Improving Speech Recognition with the Use of Context, Multitasking and Transfer Learning

***Baby Priya**
****Asst. Prof. Ayush Kumar**

ABSTRACT

The speech signal is inherently redundant and non-stationary, but due to the inertness of the vocal tract, its variations occur relatively slowly. This allows the signal to be treated as stationary over short segments. It is widely accepted that the most distinctive information in speech is captured in the short-time magnitude spectrum, which forms the basis for analyzing speech signals in a frame-by-frame manner. For this purpose, the speech signal is divided into overlapping segments, or frames, typically lasting 15–25 milliseconds, with overlaps of 10–15 milliseconds.

This paper investigates the influence of analysis window length and frame shift on speech recognition performance. The study evaluates three distinct cepstral analysis methods: mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC) and proposed perceptual linear predictive cepstral coefficients (PPLPC). The results demonstrate that the optimal speech recognition rate is achieved with an analysis window length of 10 milliseconds and a frame shift ranging from 7.5 to 10 milliseconds, irrespective of the analysis method used. Notably, the highest improvement in recognition rate was 2.5%.

These findings underline the critical role of proper parameter selection in cepstral analysis and its direct impact on enhancing the accuracy of speech recognition systems. The study emphasizes the importance of tailoring analysis settings to optimize performance, providing valuable insights for the development of more effective speech recognition models.

Keywords:- *Natural Language Processing (NLP), Computers and information processing; Speech Analysis; Speech Recognition; Speech Enhancement, Machine Learning.*

*Baby Priya, Research Scholar, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal , bpriya16cs65@gmail.com
**Asst. Prof. Ayush Kumar, Assistant Professor, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal, India

I. INTRODUCTION

As Speech analysis serves as the cornerstone of the speech recognition process, where features with high discriminative power are extracted to facilitate accurate and efficient recognition. These features are intended to convey specific linguistic information from a speech utterance, ideally without encoding attributes such as the speaker's gender, age, speaking style, or physical condition. However, in real-world scenarios, achieving this ideal is challenging, and selecting features robust to noise and speaker variability remains a critical issue in speech recognition systems. Extensive research has been devoted to developing effective feature extraction and

enhancement techniques to ensure reliable and noise-robust speech recognition systems [1], [2], [3].

Various feature systems, both static and dynamic, have been proposed, including time-domain features (e.g., frame energy, zero-crossing rate), frequency-domain features (e.g., formant-based and spectral pairs), cepstral domain features, and features modeled on human auditory perception. The first step in any feature extraction process involves segmenting the speech signal into analysis frames—overlapping signal segments—allowing the inherently non-stationary speech signal to be represented as quasi-stationary segments. Typical frames have a constant length of 15–25 milliseconds with an overlap of 10–15 milliseconds.

While these parameters have been widely used and considered effective for decades [3], [4], [5], they merit reevaluation in light of advancements in speech signal processing. Over the years, the sampling rate of speech recordings has increased significantly, from 6 kHz [6] to as high as 44.1 kHz, resulting in higher spectral resolution for the same frame window length. However, higher resolution does not necessarily translate to better speech recognition performance. In some cases, it may inadvertently encode speaker-specific information, which can detract from system robustness. Thus, the selection of analysis parameter values must be reconsidered, taking into account modern approaches to speech signal analysis.

This paper explores the manipulation of analysis parameters as a means to enhance recognition rates. The study evaluates the impact of different parameter configurations on speech recognition performance, aiming to identify optimal values that improve recognition accuracy without necessitating substantial modifications to the underlying system. This investigation provides valuable insights into refining speech analysis techniques for contemporary applications.

II. LITERATURE REVIEW

Joshi Joshi Padi et al. (2021) propose an enhanced method for speech emotion recognition by combining transfer learning with spectrogram augmentation [6]. Their approach leverages transfer learning to boost the model's capability in identifying emotions in speech, especially in low-resource settings where data is limited. By using pre-trained models on large datasets, the system can transfer knowledge to the emotion recognition task, improving its performance. Additionally, spectrogram augmentation techniques are applied to diversify the training data, further enhancing the model's robustness. This dual strategy significantly improves emotion recognition accuracy, making it more effective in real-world applications where emotional nuance is critical.

Rolland et al. (2022) investigate the application of multilingual transfer learning to enhance automatic speech recognition (ASR) systems for children [7]. Their study highlights the unique challenges of recognizing children's speech, such as higher pitch, variability in pronunciation, and limited training data. By leveraging transfer learning, they develop ASR models that can effectively adapt to children's speech patterns across various languages. This approach involves utilizing pre-trained models on adult speech and transferring the learned knowledge to improve the recognition accuracy for children's speech. Their findings demonstrate that multilingual transfer learning significantly enhances the performance of ASR systems for children, making them more reliable and effective in diverse linguistic contexts.

Latif et al. (2022) introduce a multitask learning approach that utilizes augmented auxiliary data to improve speech emotion recognition [8]. Their method integrates auxiliary data sources alongside multiple related tasks to enhance the model's capability in accurately detecting emotions in speech. By combining these elements, the approach leverages diverse data and task interactions, which collectively contribute to a more robust and precise emotion recognition system. This multitask learning framework allows the model to generalize better across different emotional contexts and improves its performance even in scenarios with limited labeled data. The study underscores the effectiveness of using augmented auxiliary data and multitask learning techniques to advance the accuracy of emotion recognition in speech applications.

Sullivan et al. (2022) examine the use of transfer learning and language model decoding to enhance automatic speech recognition (ASR) for non-native English speakers [9]. Their study focuses on adapting ASR systems to better handle the unique challenges associated with non-native speech, such as diverse accents and pronunciation variations. By applying transfer learning, they leverage pre-trained models to improve recognition accuracy for non-native speech patterns. Additionally, their approach incorporates advanced language model decoding techniques to refine the system's ability to understand and transcribe non-native speech more effectively. This dual strategy addresses common issues in ASR for non-native speakers, such as mispronunciations and accent-related distortions, resulting in a more reliable and accurate speech recognition system for diverse linguistic backgrounds.

Padi et al. (2022) present a multimodal emotion recognition approach that utilizes transfer learning from speaker recognition and BERT-based models [10]. Their method integrates various modalities to enhance the precision of emotion recognition in speech. By leveraging transfer learning from pre-trained speaker recognition models and BERT-based language models, the

approach improves the system's ability to identify and interpret emotional cues with greater accuracy. Combining these modalities allows the model to capture a broader range of emotional signals and contextual nuances, leading to more reliable emotion detection. This study highlights the effectiveness of using transfer learning in multimodal applications, demonstrating its potential to significantly advance emotion recognition technology by incorporating diverse sources of information.

Yadav and Sitaram (2022) present a comprehensive survey on multilingual models for automatic speech recognition (ASR), focusing on different transfer learning and multi-task learning strategies [11]. Their review explores the various techniques and methodologies used to develop ASR systems that can effectively handle multiple languages. They discuss the challenges encountered in multilingual ASR, such as managing linguistic diversity, accent variations, and limited data resources, while also highlighting recent advancements in the field. The survey provides an overview of how transfer learning can enhance cross-lingual performance and how multi-task learning approaches can improve language adaptability and accuracy. By covering both the current state-of-the-art methods and ongoing issues, their survey offers valuable insights into the evolution and future directions of multilingual ASR systems.

Kheddar et al. (2023) explore deep transfer learning techniques to advance automatic speech recognition (ASR), with a focus on enhancing generalization across various languages and domains [12]. Their research highlights how transfer learning can significantly improve the adaptability and robustness of ASR systems. By leveraging deep learning models pre-trained on large, diverse datasets, their approach enables ASR systems to better handle linguistic and contextual variations. The study demonstrates that transfer learning enhances the system's performance by allowing it to generalize more effectively to new languages and domains, thereby improving accuracy and reliability. This work underscores the potential of deep transfer learning to address challenges in ASR and make systems more versatile and resilient in real-world applications.

Steinmetz (2023) investigates the application of transfer learning with L2 speech to enhance automatic speech recognition (ASR) for dysarthric speech [13]. Their study focuses on transferring knowledge from models trained on non-dysarthric speech to those recognizing dysarthric speech, which is affected by motor disorders. This approach aims to improve ASR systems' ability to accurately understand and transcribe speech that has been impaired due to motor difficulties. By leveraging pre-existing models trained on clearer, non-dysarthric speech,

the method helps address the specific challenges associated with dysarthric speech, such as altered articulation and reduced clarity. Steinmetz's work demonstrates how transfer learning can be effectively utilized to bridge the gap between different types of speech, ultimately enhancing the recognition accuracy for individuals with speech impairments.

Nga et al. (2023) introduce a cyclic transfer learning approach to enhance Mandarin-English code-switching speech recognition [14]. Their method employs cyclic training between the two languages to boost recognition accuracy in code-switching situations, where speakers alternate between Mandarin and English within the same conversation. By iteratively training the model on both languages, this approach effectively captures the nuances and complexities of bilingual speech patterns. The cyclic process helps the model better understand and process the transitions between languages, improving its ability to handle mixed-language inputs. This study demonstrates the effectiveness of cyclic transfer learning in addressing the unique challenges posed by code-switching, ultimately leading to more accurate and reliable speech recognition in bilingual contexts.

Tun et al. (2023) explore the use of multimodal transfer learning for assessing oral presentations [15]. Their approach integrates both speech and visual data to improve the evaluation process. By combining these modalities, their method enhances the accuracy and comprehensiveness of oral presentation assessments. This multimodal approach leverages transfer learning techniques to incorporate pre-trained models from different domains, allowing for a more nuanced analysis of presentation skills. The study highlights how transfer learning can be effectively applied in educational settings to assess various aspects of oral presentations, such as delivery, clarity, and visual engagement. Tun et al.'s work underscores the potential of multimodal transfer learning to provide richer, more detailed evaluations, showcasing its value in enhancing educational assessment methods.

Zheng and Zhang (2023) introduce an enhanced multi-label transfer learning model designed for intelligent speech systems [16]. Their approach tackles the complexities of multi-label classification in speech applications, where multiple labels or categories need to be assigned to a single speech input. By leveraging transfer learning, their model effectively utilizes pre-trained knowledge to improve performance in these challenging scenarios. This improvement addresses issues such as overlapping labels and context-dependent classifications, showcasing how transfer learning can be adapted to handle intricate tasks in speech recognition. The study highlights the

versatility and effectiveness of transfer learning in managing complex multi-label problems, demonstrating its potential to advance the capabilities of intelligent speech systems.

Zhou et al. (2024) present a multitask co-training framework designed to enhance speech translation by simultaneously leveraging speech recognition and machine translation tasks [17]. Their innovative approach integrates these tasks to improve both the accuracy and fluency of translated speech outputs. By training models on multiple related tasks, the framework enables better alignment between spoken input and its translated text, resulting in more coherent and contextually accurate translations. This multitask co-training method helps capture the nuances of both speech recognition and translation processes, leading to improvements in overall performance. The study demonstrates how combining these tasks can address the challenges of translating spoken language, highlighting the potential for more effective and natural-sounding speech translation systems.

Ta and Le (2024) examine transfer learning techniques aimed at enhancing speech accent recognition in low-resource environments, with a focus on the Vietnamese language [18]. Their study illustrates how transfer learning can significantly boost recognition accuracy for accented speech when resources are limited. By leveraging pre-trained models and adapting them to the specific nuances of Vietnamese accents, their approach addresses the challenges of limited training data and accent variability. The research demonstrates that transfer learning effectively transfers knowledge from more resource-rich contexts to improve performance in recognizing accented speech. This approach not only improves the accuracy of speech recognition systems in low-resource settings but also highlights the potential of transfer learning to address language-specific challenges in accent recognition.

Kheddar et al. (2024) present a comprehensive survey on cutting-edge deep learning methods for automatic speech recognition (ASR) [19]. Their review explores a range of advanced techniques, focusing on transfer learning and multi-task learning, and highlights the most recent advancements in ASR technology. By examining the latest developments, their survey provides insights into how these techniques are being applied to enhance ASR systems, including improvements in accuracy, adaptability, and efficiency. The study covers various strategies for leveraging pre-trained models and integrating multiple tasks to address complex speech recognition challenges. This survey serves as a valuable resource for understanding current trends and future directions in ASR, showcasing how deep learning innovations are driving progress in the field.

Hassan et al. (2024) introduce a deep bidirectional LSTM model that is enhanced through transfer-learning-based feature extraction for dynamic human activity recognition [20]. Their method integrates transfer learning with deep learning techniques to significantly improve the accuracy of recognizing various human activities. By leveraging pre-trained models for feature extraction, their approach enables the LSTM network to better capture and interpret complex activity patterns. This combination enhances the model's ability to accurately monitor and classify dynamic activities, even in diverse and challenging conditions. Their study demonstrates how incorporating transfer learning can refine deep learning models and boost performance in activity recognition applications, offering a more reliable solution for monitoring human activities in real-world scenarios.

Kumar and Yadav (2024) investigate the use of multi view learning techniques to improve speech recognition for low-resource languages [21]. Their approach harnesses multiple perspectives or "views" of the data to enhance the accuracy of automatic speech recognition (ASR) systems. By integrating diverse types of data, such as acoustic features and linguistic information, their method addresses the specific challenges associated with recognizing speech in languages with limited resources. This multi view learning framework helps overcome the scarcity of training data and the unique linguistic characteristics of low-resource languages, leading to more accurate and robust speech recognition. Their study demonstrates how leveraging multiple data views can significantly improve ASR performance, making it a promising solution for enhancing recognition capabilities in underrepresented languages.

III. PROPOSED WORK

Our workflow for collecting and analyzing speech recognition system consists of five distinct phases:

A. Speech Signal acquisition

B. Feature Extraction

C. Acoustic Modelling

D. Language & Lexical Modelling

E. Recognition

These phases are shown in Figure 1.

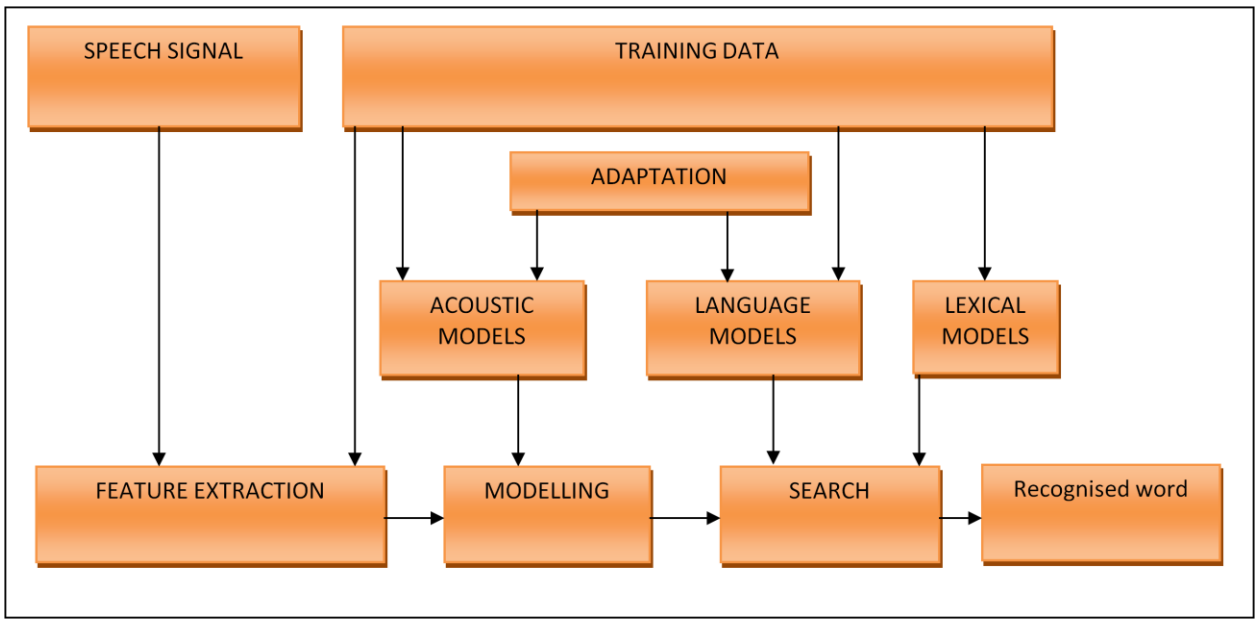

Figure 1: Proposed Architecture

A. Speech Signal acquisition

Speech signal acquisition is the initial step in speech processing, laying the foundation for applications such as speech recognition, speaker identification, and voice-controlled systems. This process involves capturing, digitizing, and pre-processing the human voice to prepare it for analysis and interpretation by computational systems.

B. Feature Extraction

Feature extraction is a critical step in speech recognition systems, where raw speech signals are transformed into a compact, informative representation suitable for processing by machine learning or deep learning models. The goal is to capture relevant linguistic information while minimizing noise, redundancy, and speaker-specific characteristics. Following are the common used features.

i) Time-Domain Features

ii) Frequency-Domain Features

iii) Cepstral Features

iv) Dynamic Features

The process involves the following steps:

i) Pre-processing: Includes noise reduction, silence removal, and pre-emphasis.

ii) Framing and Windowing: Dividing the signal into overlapping frames (e.g., 20 ms with a 10 ms overlap) to ensure stationarity. A window function (e.g., Hamming) is applied to reduce spectral leakage.

iii) Transformation: Applying methods like Fourier Transform or LPC to extract relevant frequency-domain information.

iv) Feature Computation: Generating feature vectors (e.g., MFCCs or LPCCs) for each frame.

v) Normalization: Adjusting feature values to reduce variability due to speakers or environmental conditions.

C. Acoustic Modelling

Acoustic modeling is a core component of speech recognition systems, bridging the gap between raw audio signals and linguistic units like phonemes or words. It involves creating statistical representations of the relationship between speech sounds (acoustic features) and their corresponding linguistic labels. This process is essential for accurately converting spoken language into text.

The purpose of acoustic modeling is to provide a probabilistic framework for mapping acoustic signals to phonetic units. It is typically integrated with language models and pronunciation dictionaries to build a complete speech recognition pipeline. In this purpose we used Self-Supervised Learning model. Models like Wav2Vec 2.0 pre-train on large amounts of unlabeled audio data, improving performance with limited labeled datasets.

D. Language & Lexical Modelling

Language and lexical modeling are critical components of speech recognition systems, providing the framework for interpreting and transcribing spoken language into meaningful text. These models complement acoustic models by focusing on the structure and semantics of language, ensuring the output is both accurate and contextually appropriate.

Language models (LMs) predict the likelihood of a sequence of words, ensuring the transcription aligns with linguistic patterns and real-world usage. They are essential for handling ambiguities in

speech and improving the overall accuracy of speech recognition systems. We use Neural Language Models in this work.

Lexical modeling deals with the vocabulary or dictionary used in speech recognition. It links the phonetic representation of words to their textual counterparts, ensuring accurate transcription.

E. *Recognition*

Recognition is the final and critical stage of speech recognition systems, where the processed and analyzed speech signal is converted into a meaningful sequence of words or text. This stage integrates acoustic, language, and lexical modeling outputs to identify the most probable word sequence corresponding to the input speech. Decoding is the process of mapping acoustic signals to textual outputs. It integrates probabilities from acoustic and language models to generate the most likely word sequence. We use Viterbi Algorithm for recognition. It used in Hidden Markov Models (HMMs) to find the optimal state sequence.

IV. EXPERIMENTAL SETUP AND RESULTS

The We will explore the relationship between isolated word recognition rates and the analysis window length and shift size. For this study, we used Mice for voice input. The utterances were recorded with a sampling rate of 8,000 Hz, mono channel, and 16-bit quantization.

A Dynamic Time Warping (DTW)-based recognizer was employed, a pattern comparison approach that effectively models isolated word recognition. To extract features, three different cepstral analysis techniques were used: mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC) and proposed perceptual linear predictive cepstral coefficients (PPLPC). This allowed us to apply the same Euclidean distance calculation method across all analysis types. The recognizer was implemented in .net.

The first experiment aimed to identify the analysis window length that would yield the highest recognition rate. To do this, we varied the window length from 10 to 30 ms, with increments of 2.5 ms, across all analysis types. The results of the experiment are shown in Figure 2. The highest average recognition rate was achieved using a 10 ms window length for all analysis types. This length is half the duration of the commonly used 20–25 ms window. For the PPLPC analysis, the recognition rate using the 10 ms window was 1.2% higher than that obtained with the 20 ms window.

When analyzing recognition rates for female and male speakers separately, we found that the optimal window length for female speakers was slightly shorter, ranging from 10 ms to 12.5 ms across all analysis types. For male speakers, the optimal window length was more variable, ranging from 10 ms to 17.5 ms.

Increasing the window length generally reduced the recognition rate. However, extending the window length to 15 ms in the case of PPLPC analysis and 20 ms for LPCC and MFCC analyses led to a decrease in recognition rate of less than 1%. This reduction is minimal and could be considered acceptable if needed for specific applications.

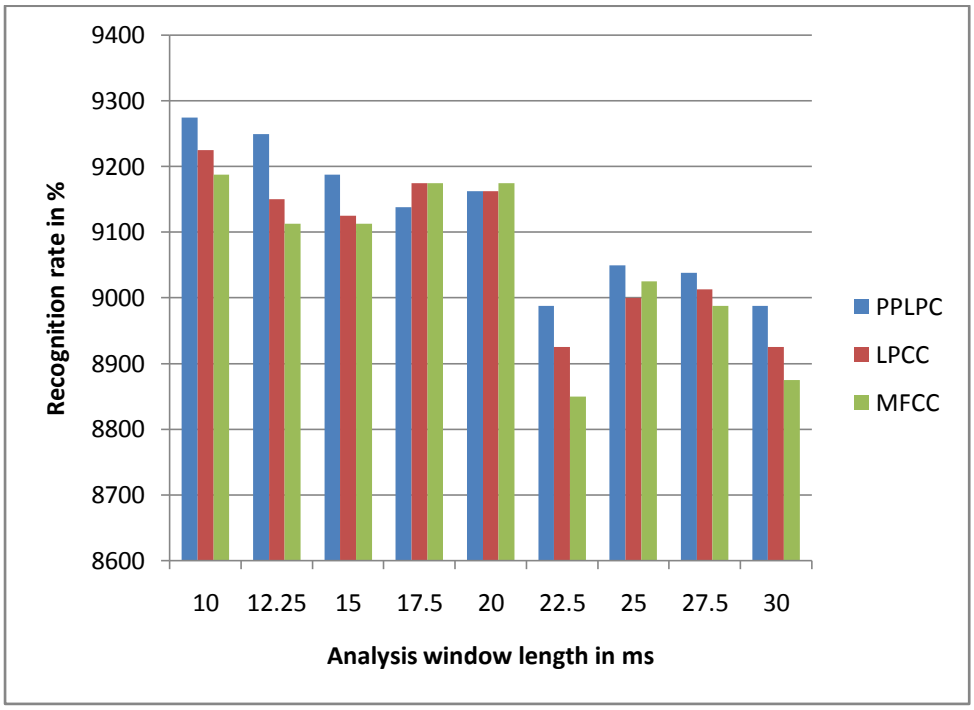

Figure 2: Recognition rate dependence on analysis window length

To further investigate the impact of frame shift, we examined the recognition rate using the maximum acceptable frame size values: 15 ms for PLP analysis and 20 ms for LPCC and MFCC analyses. Table 1 presents the combined results of frame shift values that yield the highest recognition rate for different frame sizes (expressed as a ratio to the frame size). The results indicate that the highest recognition rate is achieved when the frame shift size is between 75% and 100% of the frame size. If computational speed is a priority, setting the frame shift to 100% of the frame size is recommended.

Table 1: Frame shift values giving highest recognition rate

Analysis type	Minimal frame size		Maximal frame size	
	Frame size in ms	Frame shift in %	Frame size in ms	Frame shift in %
PPLPC	10	100	15	83
LPCC	10	75	20	100
MFCC	10	75	20	100

Figure 3 illustrates the overall improvement in recognition rate across all analysis techniques, resulting from changes in window length and frame shift size. The results show that PLPC analysis achieved the highest improvement, with a 2.5% increase in recognition rate, compared to the case of a 20 ms window length and 10 ms frame shift. In contrast, MFCC analysis was the most resilient to changes in analysis parameters, with a modest recognition rate improvement of just 0.3%.

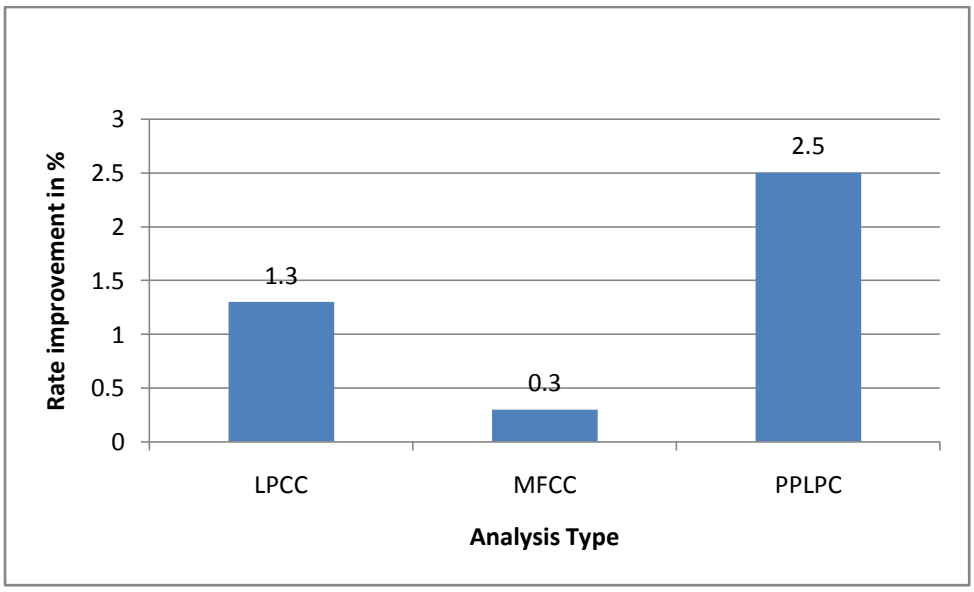

Figure 3: Speech recognition improvement.

In our experiments we registered different recognition rates for different analysis types and speakers, that is the increase/decrease of frame size and shift had diverse results for particular

speaker. This implies the speaker-dependent optimal values of frame size and shift. Thus future research ought to be directed to adapt analysis parameters to speaker characteristics.

V. CONCLUSION

In conclusion, this study highlights the importance of analysis window length and frame shift in optimizing speech recognition performance. The inherent redundancy and non-stationary nature of speech signals are addressed by treating them as stationary over short segments, with key information captured in the short-time magnitude spectrum. The investigation of three cepstral analysis techniques—MFCC, LPCC and PPLPC—reveals that a 10 ms analysis window and a frame shift between 7.5 and 10 ms offer the highest recognition accuracy across in PPLC, with a notable 2.5% improvement in recognition rate.

These results underscore the significance of parameter selection in cepstral analysis, demonstrating its direct impact on speech recognition system performance. By carefully adjusting analysis parameters, researchers and practitioners can significantly enhance the accuracy and efficiency of speech recognition models. The findings offer valuable insights for advancing speech recognition technology, highlighting the potential for further refinement and optimization in the development of more robust and accurate systems.

REFERENCES

1. Z. Jiang, H. Huang, S. Yang, S. Lu, and Z. Hao, “Acoustic Feature Comparison of MFCC and CZT-Based Cepstrum for Speech Recognition,” in Proceedings of 5th International Conference on Natural Computation, 2009, pp. 55–59.
2. L. Deng, J. Wu, J. Droppo, and A. Acero, “Analysis and comparison of two speech feature extraction/compensation algorithms,” IEEE Signal Processing Letters, vol. 12, no. 6, pp. 477–480, Jun. 2005.
3. S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, Aug. 1980.
4. J. Pelecanos, S. Slomka, and S. Sridharan, “Enhancing automatic speaker identification using phoneme clustering and frame based parameter and frame size selection,” in Proceedings of the 5th International Symposium on Signal Processing and its Applications ISSPA99 (IEEE Cat. No.99EX359), vol. 2, pp. 633–636.
5. K. Paliwal and K. Wojcicki, “Effect of Analysis Window Duration on Speech Intelligibility,” IEEE Signal Processing Letters, vol. 15, pp. 785–788, 2008.
6. Padi, Sarala, Seyed Omid Sadjadi, Ram D. Sriram, and Dinesh Manocha. "Improved speech emotion recognition using transfer learning and spectrogram augmentation." In Proceedings of the 2021 international conference on multimodal interaction, pp. 645-652. 2021.
7. Rolland, Thomas, Alberto Abad, Catia Cucchiarini, and Helmer Strik. "Multilingual transfer learning for children automatic speech recognition." (2022).

8. Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. "Multitask learning from augmented auxiliary data for improving speech emotion recognition." IEEE Transactions on Affective Computing 14, no. 4 (2022): 3164-3176.
9. Sullivan, Peter, Toshiko Shibano, and Muhammad Abdul-Mageed. "Improving automatic speech recognition for non-native English with transfer learning and language model decoding." In Analysis and Application of Natural Language and Speech Processing, pp. 21-44. Cham: Springer International Publishing, 2022.
10. Padi, Sarala, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D. Sriram. "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models." arXiv preprint arXiv:2202.08974 (2022).
11. Yadav, Hemant, and Sunayana Sitaram. "A survey of multilingual models for automatic speech recognition." arXiv preprint arXiv:2202.12576 (2022).
12. Kheddar, Hamza, Yassine Himeur, Somaya Al-Maadeed, Abbes Amira, and Faycal Bensaali. "Deep transfer learning for automatic speech recognition: Towards better generalization." Knowledge-Based Systems 277 (2023): 110851.
13. Steinmetz, Hillel. "Transfer Learning Using L2 Speech to Improve Automatic Speech Recognition of Dysarthric Speech." Master's thesis, University of Washington, 2023.
14. Nga, Cao Hong, Duc-Quang Vu, Huong Hoang Luong, Chien-Lin Huang, and Jia-Ching Wang. "Cyclic Transfer Learning for Mandarin-English Code-Switching Speech Recognition." IEEE Signal Processing Letters (2023).
15. Tun, Su Shwe Yi, Shogo Okada, Hung-Hsuan Huang, and Chee Wee Leong. "Multimodal Transfer Learning for Oral Presentation Assessment." IEEE Access (2023).
16. Zheng, Ruonan, and Rui Zhang. "Classification of intelligent speech system and education method based on improved multi label transfer learning model. "International Journal of System Assurance Engineering and Management (2023): 1-10.
17. Zhou, Yue, Yuxuan Yuan, and Xiaodong Shi. "A multitask co-training framework for improving speech translation by leveraging speech recognition and machine translation tasks." Neural Computing and Applications 36, no. 15 (2024): 8641-8656.
18. Ta, Bao Thang, and Nhat Minh Le. "Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language." Engineering Applications of Artificial Intelligence 132 (2024): 107895.
19. Kheddar, Hamza, Mustapha Hemis, and Yassine Himeur. "Automatic speech recognition using advanced deep learning approaches: A survey." Information Fusion (2024):102422.
20. Hassan, Najmul, Abu Saleh Musa Miah, and Jungpil Shin. "A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition." Applied Sciences 14, no. 2 (2024): 603.
21. Kumar, Aditya, and Jainath Yadav. "Multiview Learning-Based Speech Recognition for Low-Resource Languages." Automatic Speech Recognition and Translation for Low Resource Languages (2024): 375-403.