Speech to Text based Word and Phoneme Level

Transforming Speech Diagnosis: The Power of ASR Technologies

Timothy J Scholtz

Introduction

In this research project, our focus is on leveraging Automatic Speech Recognition (ASR) technologies for diagnosing speech disorders, language impairments, and assisting second language speakers. Our primary goal is to develop a non-invasive diagnostic system. We investigated the capabilities and limitations of existing ASR technologies while exploring their potential implementation for diagnostic purposes.

Context and Significance

Speech disorders, language impairments, and language learning create substantial communication challenges for millions worldwide. In the US alone, approximately 18.5 million individuals are affected. Moreover, the growing number of second language English speakers(approximately 1080 million) further highlights the need for effective support.

Our research project focuses on developing a non-invasive system for automatically scanning and analyzing speech for diagnostic purposes. We aim to enhance speech diagnostics, intervention outcomes, and provide valuable assistance to second language speakers in their language acquisition journey.

Potential Uses

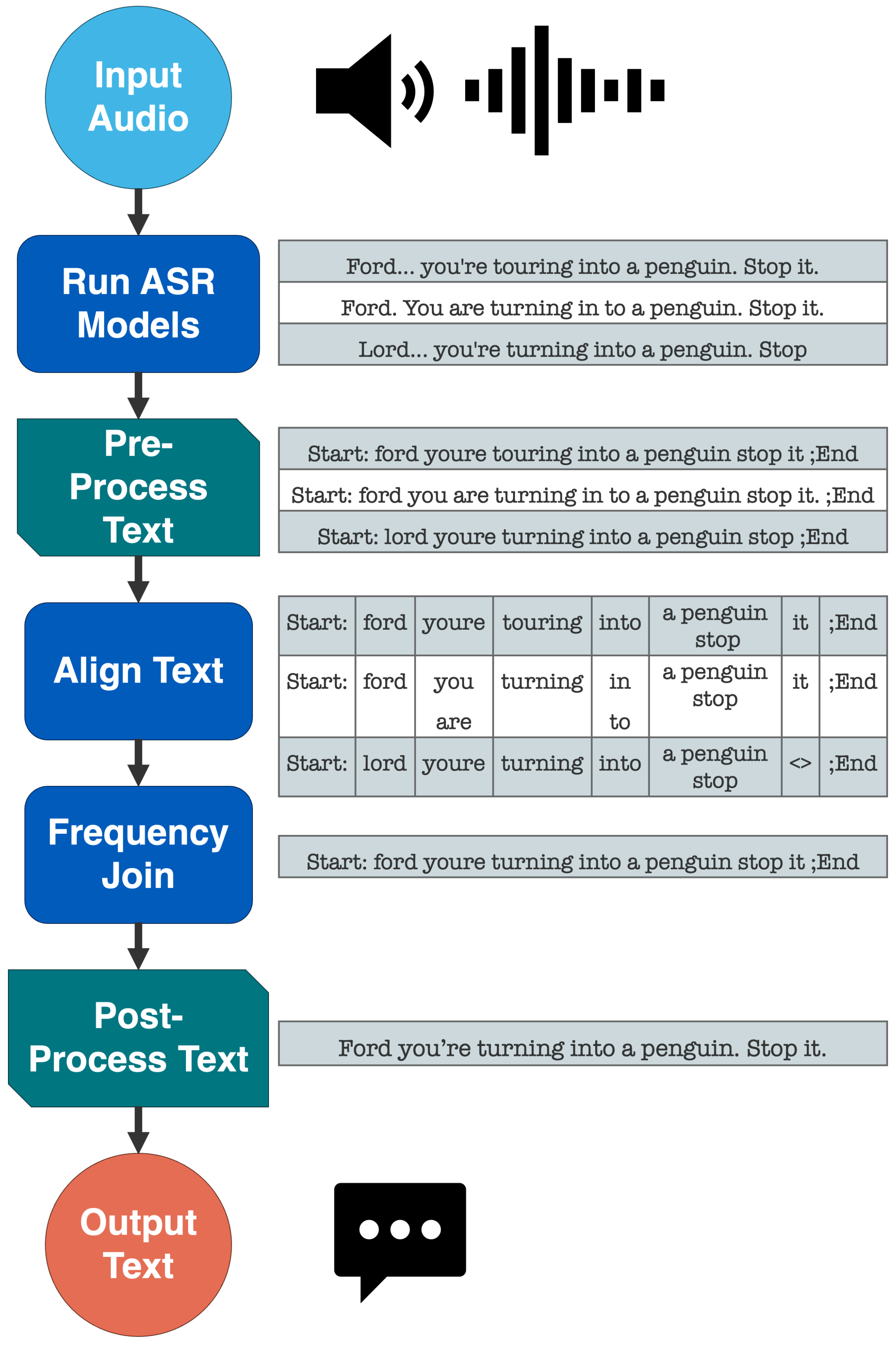
By converting audio into text, we can analyze spoken words, studying grammar, word choice, pauses, word recall, and more. A non-invasive system enables continuous monitoring, facilitating the tracking of speech development and detection of long-term changes. Applications range from general classroom diagnoses, aiding second language speakers, and detecting dementia in old age homes.

Speech To Text

After evaluating various open-source ASR models, we chose Whisper, Vosk, and ESPnet as they demonstrated superior accuracy and relative ease of implementation compared to others. Furthermore, our system also combines the outputs of these models into a unified transcript, as depicted in Diagram1.

Whisper, Vosk, and ESPnet are all End-to-End models, utilizing a process of segmenting input audio into smaller chunks, converting it into MFCC, and passing it through an encoder-decoder system

Combing Process (Diagram 1)



Ford... you're touring into a penguin. Stop it.

Ford. You are turning in to a penguin. Stop it.

Lord... you're turning into a penguin. Stop

Start: ford youre touring into a penguin stop it ;End

Start: ford you are turning in to a penguin stop it. ;End

Start: lord youre turning into a penguin stop ;End

Start:	ford	youre	touring	into	a penguin stop	it	;End
Start:	ford	you are	turning	in to	a penguin stop	it	;End
Start:	lord	youre	turning	into	a penguin stop	<>	;End

Start: ford youre turning into a penguin stop it ;End

Ford you’re turning into a penguin. Stop it.

ASR Model Testing

We tested multiple ASR models on three diverse datasets: DARPA TIMIT, PodcastFillers Dataset, and L2-ARCTIC. DARPA TIMIT served as the control dataset, while the others evaluated edge cases for the ASR models. We tested by running an audio file through the ASR models and compared the text output to the correct transcription of the audio.

WER(Word Error Rate) – the percentage of incorrect words to correct words = (S+D+I)/(N)

Δ N – Size of Correct Text	Δ S – Substitutions
Δ D – Deletions	Δ I – Insertions

About the datasets:

- DARPA TIMIT [Table 1] Comprises 707 readings from male and female speakers representing 8 dialect regions in the United States.
- PodcastFillers [Table 2] Contains transcripts from various podcasts, including non-speech sounds. We tested 104 clips. *Note: Transcriptions for this dataset were a mix of human and ASR, potentially affecting accuracy.
- L2-ARCTIC [Table 3] Non-native speakers with strong accents reading sentences. We tested 7 speakers, each with 512 recordings.

Table 1 - TIMIT

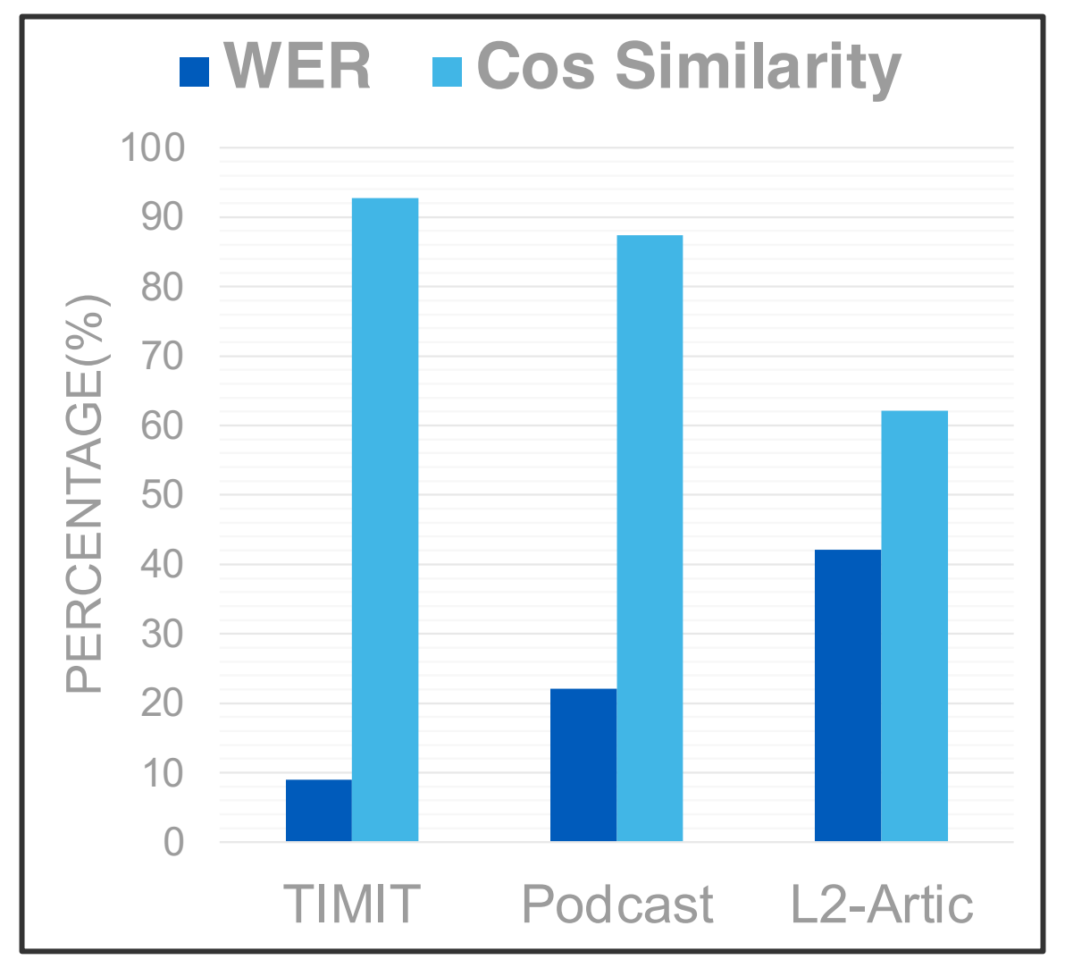
	WER(%)	Cosine Similarity (%)
Whisper	7.59	93.89
Vosk	11.16	90.37
ESPnet	10.57	91.62
Combined	6.48	94.96
Average	8.95	92.71

Table 2 - PodcastFillers

	WER(%)	Cosine Similarity (%)
Whisper	16.12	91.36
Vosk	23.84	86.54
ESPnet	29.30	82.19
Combined	19.15	89.45
Average	22.10	87.39

Table 3 - L2-ARCTIC

	WER(%)	Cosine Similarity (%)
Whisper	17.99	84.98
Vosk	32.46	74.56
ESPnet	99.23	4.19
Combined	18.61	84.63
Average	42.07	62.09



Results

Our testing revealed significant performance variations, indicating ASR models' struggles with non-speech sounds*, accents, and mispronunciations. The combined transcript sometimes outperformed individual ASR models, and potential improvements include adding more ASR models and implementing a dictionary system.

It's essential to consider that WER scores may be skewed due to varying compound nouns and non-speech sounds in the transcript. Implementing a dictionary system to address compound nouns by splitting them apart(raindrops – rain drops) could improve performance but might be challenging to cover every edge case. In real-world implementation, this specific issue becomes less relevant.

The variability in the Podcast dataset and the compound noun issue make it challenging to assess the problem’s full extent

References

- The History of Automatic Speech Recognition – Deepgram – Keith Lam
- Robust Speech Recognition via Large-Scale Weak Supervision – A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever
- Word Error Rate – Wikipedia – Various Authers
- Lemon Grad - English Language Statistics – Anil Yadav

Conclusion

Developing a system for diagnosing speech and language disorders entails several challenges, especially concerning accents, mispronunciations, and uncertainty handling for unfamiliar words. To overcome these obstacles, we propose enhancing ASR models to effectively flag mispronunciations and improve uncertainty handling. Our research concentrated on testing ASR models, showcasing their use in speech analysis. Real-world implementation holds promise, and ongoing research will advance speech diagnostics and language support.