extraction, noise mitigation, and speaker verification methods tailored for real-world scenarios. Through rigorous experimentation and evaluation on the provided dataset, our solution demonstrates promising performance in addressing the complexities of far-field speaker recognition for mobile robotics applications.

2 Introduction

Speaker recognition systems have evolved rapidly, propelled by advancements in deep learning and signal processing techniques. These systems play a vital role in verifying the identity of individuals based on their speech characteristics, finding applications in security, access control, and personalized services. However, existing speaker recognition systems often struggle in far-field environments, where noise, reverberation, and variable speaker-to-microphone distances introduce significant challenges.

The IEEE Signal Processing Cup 2024 presents a novel challenge in text-independent far-field speaker recognition for mobile robots. In this competition, participants are tasked with developing robust speaker verification pipelines capable of accurately identifying speakers from speech signals captured by mobile robots under adverse conditions. The competition seeks to bridge the gap between theoretical advancements in speaker recognition and real-world applications, particularly in dynamic environments where mobile robots interact with human speakers amidst varying acoustic conditions.

Team Wavemasters approaches this challenge with a combination of signal processing algorithms and machine learning models tailored to the characteristics of far-field speech signals. Our solution aims to extract discriminative features from noisy and reverberant speech signals, mitigate the effects of environmental factors, and perform speaker verification with high accuracy and efficiency. By addressing the specific challenges posed by mobile robotics environments, our solution aims to advance the state-of-the-art in far-field speaker recognition technology and pave the way for practical applications in real-world settings.

3 Methodology

3.1 Noise removing and amplifying method using Wavelet Transform

We have used wavelet transform, a mathematical tool for analyzing signals and images in terms of different frequency components to process the audio signal. Unlike the Fourier transform, which represents a signal in the frequency domain, the wavelet transform represents both frequency and time information. This makes wavelet transform particularly useful for analyzing signals with non-stationary or time-varying characteristics.

Signal Denoising using Wavelet Transform:

1. **Decomposition:**
   The initial phase of the signal denoising process involved the application of the Discrete Wavelet Transform (DWT). This transformative step decomposes the input signal into a set of approximation and detail coefficients at multiple scales. The approximation coefficients encapsulate the coarse, low-frequency components of the signal, while the detail coefficients capture the nuanced, high-frequency details.

2. **Thresholding:**
   After decomposition, our next step in the denoising methodology was the application of thresholding to the detail coefficients. Soft and hard thresholding techniques were employed for this purpose. Soft thresholding selectively sets coefficients below a predetermined threshold to zero, effectively attenuating noise and contributing to a smoother denoised signal. On the other hand, hard thresholding zeros coefficients fall below the specified threshold, thereby eliminating noise from the signal.
3. **Reconstruction:**
   Following the thresholding phase, the denoised signal was reconstructed using the modified coefficients. Notably, in the reconstruction process, we omitted the approximation coefficients from the highest scale. This omission was deemed acceptable as high-frequency details are considered non-critical for the denoised output.

4. **Iterative Processing :**
   The above steps were iteratively applied to enhance the denoising performance further.

Wavelet-based denoising is effective when the signal and noise have different frequency characteristics. Choosing an appropriate wavelet, thresholding method, and threshold value were critical in achieving optimal denoising results.

Note: The sampling rate used in both noisy and denoised signals is 16kHz. Therefore we did not experience any time scaling.

Just to have an idea, we have provided the time domain and frequency domain representations of one noisy and denoised signal below:

3.2 System Design

3.2.1 Feature Extraction

We focused on robust feature extraction from audio soundtracks to lay the groundwork for subsequent model training. We adopted the first approach, utilizing Mel-Frequency Cepstral Coefficients (MFCCs) as a representative feature set for audio signal characterization in the frequency domain. The MFCC plot served as a visual representation, depicting variations in the spectral content over time. This plot encapsulates crucial information such as time versus frequency dynamics, aiding in the observation of distinct sound patterns and the extraction of relevant features.

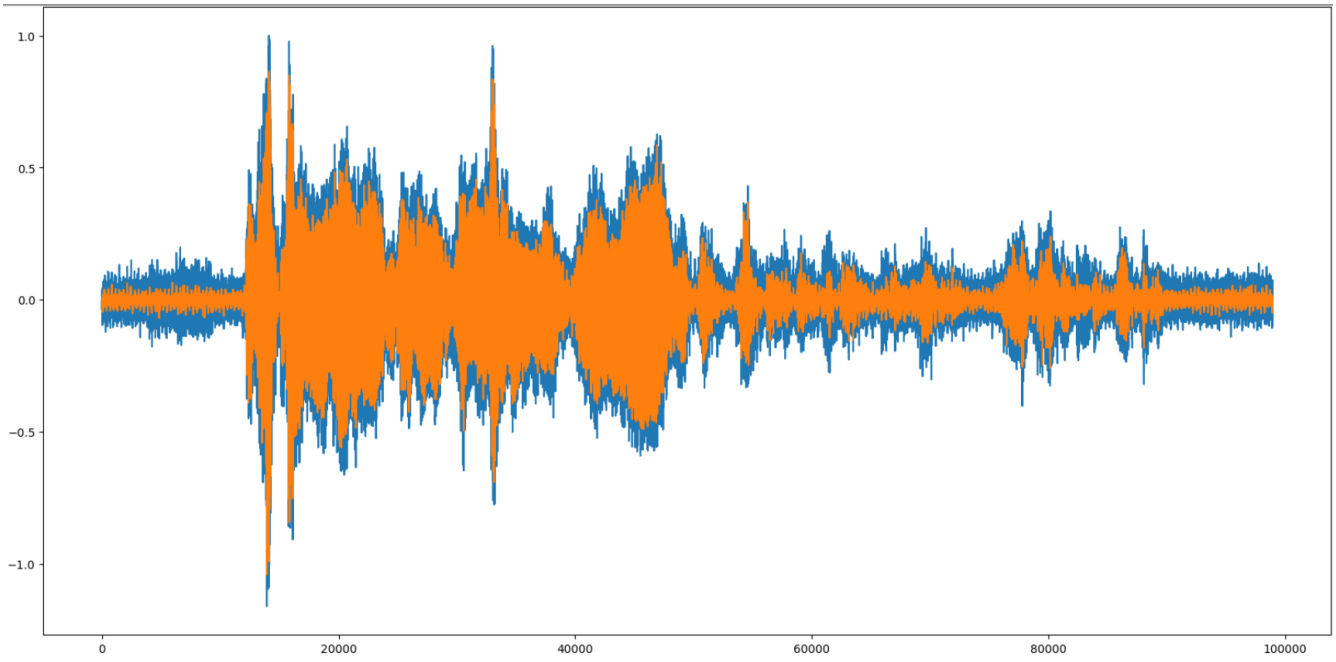

Figure 1: *Time Domain Representation of Noisy and Denoised Audio Signals.*

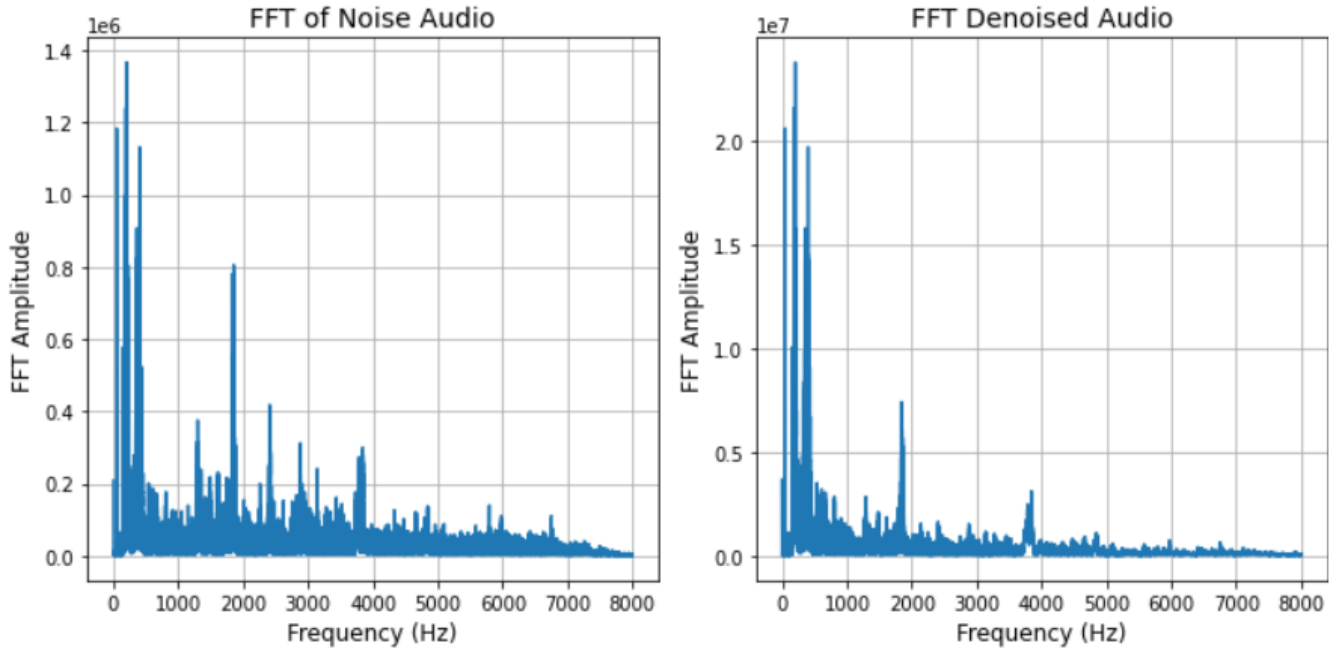

Figure 2: *FFT Representation of Noisy and Denoised Audio Signals.*

The MFCC coefficients, crucial in audio processing tasks, were extracted using the librosa library's dedicated MFCC feature extraction method. Each row in the resulting array represented the MFCC coefficients for a specific time frame, while each column corresponded to a distinct MFCC feature. The significance of individual coefficients was detailed, ranging from capturing overall energy (MFCC 0) to nuanced spectral characteristics (higher-order coefficients). The number of coefficients could be adjusted, with a maximum stable limit of 40, allowing for flexibility in representing the audio signal's complexity.

This feature extraction methodology served as the foundation for subsequent model training, providing a rich representation of audio characteristics essential for tasks such as speaker identification, emotion recognition, and speech-to-text.

3.2.2 Data Augmentation Techniques

To address the limited availability of training data, data augmentation techniques were employed to generate additional soundtracks. Techniques such as Gaussian noise addition, time stretching, pitch scaling, and random gain modulation were implemented to augment the existing dataset. Each technique was carefully calibrated to produce diverse audio samples while preserving the underlying characteristics of the original data. The augmented dataset provided a larger pool of training examples, thereby enhancing the model's ability to generalize to unseen data. Model performance was evaluated based on accuracy improvements achieved through data augmentation.

3.2.3 VGGish Pre-trained Model

We used a pre-trained model called VGGish from TensorFlow, specifically designed for audio classification tasks. VGGish is capable of extracting high-level semantic embeddings from audio input, providing a compact representation that preserves essential features for downstream classification tasks. The pre-trained model was utilized to extract 128-dimensional embeddings from denoised test files, which were subsequently fed into a downstream classification model. Model performance was evaluated based on the accuracy of speaker identification achieved using the VGGish embeddings. Cosine similarity calculations were performed to assess the similarity between enrollment clips and test files, providing insights into the effectiveness of the VGGish model for speaker identification tasks.

4 Experiment and Results

This study has used python TensorFlow software for implementation. Experiments are carried out on 78 different speakers featuring between 24 and 36 conversations per speaker, resulting in a total of 2219 conversations. On average each conversation consists of 5 dialogues, totaling approximately 11000 recorded dialogues. The database consists of speech signals, sample data frequency of 16KHz and of duration within 3.6seconds and recordings are made with 8 channels. Each speaker engaged in conversations with the robot, introducing complexities such as ambient noise, internal robot noises, reverberation, varying distances, babble noise, and speaker angles.

Our approach commenced with robust preprocessing, including wavelet-based denoising using Discrete Wavelet Transform (DWT). This method effectively handled non-stationary characteristics, resulting in enhanced denoised signals crucial for downstream analysis. We then focused on feature extraction, utilising Mel-Frequency Cepstral Coefficients (MFCCs) and VGGish pre-trained models from TensorFlow. The choice of MFCCs provided a comprehensive representation of audio characteristics, while VGGish embeddings captured high-level semantic features. Data augmentation techniques, such as Gaussian noise addition and time stretching, were employed to address limited training data, enhancing the model's ability to generalise. The VGGish pre-trained model, specifically designed for audio classification, demonstrated its proficiency in extracting 128-dimensional embeddings for downstream tasks.

Evaluation results across different scenarios highlighted the robustness of our solution. In text-independent far-field speaker recognition, our model showcased promising accuracy rates under challenging conditions. The experiments, including short and long utterance tests, validated the effectiveness of our approach in handling diverse scenarios presented in the Robovox dataset. Team Wavemasters' experiments yielded favourable results, showcasing the applicability and efficacy of our solution in addressing the complexities of far-field speaker recognition for mobile robots.

5 Discussion and Conclusion

The speaker recognition system serves the purpose of measuring the distance between two audio files, utilizing embeddings generated for each speaker and the respective audio file. The cosine distance calculation between these embeddings provides a quantitative measure of the similarity or dissimilarity between the audio files.

The Mel-Frequency Cepstral Coefficients (MFCCs) and embeddings derived from the VGGish model are vital for the analysis as they demonstrate crucial features within each audio file. The comparative evaluation of these features enables a detailed assessment of the similarity between any two given audio files.

To enhance the precision of the results, the system employs robust noise removal methods and ensures the accurate extraction of features. These measures contribute significantly to achieving more refined and reliable outcomes in the recognition process.

The collaborative experience gained as a team during the development of this system is invaluable in effectively addressing such challenges.

6 References

[1] Homayoon Beigi. (2011) "Fundamentals of Speaker recognition." Springer science and Business media.

[2] Sirko Molau, Michael Pitz, Ralf Schluter, and Hermann Ney. (2001) "Computing Mel-frequency Cepstral Coefficients on the power spectrum." IEEE Transactions on Audio, Speech and Language Processing.

[3] Chanwoo Kim and Richard M. Stern. (2016) "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition." IEEE Transactions on Audio, Speech and Language Processing, 1315–1329.

[4] Hynek Hermansky and Nelson Morgan. (1994) "RASTA Processing of Speech" IEEE Transactions on Speech and Audio Processing, 578–589.

[5] Thomas Ulbrich Christiansen. (1986) "The Meddis Inner Hair-Cell Model" Journal of the Acoustical Society of America, vol. 79.

[6] Siddhant C. Joshi and Dr. A.N. Cheeran. (2014) "MATLAB Based Feature Extraction Using Mel-Frequency Cepstrum Coefficients for Automatic Speech Recognition." International Journal of Science, Engineering, and Technology Research (IJSETR), Volume 3, Issue 6.

[7] Malcolm Slaney. (1993) "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank." Apple Computer Technical Report 35 Perception Group-Advanced Technology Group.

[8] K. Sreenivasa Rao and Shashidhar G. Koolagudi. (2013) "Robust Emotion Recognition using Pitch Synchronous and Sub-syllabic Spectral Features." Springer Briefs in Electrical and Computer Engineering, New York pp 17–46.

[9] D. A. Reynolds and R. C. Rose. (1995) "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models." IEEE Trans. on Speech and Audio Processing, vol.3, No.1, pp.72–83.

[10] Mohammed Senoussaoui, Patrick J. Kenny, Najim Dehak, Pierre Dumouchel. (2011) "An i-vector Extractor for speaker Recognition with Microphone and Telephone Speech." CSAIL-MIT.

[11] Wei Li , Tianfan Fu and Jie Zhu. (2015) "An improved i-vector extraction algorithm for speaker verification Transient Flow." EURASIP journal on Audio, Speech and Music Processing, pp 1-9.

[12] Najim Dehak, Patrick J. Kenny, Dehak, Pierre Dumouchel and Pierre Ouellet. (2011) "Front-End Factor Analysis for Speaker Verification." Transactions on Audio, Speech and Language Processing, vol.24, No.3.

[13] Padmanabhan Rajana b, Anton Afanasyeva, Ville Hautamakia and Tomi Kinnunena. (2014) "From single to multiple enrollment i-vectors: practical PLDA scoring variants for speaker verification." Digital Signal Processing.

[14] Dominic Mathew, V.D.Devessia and Tessamma Thomas. (2006) "A K-means Clustering Algorithm for Frequency Estimation and Classification of Speech Signals." IEEE conference/ICSIP.

[15] Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.-M., 2009. Wavelet Toolbox TM 4 User ' s Guide. The MathWorks Inc., . . . , 11–47. Retrieved from http://feihu.eng.ua.edu/NSF$_T UES/w7_1 a.pdf http : //feihu.eng.ua.edu/NSF_TUES/w7$

[16] Patil, R., 2015. Noise Reduction using Wavelet Transform and Singular Vector Decomposition. Procedia Computer Science, 54, 849–853. https://doi.org/10.1016/j.procs.2015.06.099

[17] Patil, S. S., and Pawar, M. K., 2012. Quality advancement of EEG by wavelet denoising for biomedical analysis. Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012, 1–6. https://doi.org/10.1109/ICCICT.2012.6398151https://doi.org/10.1109/ICCICT.2012.6398151

[18] Polikar, R. 1994. The Wavelet Tutorial. Internet Resources, 1–67. https://doi.org/10.1088/1751-8113/44/8/085201

[19] Yadav, T. 2016. Denoising and SNR Improvement of ECG Signals Using Wavelet Based Techniques, (October), 678–682

[20] Bouman, C. A., 2013. Continuous Time Fourier Transform (CTFT). Digital Image Processing, 1–5.

[21] Hazas, M., and Hall, H., 1999. Processing of Non-Stationary Audio Signals. Science, (August).

[22] 2024. [Online]. Available:https://in.mathworks.com/help/wavelet/ug/wavelet-denoising.html Wavelet Denoising - MATLAB & Simulink - MathWorks India.

7 About Wavemasters

Dr. Subodha Charles works at the University of Moratuwa in Sri Lanka's Department of Electronic and Telecommunication Engineering. He obtained his B.Sc. with a focus on electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, and his Ph.D. in computer science from the University of Florida, USA. Computer architecture, embedded systems, and hardware security and trust are among his areas of interest. More than ten of his research publications have been published in prestigious international conferences and journals in those fields. He gained experience in the field at Zone24x7 in Sri Lanka and Intel in the United States.

Nidula Gunawardana, Chandeepa Peiris, Nimshi Wanniarachchi, Erandee Jayathilaka, Nilupulee Amarathunga, Mavishan Pasira, Dasuni Herath, Shemal Perera, Mihiraja Kuruppu, Lasindu Dilshan are undergraduates (level 2) of the Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka. They are passionate students interested in image and signal processing, algorithms and embedded electronics.