**SpeCor: Speech Recognition and Stutter Correction**

**Dr. Manal Mohsen Tantawi**

Associate professor,

Scientific Computing Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

**TA. Manar Sultan**,

Scientific Computing Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

Abdelrahman Fakhry Hussein   
Scientific Computing  
CS Ain Shams UniversityGiza, Egypt  
20201701097@cis.asu.edu.eg

Mohamed Abdelmonem Abdelrahman   
Scientific Computing  
CS, Ain Shams UniversityCairo, Egypt  
20201700706@cis.asu.edu

Mohammed Mahmoud Hassan   
Scientific Computing  
CS Ain Shams UniversityGiza, Egypt  
20201700733@cis.asu.edu.eg

Tarek Abo Elenen Ali   
Scientific Computing  
CS Ain Shams UniversityCairo, Egypt  
20201700423@cis.asu.edu.eg

Eman Basem Ezzat   
Scientific Computing  
CS Ain Shams UniversityCairo, Egypt  
20201700164@cis.asu.edu.eg

Basma Ahmed Sayed   
Scientific Computing  
CS Ain Shams UniversityCairo, Egypt  
20201700185@cis.asu.edu.eg

# **Abstract**

SpeCor is an innovative Automatic Speech Recognition (ASR) technology designed to improve accuracy for individuals with speech disorders like stuttering and lisping. It integrates specialized correction mechanisms tailored for these disorders, enhancing ASR performance and accessibility. The system employs a minimalist preprocessing approach and emphasizes robust training techniques with noise and data augmentation to mimic real-world conditions. Key to its operation are Mel-Frequency Cepstral Coefficients (MFCCs) for effective feature extraction, focusing on capturing essential phonemic characteristics.

Using the Whisper model and advanced GPT-2 tokenizer, SpeCor processes diverse speech patterns effectively. Evaluation on datasets like LibriSpeech and LibriStutter demonstrates substantial reductions in Word Error Rate (WER) for speech including impediments, improving from 95.5% to 16% WER when trained on inclusive datasets. The architecture combines convolutional layers and transformer blocks for stable and efficient training.

SpeCor not only enhances user experience for those with speech impediments but also advances inclusive voice-based technologies, promising a more reliable and accessible digital future for all users.

# **Introduction**

Voice-based communication technologies have become an integral part of our daily lives, playing crucial roles in applications ranging from virtual assistants and transcription services to customer support and accessibility tools. Despite their widespread adoption, these technologies often fall short for individuals with speech impediments such as stuttering and lisping. These speech patterns can lead to significant inaccuracies in transcription, causing limiting the usability of Automatic Speech Recognition (ASR) systems for these users.

Current ASR systems are primarily designed to handle typical speech patterns, making them inadequate for accurately transcribing speech from individuals with stuttering and lisping. Stuttering is characterized by repetitions, prolongations, and blocks that disrupt the flow of speech, while lisping involves the substitution of sibilant sounds with acoustically similar but distinct sounds. These distortions significantly affect the acoustic features of speech, leading to poor recognition accuracy. Consequently, the effectiveness of ASR technology for individuals with these speech disorders is limited, restricting their ability to benefit from advancements in voice-based communication.

The primary objective of this research is to develop an advanced ASR system named SpeCor, tailored to the needs of individuals with speech impediments, particularly those who stutter or have lisps. SpeCor aims to Develop a robust speech-to-text conversion system that effectively translates spoken input into accurate transcription text., and ensure accurate speech recognition, when the speech includes stuttering or lisping.

By addressing these objectives, SpeCor aims to enhance the ability of individuals with speech impediments to interact with technology and access digital services seamlessly, fostering a more inclusive technological environment

**The remanning of this paper is arranged as follows:**

Section 2: Related Works

Section 3: System Architecture

section 4: Results

section 5: Conclusion

# **Related Work**

This section provides an overview of key studies and previous research in Automatic Speech Recognition (ASR), focusing on the techniques and methodologies employed to address its challenges. It synthesizes foundational literature that has contributed to the evolution of ASR technology, highlighting advancements in feature extraction, pattern recognition, and machine learning models. By reviewing these works, this section sets the stage for understanding the current landscape of ASR research and identifies gaps and opportunities for further exploration in improving accuracy, robustness, and applicability across various domains and languages.

[(Z. Yao et. al.,)](#s1)[1] presents ZIPFORMER, an advanced encoder for speech recognition, demonstrating improved performance over existing models such as E-Branchformer, Conformer, and Squeeze Former. Evaluations on datasets like LibriSpeech, Aishell-1, and WenetSpeech show ZIPFORMER achieving the lowest Word Error Rate (WER) of 4.38 compared to 4.55, 5.55, and 5.97 for E-Branchformer, Conformer, and SqueezeFormer, respectively.

[(D. Prabhu et al.,)](#s2)[2] explores methods like Codebook Attend, DAT, and MTL to improve speech recognition for accents using datasets like Common Voice Corpus and L2-ARCTIC. It reports Word Error Rates (WER) for these methods: Codebook Attend achieved 18.2%, DAT achieved 18.7%, and MTL achieved 18.9%, highlighting their effectiveness in handling accented speech.

[(H. Toyin et. al.,)](#s3)[3] introduces ArTST, a model designed for Arabic text and speech processing, achieving a Word Error Rate (WER) of 17.27% and a Character Error Rate (CER) of 9.99%. It utilizes techniques like Audio Sampling, Mel-Frequency Cepstral Coefficients, and Log Mel-filter banks to enhance accuracy in Arabic speech recognition tasks. It compares ArTST with SpeechT5, highlighting significant performance differences; SpeechT5 achieves a higher WER of 53.19% and a CER of 19.01%. ArTST's success in handling Arabic speech recognition tasks, including removing punctuation marks, underscores its effectiveness for applications requiring precise transcription and understanding of spoken Arabic.

[(by C. Wang et. al.,)](#s4)[4] introduces fairseq S2T, a framework optimized for rapid speech-to-text modeling using the fairseq toolkit. It emphasizes efficiency in converting spoken language into text, supporting various applications requiring real-time or high-throughput speech recognition. fairseq S2T leverages advanced neural network architectures and optimization techniques to achieve competitive performance in speech recognition tasks. The framework is designed to handle large-scale datasets efficiently, making it suitable for both research and practical deployment scenarios.

[(by J. Ao et. al.,)](#s5) [5] SpeechT5 introduces a unified encoder-decoder architecture tailored for spoken language processing, building upon the successful T5 model. It pre-trains on diverse spoken language data, enhancing capabilities in tasks like speech recognition, synthesis, and understanding with promising results across various benchmarks. The paper highlights SpeechT5's effectiveness in leveraging a unified approach to process different modalities of spoken language data. It achieves competitive performance in tasks such as speech-to-text and text-to-speech, demonstrating its potential for advancing spoken language processing applications.

[(by Y. Leng et. al.,)](#s6)[6] introduces SoftCorrect, a novel approach aimed at improving automatic speech recognition (ASR) systems. SoftCorrect utilizes soft detection mechanisms and deep learning techniques to identify and correct errors in ASR outputs with enhanced accuracy and efficiency. It dynamically adjusts confidence thresholds based on input characteristics, effectively mitigating recognition errors compared to traditional methods. The study demonstrates superior performance of SoftCorrect in error correction tasks, highlighting its potential to advance ASR technology by improving reliability and user experience in various real-world applications.

[[(by A. Radford et. al.,)](#s7)[](#s7)7] explores a novel approach to improving speech recognition by leveraging large-scale weak supervision. It introduces a methodology that utilizes extensive amounts of weakly labeled data to train models, aiming to enhance performance and generalization. This strategy enables the model to learn from diverse speech variations and environments, leading to improved accuracy and robustness across different conditions compared to traditional supervised learning methods. The study demonstrates significant advancements in speech recognition capabilities, highlighting the effectiveness and scalability of training models with abundant and varied data sources under weak supervision.

[(by S. Dutta et. al.,)](#s8)[8] by S. Dutta et al. introduces a sequence-to-sequence (Seq2Seq) approach for enhancing Automatic Speech Recognition (ASR) systems. The Seq2Seq models, equipped with attention mechanisms, are specifically designed to identify and correct errors in speech transcripts. By training on large-scale datasets, the models effectively improve the accuracy of ASR outputs by accurately aligning and rectifying errors caused by factors such as noise and accents. The study demonstrates superior performance of the Seq2Seq model compared to traditional methods in error correction tasks, highlighting its potential to enhance the reliability and usability of ASR technology in real-world applications.

[(by L. Meng et. al.,)](#s9)[9] introduces MIXSPEECH, a data augmentation technique designed to enhance Automatic Speech Recognition (ASR) in low-resource settings. MIXSPEECH combines mixup and SpecAugment methods to generate diverse training examples: mixup blends audio samples and labels, while SpecAugment augments spectrograms by masking frequency bands and time segments. This approach enriches the training dataset with varied speech signals, improving ASR model robustness and performance. The study demonstrates significant enhancements in ASR accuracy and generalization compared to traditional methods, addressing challenges posed by limited data availability in low-resource scenarios.

[(by Y. Hu et. al.,)](#s10)[10] explores the application of pre-trained contextual language models, such as BERT, for improving misspelling correction. The study demonstrates how these models, fine-tuned for the task, effectively detect and rectify misspelled words by leveraging their ability to understand contextual semantics. Compared to traditional rule-based methods, the approach achieves higher accuracy and robustness in correcting various types of spelling errors, showcasing the potential of advanced language models in enhancing text processing capabilities, particularly in automated spelling correction systems.

[(by Y. Leng et. al.,)](#s11)[11] introduces FastCorrect, a method designed to improve the efficiency and accuracy of error correction in Automatic Speech Recognition (ASR) systems. FastCorrect utilizes edit alignment techniques to swiftly identify and correct errors between ASR-generated transcripts and their corresponding ground truth. By employing fast algorithms, the approach enhances ASR performance by quickly adjusting transcripts to accurately reflect spoken content, reducing computational overhead and enhancing real-time processing capabilities. This research aims to streamline error correction in ASR, making it more effective and suitable for applications requiring rapid and reliable transcription.

[(by et. O. Basystiuk al.,)](#s12)[12] outlines the creation of a system designed to automatically convert audio recordings into text. It discusses the methodology used to develop the system, focusing on technologies likely related to Automatic Speech Recognition (ASR) and possibly machine learning. The study emphasizes the system's architecture, algorithms for accurate transcription, and its potential applications in various fields requiring efficient handling and analysis of spoken content. Overall, the paper contributes to advancing speech processing technology by detailing the implementation and capabilities of an automated audio-to-text conversion system.

[(by A. Mohamed et. al.,)](#s13)[13] introduces a hybrid model that combines transformer architectures with convolutional neural networks (CNNs) to improve Automatic Speech Recognition (ASR). This approach aims to leverage transformers' ability to capture long-range dependencies and contextual information, while integrating CNNs for efficient local feature extraction. The study focuses on enhancing ASR accuracy and computational efficiency, particularly beneficial for handling diverse speech patterns and large datasets. It evaluates the hybrid model against traditional transformer-based ASR systems, highlighting potential improvements in transcription accuracy and robustness to noise and accents. Overall, the research contributes to advancing ASR technology by proposing a hybrid architecture that effectively integrates the strengths of transformers and CNNs to optimize speech recognition performance.

In the realm of automatic speech recognition (ASR), researchers frequently leverage popular datasets such as LibriSpeech, Common Voice Corpus, Wall Street Journal, and Aishell-1. These datasets provide diverse collections of speech data that facilitate the development and evaluation of ASR models across different languages and domains.

Preprocessing techniques are crucial in ASR research, with common approaches including Data Augmentation and Data Normalization, alongside feature extraction. Mel-Frequency Cepstral Coefficients (MFCC) stands out as the predominant feature extraction technique employed due to its effectiveness in capturing speech characteristics.

Recent advancements in ASR have prominently featured transformers, which have significantly enhanced model performance by enabling better handling of long-range dependencies in speech data.

For evaluating the accuracy and efficacy of ASR systems, researchers commonly use metrics such as Word Error Rate (WER) and Character Error Rate (CER). These metrics provide standardized measures to assess the transcription quality of speech recognition models.

# **System architecture**

## A diagram of a computer program Description automatically generated

Figure 3.1 Model Architecture

The model architecture for speech recognition uses an encoder-decoder Transformer, exemplified by the Whisper model. This architecture handles various tasks like multilingual speech recognition and translation by representing tasks as sequences of tokens. The encoder processes input features using convolutional layers and Transformer blocks with sinusoidal position embeddings. The decoder, with learned position embeddings and cross-attention layers, generates the output sequence. The architecture includes pre-activation residual connections and layer normalization for efficient training and stability and uses a Key-Value (KV) cache for faster inference.

## **Preprocessing**

In the recent shift towards minimalist preprocessing for speech-to-text conversion, the focus is on enhancing the model's robustness by introducing noise during training and using data augmentation techniques. This approach aims to simulate real-world conditions where background noise is common, reducing the need for extensive noise removal and enabling more efficient real-time transcription. By training on noisy data, the model learns to handle diverse acoustic environments, improving its performance in practical scenarios.

## **Feature Extraction**

The feature extraction process in speech recognition involves computing Mel-Frequency Cepstral Coefficients (MFCCs) through several steps. Initially, the audio signal undergoes pre-emphasis to boost high frequencies, followed by windowing to minimize spectral leakage. The Fast Fourier Transform (FFT) converts the signal to the frequency domain, and the Log Mel Filter Bank applies a log scale mimicking human auditory perception. Finally, the Discrete Cosine Transform (DCT) compresses the features, producing MFCCs that capture essential sound characteristics for tasks like speech recognition and speaker identification.

## **Tokenizer for Speech-to-Text Conversion**

Tokenizers convert text into sequences of tokens, aiding model training and inference. The GPT-2 tokenizer, known for its byte pair encoding (BPE) and contextual understanding, breaks down text into sub words or characters. This improves the model's handling of rare and complex words, ensuring efficient and accurate processing. Its advanced tokenization capabilities enhance the robustness and precision of the speech-to-text conversion, making it a preferred choice for capturing diverse vocabulary and linguistic structures.

## **Inference**

Inference in speech-to-text models involves using a trained model to predict text from new audio data. The process starts with capturing and preprocessing the audio, extracting features, and passing them through the model's encoder and decoder to generate text tokens. Decoding can be done using greedy decoding, which is simple and fast but may miss optimal sequences, or beam search, which considers multiple sequences for better accuracy but is slower and more complex. This process enables the model to efficiently convert spoken language into accurate written text.

# **Experimental results**

## **Datasets**

* + 1. **LibriStutter Data**

The LibriStutter Data, derived from the LibriSpeech corpus, aids stuttering research by providing a dataset with artificial stuttering patterns, simulating common stuttering behaviors such as repetitions, prolongations, and blocks. The original LibriSpeech dataset is a large corpus of read English speech from audiobooks, featuring around 1,000 hours of audio, diverse speaker demographics, and accurate transcriptions. LibriStutter aims to help develop ASR systems that can handle speech disfluencies, particularly stuttering. This dataset is valuable for speech recognition research, allowing for the development of models robust to speech disfluencies, and for linguistic studies, enabling the analysis of the impact of stuttering on speech processing.

* + 1. **LibriSpeech Dataset**

The LibriSpeech ASR dataset, sourced from audiobooks, is a high-quality, large-scale corpus of read English speech used for ASR research and benchmarking. Introduced in 2015, LibriSpeech provides substantial, high-quality speech data for ASR model training, sourced from public domain audiobooks. The dataset includes transcriptions that provide the ground truth for ASR models, aligned with audio segments. It is ideal for training and benchmarking ASR models, training models for speaker identification and verification, and analyzing speech patterns and pronunciation variations in linguistic research. However, LibriSpeech represents read speech, which may not generalize to conversational speech, and lacks full accent and dialect diversity.

## **Performance Matric**

## We used Word Error Rate (WER) to evaluate the model

Where:

* **S:** is the number of substitution errors.
* **D:** is the number of deletion errors.
* **I**: is the number of insertion errors.
* **N:** is the total number of words in the reference transcription.

## **Results**

ASR systems were evaluated using LibriSpeech and LibriStutter datasets, with WER results indicating the effectiveness of different training approaches.

### Table 4.1 Results Using WER

|  |  |  |
| --- | --- | --- |
| **Train Test** | **LibriSpeech** | **LibriStutter** |
| **Using LibriSpeech Only** | **8.3** | **95.5** |
| **Using LibriStutter Only** | **445.39** | **212.05** |
| **Using LibriSpeech & LibriStutter** | **9.24** | **16** |

Using the LibriSpeech dataset alone resulted in a WER of 8.3 on its test set, indicating high accuracy in recognizing non-stuttered speech typical of fluent speech conditions. This dataset is well-suited for general ASR tasks focused on fluent speech.

In contrast, training the model solely on the LibriStutter dataset led to a significantly higher WER of 445.39 on the LibriSpeech test set, highlighting poor performance in recognizing both stuttered and non-stuttered speech patterns. This indicates limitations in handling diverse speech conditions effectively when using only stuttered speech data.

Combining both LibriSpeech and LibriStutter datasets during training resulted in a WER of 9.24, demonstrating improved accuracy in recognizing stuttered speech while maintaining reasonable performance for fluent speech. This approach shows promise in enhancing the model's capability to handle a broader range of speech patterns.

Evaluation on the LibriStutter dataset using only the LibriSpeech-trained model showed a WER of 95.5, underscoring a significant accuracy gap between stuttered and non-stuttered speech recognition. This disparity suggests potential challenges and limitations in recognizing speech impediments when relying solely on ASR systems trained on fluent speech data.

Conversely, training exclusively on the LibriStutter dataset resulted in a WER of 212.05, indicating a bias towards recognizing stuttered speech patterns but achieving lower overall accuracy. This reinforces the need for balanced training data encompassing both speech types to improve overall ASR performance.

By integrating both LibriSpeech and LibriStutter datasets, the combined approach achieved a WER of 16, demonstrating favorable outcomes for both speech types and emphasizing the benefits of using comprehensive training data to develop more inclusive and accurate ASR systems.

# **Conclusion**

SpeCor is an innovative solution to the challenges of Automatic Speech Recognition (ASR) technology, particularly in addressing the accurate transcription of speech impediments such as stuttering and lisping. Traditional ASR systems struggle with atypical speech patterns, leading to inaccuracies in transcription for individuals with speech disorders. Stuttering and lisping significantly distort the acoustic features of speech, confusing ASR systems and reducing transcription accuracy.  
  
SpeCor addresses these limitations by incorporating mechanisms to correct stuttering and lisping, enhancing the accuracy of speech recognition. The system integrates advanced machine learning algorithms trained on diverse datasets, enabling it to recognize and correct errors associated with speech disorders and accents.  
  
Broad applications and benefits of SpeCor include improved communication, accessibility, education and employment, content creation and social interaction, and content creation and social interaction. The objectives and approaches of SpeCor include speech-to-text conversion, speech recognition, real-time speech correction, and user-friendly design and versatile applications.  
  
Case studies and supporting research have demonstrated the effectiveness of Whisper models in generalization and accuracy, machine learning for stuttering identification, speech to text and text to speech recognition systems, speech-based text correction for the visually impaired, and FastCorrect, which introduced a novel non-autoregressive error correction model that significantly improves ASR error correction efficiency and accuracy.  
  
Future work should focus on expanding the variety of speech patterns in training datasets, improving model architectures to handle diverse speech more effectively, and addressing biases in ASR systems to ensure fair and accurate recognition for all users.

# **References**

[1] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, D. Povey, “Zipformer: A faster and better encoder for automatic speech recognition,” arXiv.org. https://arxiv.org/abs/2310.11230 (accessed Jun. 27, 2024).

[2] D. Prabhu, P. Jyothi, S. Ganapathy, and V. Unni, “Accented Speech Recognition With Accent-specific Codebooks,” arXiv.org. https://arxiv.org/abs/2310.15970 (accessed Oct. 26, 2023).

[3] H. O. Toyin, A. Djanibekov, A. Kulkarni, and H. Aldarmaki, “ArTST: Arabic Text and Speech Transformer,” arXiv.org. https://arxiv.org/abs/2310.16621v1 (accessed Jun. 27, 2024).

[4] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. M. Pino, “fairseq S2T: Fast Speech-to-Text Modeling with fairseq,” arXiv (Cornell University), Oct. 2020, doi: https://doi.org/10.48550/arxiv.2010.05171.

[5] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, “SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing,” arXiv:2110.07205 [cs, eess], May 2022, Available: https://arxiv.org/abs/2110.07205.

[6] Y. Leng, X. Tan, W. Liu, K. Song, R. Wang, X.-Y. Li, T. Qin, E. Lin, And T.-Y. Liu, “SoftCorrect: Error Correction with Soft Detection for Automatic Speech Recognition,” arXiv.org, Dec. 02, 2022. https://arxiv.org/abs/2212.01039

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv:2212.04356 [cs, eess], Dec. 2022, Available: https://arxiv.org/abs/2212.04356.

[8] S. Dutta, S. Jain, A. Maheshwari, G. Ramakrishnan, and Preethi Jyothi, “Error Correction in ASR using Sequence-to-Sequence Models,” arXiv (Cornell University), Feb. 2022, doi: https://doi.org/10.48550/arxiv.2202.01157.

[9] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, “MixSpeech: Data Augmentation for Low-resource Automatic Speech Recognition,” arXiv:2102.12664 [cs, eess], Feb. 2021, Available: https://arxiv.org/abs/2102.12664

[10] Y. Hu, X. Jing, Y. Ko, and J. T. Rayz, “Misspelling Correction with Pre-trained Contextual Language Model,” arXiv:2101.03204 [cs], Jan. 2021, Available: https://arxiv.org/abs/2101.03204

[11] Y. Leng, X. Tan, L. Zhu, J. Xu, R. Luo, L. Liu, T. Qin, X.-Y. Li, E. Lin, and T.-Y. Liu, “FastCorrect: Fast Error Correction with Edit Alignment for Automatic Speech Recognition,” arXiv (Cornell University), May 2021, doi: https://doi.org/10.48550/arxiv.2105.03842.

[12] O. Basystiuk, N. Shakhovska, V. Bilynska, O. Syvokon, O. Shamuratov, and V. Kuchkovskiy, “The Developing of the System for Automatic Audio to Text Conversion,” ceur-ws.org. https://ceur-ws.org/Vol-2824/, Paper 1, (2018)

[13] A. Mohamed, Dmytro Okhonko, and L. Zettlemoyer, “Transformers with convolutional context for ASR,” arXiv (Cornell University), Jan. 2019, doi: https://doi.org/10.48550/arxiv.1904.11660.

[14] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” arXiv:2005.08100 [cs, eess], May 2020, Available: https://arxiv.org/abs/2005.08100

[15] M. C. Stoian, S. Bansal, and S. Goldwater, “Analyzing ASR pretraining for low-resource speech-to-text translation,” arXiv.org, Feb. 09, 2020. https://arxiv.org/abs/1910.10762

[16] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, “Jasper: An End-to-End Convolutional Neural Acoustic Model,” arXiv:1904.03288 [cs, eess], Aug. 2019, Available: https://arxiv.org/abs/1904.03288

[17] A. Dash, N. Subramani, T. Manjunath, V. Yaragarala, and S. Tripathi, “Speech Recognition and Correction of a Stuttered Speech.” Accessed: Jun. 27, 2024. [Online]. Available: https://e-tarjome.com/storage/panel/fileuploads/2022-07-06/1657094472\_e16845.pdf

[18] Y. Miao, M. Gowayyed, and F. Metze, “EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding,” arXiv (Cornell University), Jul. 2015, doi: https://doi.org/10.48550/arxiv.1507.08240.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, doi: https://doi.org/10.1109/icassp.2015.7178964.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline,” arXiv:1709.05522 [cs], Sep. 2017, Available: https://arxiv.org/abs/1709.05522

[21] T. Kourkounakis, A. Hajavi and A. Etemad, "FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2986-2999, 2021, doi: 10.1109/TASLP.2021.3110146.

[22] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, “Common Voice: A Massively-Multilingual Speech Corpus,” arXiv.org, Mar. 05, 2020. https://arxiv.org/abs/1912.06670

[23] R. Cattoni, M. A. Gangi, L. Bentivogli, M. Negri, and M. Turchi, “MuST-C: A multilingual corpus for end-to-end speech translation”. In Computer Speech & Language Journal, 2020, doi: https://doi.org/10.1016/j.csl.2020.101155

[24] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, Z. Peng, “WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition,” arXiv.org, Feb. 23, 2022. https://arxiv.org/abs/2110.03370 (accessed Jun. 27, 2024).

[25] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” arXiv:2006.11477 [cs, eess], Oct. 2020, Available: https://arxiv.org/abs/2006.11477