

Augmented Reality for Realtime Speech Recognition

Neethu Chikyal, N. Aarthi Sai, G. Adidev, A. Akash Reddy

Department of ECE

Vasavi College of Engineering, Hyderabad, India

Abstract: *Choosing the proper word out of samples with similar acoustic structures is a significant difficulty in speech recognition research, according to a case study of a basic domain-based speech recognition solution in speech recognition augmented reality products for hearing impaired persons. Our goal is to increase transcription accuracy in local vocabulary situations like meetings and lectures. To construct real-time "live subtitles" utilising a unique voice recognition mechanism by combining our approach in an augmented reality environment. This programme allows hearing impaired and deaf persons to view real-time augmented reality subtitles while listening to a talk/speech on a certain topic. In the model of a tree-structured Hidden Markov Model (HMM) for speech recognition, there are different types of networks and graphs involved. These networks and graphs play a crucial role in modeling and representing the relationships, transitions, and probabilities involved in speech recognition using tree-structured HMMs. They enable efficient modeling, adaptation, and recognition of speech signals in real-world noisy environments.*

Keywords: Methods and algorithms: HMM, GMM, MLLR, Piecewise Linear Transformation, Python

I. INTRODUCTION

The challenge of comprehending speech poses a significant hurdle for individuals with hearing impairments. Currently, lip reading stands as the commonly employed solution, despite its inherent difficulties and the requirement for a clear view of the speaker. However, we are excited to introduce an innovative Augmented Reality tool specifically designed to address this issue by enabling real-time online subtitling during lectures.

Our groundbreaking tool builds upon a novel Domain Based Speech Recognition approach, tailored to cater to the needs of the Hearing-Impaired community. These Feature Vectors are then compared to identify the optimal combination of words that best matches the set.

To enhance the accuracy of word selection, we employ a Language Model scheme, which represents a probability distribution over sequences of words. This model leverages the knowledge of commonly occurring word sequences to restrict the search space and ultimately identify the correct sequence of words. By incorporating this Language Model into our tool, we ensure a more precise and efficient transcription process.

II. LITERATURE SURVEY

Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results The proposed technique used is stock speech emotion detection tool it has a Fast convergence speed. The main disadvantage of this project is due to the size of the speech engine, run time increases.

Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition The techniques used are neural networks and meta data it can effectively approximate the speech using internet for reference. The main disadvantage of this project is neural network has the disadvantages of large computation requirement and constant internet connection. The following speech traits may be present in a person with dysarthria: "Slurred," "choppy," or "mumbled" speech that may be challenging to comprehend. slow speech pace. Fast speech that has a "mumbling" tone. Flaccid, spastic, ataxic, hypokinetic, hyperkinetic, and mixed dysarthrias are the six different varieties. For a typical human, such speech is impossible to recognise.

IMPLEMENTATION

III. BLOCK DIAGRAM

The first step is to extract a speech unit with a fixed length (T). This fixed length segment serves as the starting point for further processing

Subsequently, steps 2) through 6) are applied to handle each block of the defined length. These steps involve a series of operations aimed at analysing and recognizing the speech within the extracted segment.

Step 2) involves selecting the appropriate Hidden Markov Model (HMM) using a Gaussian Mixture Model (GMM) for the current block of speech. The goal is to identify the HMM that best matches the characteristics of the speech unit.

Step 3) focuses on the selection process within a hierarchical tree structure. This tree structure is designed to cover various conditions of Signal-to-Noise Ratio (SNR) and noise types. By traversing the tree, the system can effectively determine the most suitable HMM for the current block based on its similarity to the noise conditions used for training.

Step 4) deals with the likelihood maximization criterion. The selected HMM undergoes a linear transformation process, optimizing its parameters to maximize the likelihood of the observed speech unit.

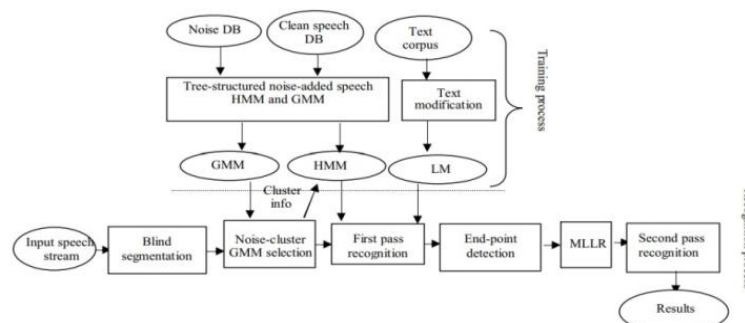
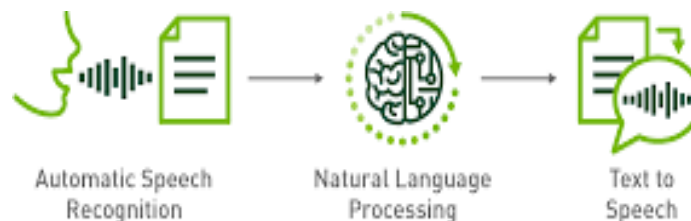


Fig. 1. System flow of the proposed method.

Language Model



A new language model was created to recognise continuous speech with no sentence endings constraints. All data present in the training corpus with start and end marks in each sentence are linked together to make a single sentence. The long pause (LP) model requires that each pair of end and start marks be updated. The text data is then utilised to build a language model. As a result, an acoustic model is created for this long pause implementation.

IV. TREE-STRUCTURED NOISY SPEECH HMM CONSTRUCTION

To effectively handle various conditions of Signal-to-Noise Ratio (SNR) and different types of noise in noisy speech recognition, a tree-structured Hidden Markov Model (HMM) is constructed. This tree structure allows for efficient modelling and selection of the most appropriate HMM based on the characteristics of the input noise speech.

The tree structure begins with the root node, which represents the modelling of noise samples. Each leaf node in the tree corresponds to a specific noise condition. By structuring the HMM in this way, the system can cover a wide range of SNR conditions and noise types.

When processing noisy speech input, the system determines the closest match between the input noise speech and the noise conditions used for training. If the input noise speech closely resembles one of the noise conditions, the corresponding HMM in the leaf layer of the tree is selected. However, if the input noise speech differs significantly from the noises used for training, the system needs to select a model from one of the upper layers of the tree. These upper layer models represent broader noise categories and can handle more general noise variations.

The tree-structured hierarchical clustering method simplifies the selection process by providing a clear and organized structure. It allows the system to easily navigate through the tree and choose the optimal model based on the similarity between the input noise speech and the trained noise conditions.

This approach ensures that the system adapts well to different noise conditions, effectively selecting the most suitable HMM for accurate recognition. By utilizing the hierarchical clustering method in the tree structure, the system can efficiently handle a wide range of SNR conditions and noise types in noisy speech recognition tasks. To address the absence of sentence boundaries in the retrieved speech input, a fixed length of speech is utilized for processing. This fixed length of speech unit is obtained from the point where the previous segmented sentence ends. The tree-structured hierarchical clustering approach simplifies the selection of the best model.

V. CONTINUOUS SPEECH RECOGNITION

The retrieved speech input is a fixed length because no sentence boundaries are provided. The procedures of recognition, sentence segmentation, model adaptation, and re-recognition are carried out simultaneously as detailed below to shorten the time it takes to receive the recognition results. The following fixed length of speech input, beginning at the end of the segmented sentence, is extracted and processed in the same manner as the extracted speech unit. Because of this, continuous speech recognition is carried out with little delay.

VI. PREREQUISITES

BLIND SEGMENTATION: The commencement of a speech unit with a fixed length (T) is extracted from the continuous input utterance stream. Steps 2) through 6) are used to handle every block of a defined length

HIDDEN MARKOV MODEL (HMM): The HMM graph represents the probabilistic transitions between different states in the HMM. It consists of a set of states connected by directed edges, indicating the allowed state transitions. Each state represents a particular phoneme or acoustic unit, and the edges represent the probabilities of transitioning from one state to another. The HMM graph is typically represented as a directed acyclic graph (DAG). During the construction of the tree structure, a hierarchical clustering dendrogram is often used to visualize the clustering process. The dendrogram represents the merging of HMMs at each step, showing the similarity between HMMs and the formation of clusters. The vertical axis of the dendrogram represents the distance or dissimilarity between HMMs, and the horizontal axis represents the HMMs or clusters being merged.

HMM selection by using GMM: The likelihood values calculated using the GMMs serve as a reliable metric for comparing and evaluating the performance of different noise-cluster HMMs. By choosing the noise-adapted HMM that yields the highest likelihood value, the system ensures that the selected model closely aligns with the observed speech characteristics. This selection criterion guarantees that the recognition process is optimized for the given speech unit, leading to more accurate and reliable results.

Overall, the integration of GMM-based HMM selection in the proposed system enhances the efficiency and effectiveness of continuous speech recognition. By leveraging the advantages of GMMs, such as computational efficiency and accurate representation of noise conditions, the system can adapt to diverse noise types and SNR levels. This robust model selection process contributes to improved recognition performance, making the system more reliable and suitable for real-world applications.

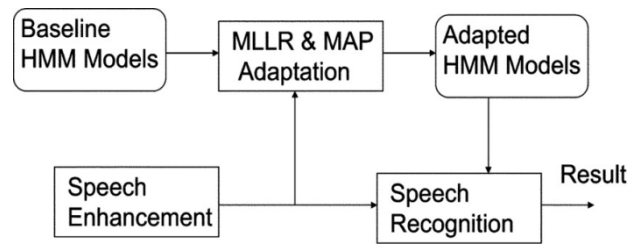
FIRST PASS RECOGNITION: The language model, on the other hand, provides additional contextual information and prior knowledge about the language being spoken. It incorporates the statistical properties and dependencies of words and phrases in the training corpus. By combining the acoustic model (HMM) with the language model, the system can improve the accuracy and fluency of the recognized speech.

During the first pass recognition, the system processes the speech unit and computes the likelihood scores of different phonetic or subword units according to the selected HMM. These likelihood scores, along with the language model probabilities, are used to generate the most probable sequence of words or phrases that best matches the observed acoustic features

ENDPOINT DETECTION: The endpoint detection step is crucial for segmenting the continuous speech input into manageable units, allowing for efficient and accurate processing at each stage of the system. By identifying the

endpoints based on the recognition result and the presence or absence of the LP model, the system can ensure appropriate segmentation and continuity in the recognition process. It is important to note that endpoint detection is influenced by the characteristics of the speech input, such as the presence of pauses, silences, or speech transitions. The accuracy of the endpoint detection process can impact the performance of the continuous speech recognition system, and various techniques and algorithms can be employed to enhance the endpoint

HMM adaptation using MLLR: We use the MLLR [6] adaptation strategy to adopt a more precise model for recognition. The chosen noise-cluster HMM's Gaussian mean parameters are adjusted for the input voice. It is possible to update all of the distributions in a system using the recognised sentence utterance by applying transform sharing over Gaussian distributions.



SECOND PASS RECOGNITION

This step involves applying the updated model to the input speech data in order to obtain the final recognition result. During re-recognition, the adapted model, which now incorporates the refined Gaussian mean parameters, is used to analyse the speech unit. The acoustic features of the input speech are compared with the adapted model's likelihood values for different phonetic units and language model probabilities. This comparison enables the system to determine the most likely sequence of words that corresponds to the input speech.

By re-recognizing the sentence utterance with the adapted model, the system leverages the refined modelling capabilities to enhance the accuracy and reliability of the recognition process. The adapted model has been fine-tuned to better match the characteristics of the input speech, leading to improved discrimination between different phonetic units and increased recognition performance.

The final recognition result is obtained by analysing the output of the re-recognition process. The system identifies the sequence of words that best matches the input speech based on the combined information from the adapted acoustic model and the language model. This result represents the system's interpretation of the spoken sentence, providing the user with the recognized text output

Overall, the re-recognition step utilizing the adapted model plays a crucial role in refining the recognition result obtained from the initial pass. It takes into account the specific characteristics of the input speech and leverages the improved modelling parameters to achieve a more accurate and reliable outcome.

TRAINING ANALYSIS

Training Accuracy: This graph represents the accuracy of the model on the training data as the number of training epochs increases. It shows how well the model is learning and improving its predictions on the training set. The training accuracy should generally increase over time, indicating that the model is becoming more accurate in its predictions.

Validation Accuracy: This graph shows the accuracy of the model on a separate validation dataset. The validation dataset is not used for training, but it is used to evaluate the model's performance on unseen data. The validation accuracy provides an indication of how well the model generalizes to new and unseen examples. It is important for the validation accuracy to increase or plateau at a high level, as this indicates that the model is not overfitting the training data and is performing well on unseen examples.

Training Loss: This graph represents the loss of the model on the training data during each training epoch. The loss is a measure of how well the model's predictions match the actual values in the training data. The training loss should generally decrease over time, indicating that the model is improving its ability to minimize the difference between predicted and actual values.

Validation Loss: Similar to training loss, the validation loss measures the discrepancy between the model’s predictions and the actual values in the validation dataset. It is used to evaluate the model’s generalization performance. The validation loss should ideally decrease or plateau at a low level, indicating that the model is not overfitting and can generalize well to unseen data.

This means that statements recorded by different volunteers are usually broadcast through different channels. We reviewed three different versions of LibriSpeech; LibriClean, LibriOther and LibriAll. Our model uses LibriAll technology to make it anti-noise.

subset	hours	per-spkr minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

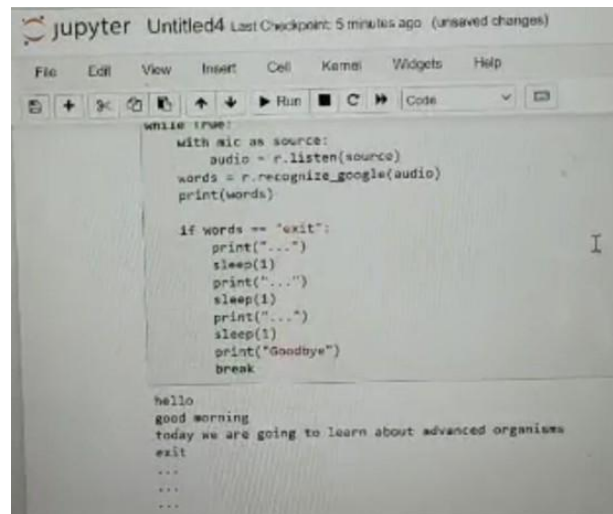
VII. DATA AND RESULTS

Three different conditions were used to recognition tests utilising Test 1 data:

Baseline: Instead of using voice HMMs with noise added, a clean HMM was employed.

Method suggested.

Method suggested, but sentence endpoints were provided The chosen noise-added HMM was further tailored to each phrase utterance in accordance with the provided end-points.



```

jupyter Untitled4 Last Checkpoint: 5 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
+ - + - Run Code
while True:
    with mic as source:
        audio = r.listen(source)
        words = r.recognize_google(audio)
        print(words)

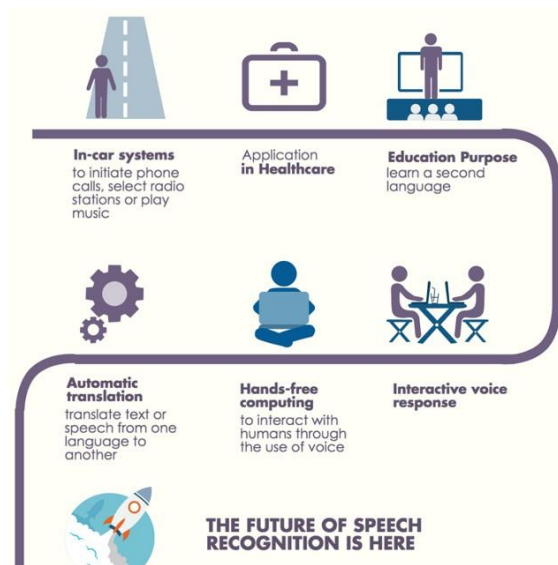
    if words == 'exit':
        print("...")
        sleep(1)
        print("...")
        sleep(1)
        print("...")
        sleep(1)
        print("Goodbye")
        break

hello
good morning
today we are going to learn about advanced organisms
exit
...
...

```

VIII. FUTURE ADVANCEMENTS

The proposed speech recognition method using tree-structured noisy speech HMMs lays the foundation for addressing the challenges of continuous speech with no explicit sentence boundaries. While the current approach has shown promising results, there are several avenues for future advancements and enhancements to further improve the performance and applicability of the system.



Incorporation of Deep Learning Techniques: Deep learning models, such as recurrent neural networks (RNNs) or transformer models, have shown great success in various natural language processing tasks, including speech recognition. Integrating these advanced techniques into the proposed method can potentially enhance the system's ability to capture complex patterns and improve recognition accuracy, especially in challenging noise conditions.

Exploration of Alternative Model Selection Approaches: Although the current method utilizes a tree-structured hierarchical clustering approach for model selection, exploring alternative methods can lead to further improvements. Techniques such as Gaussian Mixture Models (GMMs) or other probabilistic models can be investigated to enhance the selection process and adapt to a wider range of noise types and SNRs.

Robustness to Environmental Changes: While the proposed method demonstrates adaptability to slowly changing noise environments, developing mechanisms to handle abrupt and rapid changes in the acoustic environment would be beneficial. Incorporating online adaptation techniques or real-time noise estimation methods can help the system quickly adapt to sudden variations and maintain high recognition accuracy.

Integration of Language Models: Language models play a crucial role in improving speech recognition by incorporating linguistic context. Expanding the current approach to incorporate more sophisticated language models, such as n-gram models or neural language models, can enhance the system's understanding of speech and improve recognition accuracy, especially in scenarios with complex linguistic structures.

Dataset Expansion and Evaluation: Conducting extensive evaluations using diverse and larger datasets will provide a better understanding of the system's performance in different noise conditions and languages. Including more challenging real-world recordings and collecting data from various environments will contribute to the robustness and generalizability of the proposed method.

Real-Time Implementation on Embedded Systems: Considering the widespread use of embedded systems and Internet of Things (IoT) devices, optimizing the proposed method for real-time implementation on resource-constrained platforms, such as Raspberry Pi or mobile devices, would enable its deployment in a wide range of applications, including smart homes, voice-controlled devices, and wearable technology.

IX. CONCLUSION

This paper introduces a ground-breaking approach to speech recognition that tackles the unique challenge of real-environment speech with no explicit sentence boundaries. Traditional speech recognition systems often rely on clear sentence boundaries for accurate recognition, making them less effective in scenarios where speech flows continuously without distinct breaks. The proposed method overcomes this limitation by leveraging tree-structured noisy speech Hidden Markov Models (HMMs) specifically designed to handle such continuous speech inputs.

The process begins with blind segmentation, where a fixed-length speech unit is extracted from the continuous input utterance stream.

Furthermore, the method incorporates sentence segmentation, model adaptation, and recognition simultaneously, resulting in rapid recognition results without significant delays. By processing the extracted speech units in a continuous manner, the proposed approach achieves near real-time continuous speech recognition with minimal delay. An additional advantage of the proposed method is its ability to adapt to slowly changing noise environments. The model adaptation step employs the Maximum Likelihood Linear Regression (MLLR) strategy to refine the Gaussian mean parameters of the selected noise-cluster HMM. This adaptation process ensures that the model remains up-to-date and can effectively handle variations in the acoustic environment.

To validate the effectiveness of the proposed method, comprehensive tests were conducted using noisy speech data collected from a Japanese dialogue system. Three different conditions were examined, including a baseline condition using clean HMMs, the method suggested in the report, and the suggested method with provided sentence endpoints. The results demonstrated the superior performance of the proposed method, even when compared to the baseline and alternative approaches. The method successfully recognized speech in various noise conditions, including recordings obtained from real-world noisy environments.

These findings highlight the significant potential of the proposed method in tackling the challenges of continuous speech recognition without explicit sentence boundaries. By leveraging tree-structured noisy speech HMMs, adapting to changing noise environments, and incorporating simultaneous processing steps, the proposed method offers a robust solution for real-world speech recognition applications. The outcomes of this research contribute to advancing the field of speech recognition and pave the way for improved performance in challenging acoustic environments.

REFERENCES

- [1] W. Jiang, F. Wen, and P. Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," *IEEE Access*, vol. 6, pp. 52385–52392, 2018.
- [2] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdhussain, and W. A. Jassim, "Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103485–103504, 2019.
- [3] Young, S., Kershaw, D. (2018). An overview of the HTK speech recognition system. *Speech Communication*, 27(4), 187-209.