UE21CS390A – Capstone Project Phase – 1

Project Progress Review #2
(Project Requirements Specification and Literature Survey)

Project Title : Generating Clear speech from Speech impaired Audio
Project ID : PW24_RS_03
Project Guide : Dr. Ramamoorthy Srinath
Project Team with SRN : Roseline Jerry A - PES1UG21CS500
T P Shriambhikesh - PES1UG21CS659
Tanmay Praveen Udupa - PES1UG21CS662
Vandana S - PES1UG21CS697

Agenda

- Abstract
- Suggestions from Review-1
- Constraints
- Assumptions
- Risks
- Functional Requirements
- Non functional Requirements
- Literature Survey
  - Parrotron Paper
  - Project Euphonia
  - Articulatory features of ASR for Pathological speech
  - Arabic Dysarthric Speech Recognition Using Adversarial and Signal-Based Augmentation
- Conclusions
- References

Abstract

- Effective communication is essential for individuals to express themselves, convey ideas, and connect with others. However, for people with speech impairments this can be challenging. Whether caused by medical conditions, developmental disorders, or other factors, speech difficulties often create barriers to clear and understandable communication.

- In response to this challenge our project aims to develop a solution that enhances impaired speech outputs, ensuring clearer communication for these individuals.

Abstract

Overview of Scope:

1. Development of generic speech synthesis system
2. Providing clear and natural sounding speech
3. Real time Adaptations
4. User friendly interactions and Customization
5. Integration of feedback mechanism
6. Ethical standards and privacy considerations
7. Collaboration with Healthcare professionals

Suggestions from Review - 1

1. **Treating impaired speech as a Different language :**

   It was suggested to treat impaired speech as a distinct language to better understand its characteristics. This approach could involve analysing speech patterns, phonetic variations unique to individuals with speech impairment.

1. **Ensuring effectiveness of feedback mechanism:**

   These mechanisms play a crucial role in gathering user input, evaluating performance of the ML model and identifying areas of improvement. It's important that the feedback mechanism is user friendly and capable of catching relevant insights to further improve it.

1. (multimodality)

Constraints

Legal implications:

The project must adhere to legal regulations and standards governing speech synthesis, privacy, and data protection.Intellectual property rights must be considered when utilizing proprietary algorithms or datasets in the development of the speech synthesis system.

Software/Hardware requirements:

This includes access to computational resources for training and deploying the speech synthesis model, as well as compatibility with operating systems and development environments. Dependencies on third-party libraries, frameworks, or APIs must also be considered and managed effectively.

Ethical Considerations, Privacy and Security etc

Dependencies

Hardware and Software Infrastructure:

The software's performance relies on the host computer's hardware and network infrastructure, necessitating their availability and proper functioning.

Speech Synthesis Algorithms:

The project relies on effective speech synthesis algorithms to generate natural-sounding speech output, with a focus on optimizing them for individuals with speech impairments.

User Engagement and Feedback:

The success of the project relies on active user engagement and feedback, requiring effective strategies for soliciting and incorporating user insights into testing, evaluation, and improvement process

Assumptions

Availability of sufficient data:

This project assumes access to sufficient and diverse datasets of impaired speech samples, for training and testing the speech synthesis model. This includes data representing various types and degrees of speech impairments to ensure the robustness and generalizability of the system.

Collaboration with Health professionals:

The project assumes collaboration with speech therapists, healthcare professionals, and other domain experts to validate the effectiveness of the speech synthesis system and align with therapeutic goals. This collaboration is essential for understanding user needs, gathering feedback, and ensuring the ethical and responsible development of the application.

Risks

- **Low Testing Accuracy:** Variability in speech patterns, background noise, and individual speech impairments may lead to lower accuracy rates during testing.
- **Errors/Bugs:** Algorithm implementation issues, data preprocessing problems, or integration errors with other software components could result in functional and reliability issues, leading to incorrect speech conversions or system failures.
- **Data Bias and Imbalance:** Training data may exhibit bias or imbalance, such as overrepresentation or underrepresentation of certain speech patterns or demographics, impacting the model's performance, especially for individuals with less common speech impairments or diverse linguistic backgrounds.
- **Regulatory Compliance:** Compliance with data protection laws and medical device regulations is crucial, especially for healthcare applications. Non-compliance may lead to legal consequences, financial penalties, or reputational damage.

Functional Requirements

1. Speech-to-Text Conversion
- Accurate speech recognition technology for various accents, dialects, and speech impairments.
- Robust handling of background noise and environmental factors.
- Ability to detect and handle multiple speakers in a conversation or audio stream.

Functional Requirements

2. Text Correction:

- Natural language processing techniques to identify and correct errors, and impaired speech patterns.
- Ability to learn and adapt to individual users' speech patterns and improve correction accuracy over time.

Functional Requirements

3. Text-to-Speech Conversion:

- High-quality speech synthesis with natural intonation and pronunciation.
- Ability to handle complex text inputs, including proper names, abbreviations, numbers, and special characters.
- Efficient memory and resource management for handling large or continuous text inputs.

Functional Requirements

4. User Feedback on Speech Output:
   - Incorporation of user feedback into machine learning models or algorithms to improve speech synthesis.
   - Ability to capture and analyze user feedback data to identify areas for improvement.

Non - Functional Requirements

1. **Performance:** The system should process speech in real-time or near-real-time to ensure timely responses.
2. **Accuracy:** The generated clear speech should closely match the intended message of the disordered speech, minimizing errors and distortions.
3. **Robustness:** The system should be capable of handling various types and severities of speech disorders effectively without significant degradation in performance.
4. **Scalability:** The system should be able to handle varying loads of speech processing tasks, accommodating increasing demands without sacrificing performance.
5. **Compatibility:** It should integrate seamlessly with existing speech recognition systems or other applications where clear speech generation is needed.
6. **Security:** Measures should be in place to ensure the security and privacy of the processed speech data, especially if it contains sensitive information.

Non - Functional Requirements

7. User Interface: The system should have a user-friendly interface, making it easy for users to interact with and configure settings as needed.

8. Accessibility: The system should be accessible to users with disabilities, such as providing alternative input methods for those who cannot produce clear speech.

9. Adaptability: The system should be adaptable to different languages, dialects, and speech disorders, allowing for broader usage across diverse populations.

10. Reliability: The system should operate reliably under varying conditions, minimizing downtime and ensuring consistent performance.

11. Ethical Considerations: The project should adhere to ethical guidelines, ensuring that the generated clear speech is used responsibly and ethically, without causing harm or perpetuating biases.

Literature Survey

1. Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation [3]

Tanmay Praveen Udupa (PES1UG21CS662)

Literature Survey

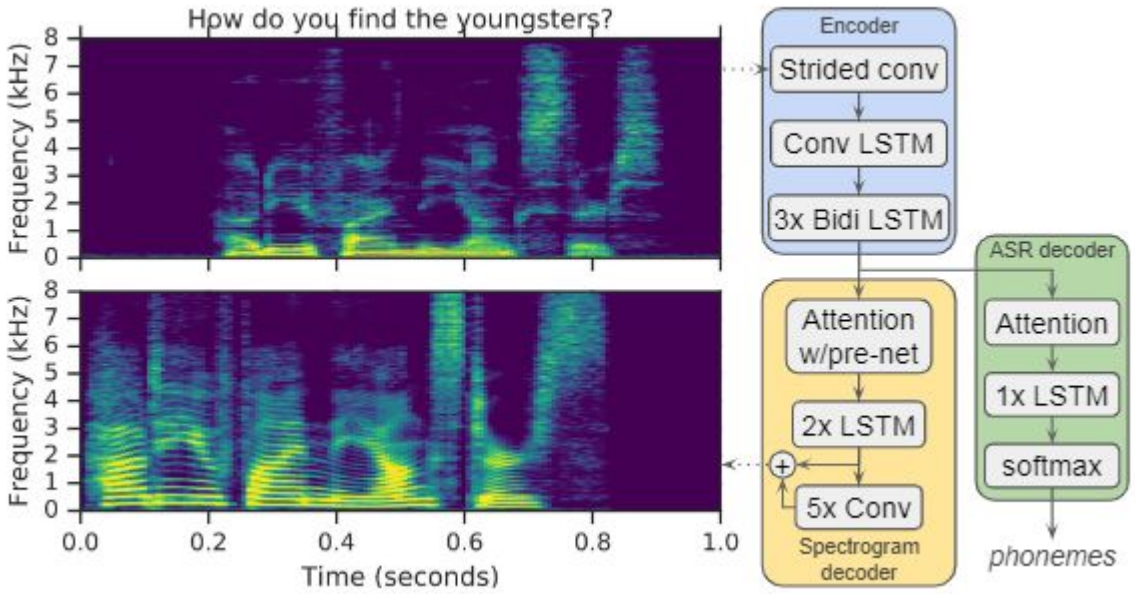
Abstract

- Parrotron is an end-to-end-trained speech-to-speech conversion model that maps an input spectrogram directly to another spectrogram, without utilizing any intermediate discrete representation.

- This model can be trained to normalize speech from any speaker regardless of accent, prosody, and background noise, into the voice of a single canonical target speaker.

Model architecture

- Composed of an encoder and a decoder with attention, followed by a vocoder to synthesize a time-domain waveform.
- The encoder converts a sequence of acoustic frames into a hidden feature representation which the decoder consumes to predict a spectrogram.



Literature Survey

Applications

- **Voice normalization**
  - Addresses the task of normalizing speech from an arbitrary speaker to the voice of a predefined canonical speaker.
  - Data: Model is trained on a ~30,000 hour training set consisting of about 24 million English utterances which are anonymized and manually transcribed.
  - Metric used: Word Error Rates (WERs)
  - Metric score: The WER on the original speech (matched condition) is 8.3%, which is viewed as an upper bound. The best-performing Parrotron model yields 17.6% WER.

Literature Survey

- **Normalization of hearing-impaired speech**
  - Investigates whether the normalization model can be used to convert atypical speech from a deaf speaker into fluent speech.
  - Data: 15.4 hours of speech, corresponding to read movie quotes.
  - Metric used: Word Error Rates (WERs)
  - Metric score: The best finetuning strategy was adapting all parameters, and dramatically reduced the WER from 89.2% (real speech) to 32.7%.
- **Speech separation**
  - Addresses the task of reconstructing the signal from the loudest speaker within a mixture of overlapping speech.

PES UNIVERSITY

Literature Survey

Future work

In the future, they plan to test it on other speech disorders, and adopt techniques from [5, 6] to preserve the speaker identity.

Relevance of the paper to this project

- This paper is relevant to this project as it demonstrates the ability of the model to handle speech from a deaf speaker, which is an example of an impaired speech
- Since the model has a possible application in separating the loudest speaker among overlapping speakers, we could use this model to remove the background noise in the input speech thereby improving the speech clarity.

2. Project Euphonia: Personalizing ASR for Dysarthric and Accented Speech with Limited Data [4]

T P Shriambhikesh (PES1UG21SC659)

Literature Survey

Abstract

- This paper presents and evaluates **fine tuning techniques** to improve ASR for users with non-standard speech.
- Personalized models trained using this approach bring 62% and 35% relative WER improvement over the standard ALS. Bringing the absolute WER down to **10% in mild dysarthria and 20% in severe dysarthria.**
- Proves that **finetuning a particular subset of layers** often gives better results than finetuning the entire model.

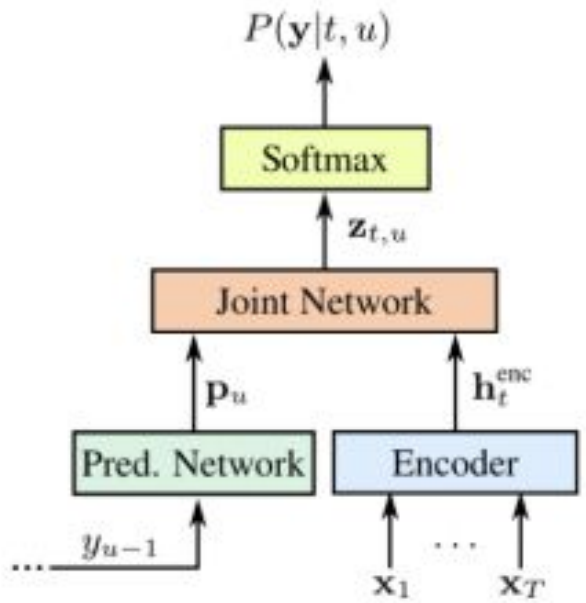
Literature Survey

Model architecture

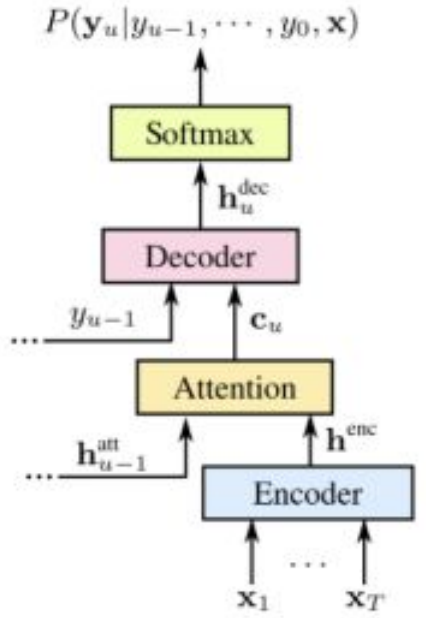
- A personalized ASR model by starting with a **base model trained on standard, unaccented speech.**(More resource efficient than retraining the entire model.)
- This paper experiments with two different base models:
  - Bi-directional RNN-Transducer [i]
  - Listen, Attend and Spell(LAS) [ii]
- Both are end to end sequence to sequence models

[i] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, “A comparison of sequence-to-sequence models for speech recognition.” in Interspeech, 2017, pp. 939–943.
[ii] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.

Schematic Diagrams



RNN-T architecture



LAS architecture

Finetuning

- With **RNN-T architecture**, 1st, 2nd, and 3rd layer was finetuned in fixed combination on both datasets. Training from E_0 up (Where lower number represents layer closer to the input) with or without the joint layer produced promising results.

- With **LAS architecture**, various combinations were tested and consistently found that best results from this network came from finetuning the entire network.

Literature Survey

Datasets

- **ALS**
  - 36.7 hours of audio from 67 people with ALS, in partnership with the ALS Therapy Development Institute.
- **Accented Speech**
  - L2 Arctic dataset of non native speech, consists of 20 speakers with approximately 1 hour speech per speaker.

Table 1: *Average WER Improvements*

	Cloud	RNN-T		LAS	
		Base	Finetune	Base	Finetune
Arctic[1]	24.0	13.3	8.5	22.6	11.3
ALS[2]	42.7	59.7	20.9	86.3	31.3
ALS[3]	13.1	33.1	10.8	49.6	17.2

[1] Non-native English speech from the L2-Arctic dataset. [24]
[2] Low FRS (ALS Functional Rating Scale) intelligible with repeating, Speech combined with nonvocal communication.
[3] FRS-3 detectable speech disturbance. [4]

Literature Survey

Future Work

- Open question on **additional techniques** that can be helpful in the low data regime (Virtual Adversarial Training, data augmentation, etc.)
- Explore **pooling data from multiple speakers** with similar conditions which where out of scope with respect to this paper.

Relevance

- Explores and evaluates the technique of fine tuning already existing models using smaller datasets.

3. Articulatory Features for ASR of Pathological Speech [4]

Vandana S (PES1UG21CS697)

Literature Survey

Abstract:

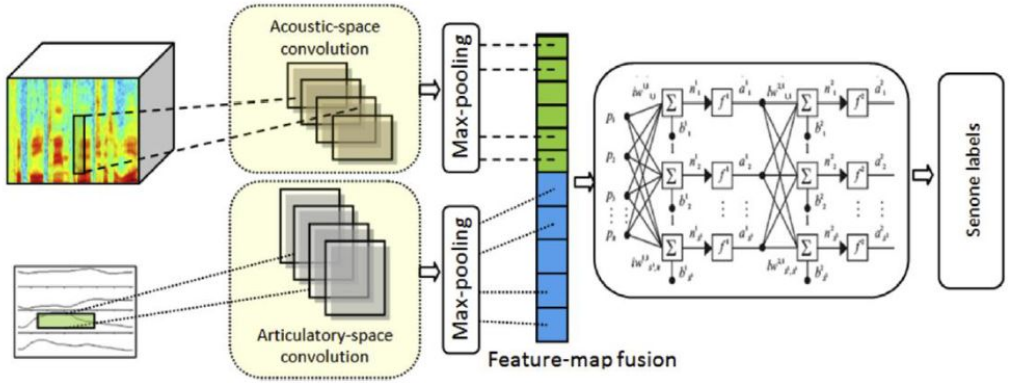
This study investigates the joint use of Articulatory and Acoustic features in **Automatic Speech Recognition (ASR) systems** for pathological speech, particularly dysarthria. Despite extensive efforts, current ASR systems struggle with lower performance on pathological compared to normal speech. This is due to the high variability caused due to articulatory impairments.

To address this, a fused-featuremap convolution neural network (fCNN) that **integrates articulatory and acoustic information** is proposed. The fCNN is evaluated against conventional CNNs and time frequency convolutional networks (TFCNNs) using Dutch and Flemish pathological speech corpora.

Literature Survey

Model Architecture:

- The fCNN model proposed for dysarthric speech recognition combines acoustic and articulatory features to address ASR challenges.
- It consists of two specialized convolutional layers: one for acoustic features, derived from filterbank energy, and another for articulatory features, specifically tongue and lip movement trajectories.
- Max-pooling is applied post-convolution to extract salient features, which are fused before input to a single neural network.
- This design enables comprehensive integration of acoustic and articulatory information, enhancing ASR performance for dysarthric speech.



Literature Survey

Training Dataset :

- The EST Dutch dysarthric speech database. [10]
- Components: Read speech, spontaneous conversations, interviews, and discussions.
- Duration: Normal Flemish (FL) speech data: 186.5 hours, Northern Dutch (NL) speech data: 255 hours, EST Dutch dysarthric speech database (Dys. NL): 6 hours and 16 minutes from 16 speakers.
- Exclusions: Segments with pronunciation errors, single words, and pseudowords were excluded to maintain the integrity of ASR performance evaluation.

Testing Dataset:

- CHASING01 Dutch Dysarthric Speech Data [9]
- Flemish COPAS Database [11]

Literature Survey

Results:

- The results show that combining articulatory features with acoustic features improves speech recognition accuracy, especially for dysarthric speech.
- In Dutch test sets, the fCNN model achieved the best performance with a Word Error Rate (WER) of 19.1%, surpassing other models like CNN and TFCNN.
- In Flemish test sets, the fCNN model consistently outperformed others, especially when trained on combined Flemish and Dutch data.
- This highlights the potential of using articulatory features to enhance ASR performance in pathological speech.

Literature Survey

Metrics- Word Error Rate (WER): WER is used as the primary metric to evaluate the performance of ASR systems. It measures the percentage of words in the recognized output that differ from the reference transcription.

Future Works-

- Further Investigation of Articulatory Features: Explore additional uses and benefits of articulatory features in ASR systems.
- Model Optimization and Fine-Tuning: Improve performance and efficiency by optimizing hyperparameters and refining model architectures.
- Robustness and Generalization: Assess how ASR systems handle variations in dysarthric speech characteristics and evaluate their ability to work across different languages and dialects.
- User-Centric Evaluation: Conduct evaluations focusing on user needs and experiences to ensure the ASR systems meet their requirements effectively.
- Integration with Assistive Technologies: Investigate integrating ASR systems with assistive technologies to improve accessibility for individuals with dysarthria. Additionally, explore real-time ASR applications for smoother communication in everyday situations.

Literature Survey

Relevance to Project:

- Incorporating articulatory features alongside acoustic ones in ASR systems is crucial for addressing challenges in pathological speech.
- This comprehensive approach enables a deeper understanding of unique speech patterns like dysarthria, leading to clearer communication.
- Articulatory features also allow for tailored solutions, ensuring personalized communication for individuals with speech difficulties.

Literature Survey

4.Arabic Dysarthric Speech Recognition Using Adversarial and Signal-Based Augmentation [1]

Roseline Jerry A (PES1UG21SC500)

Literature Survey

Abstract:

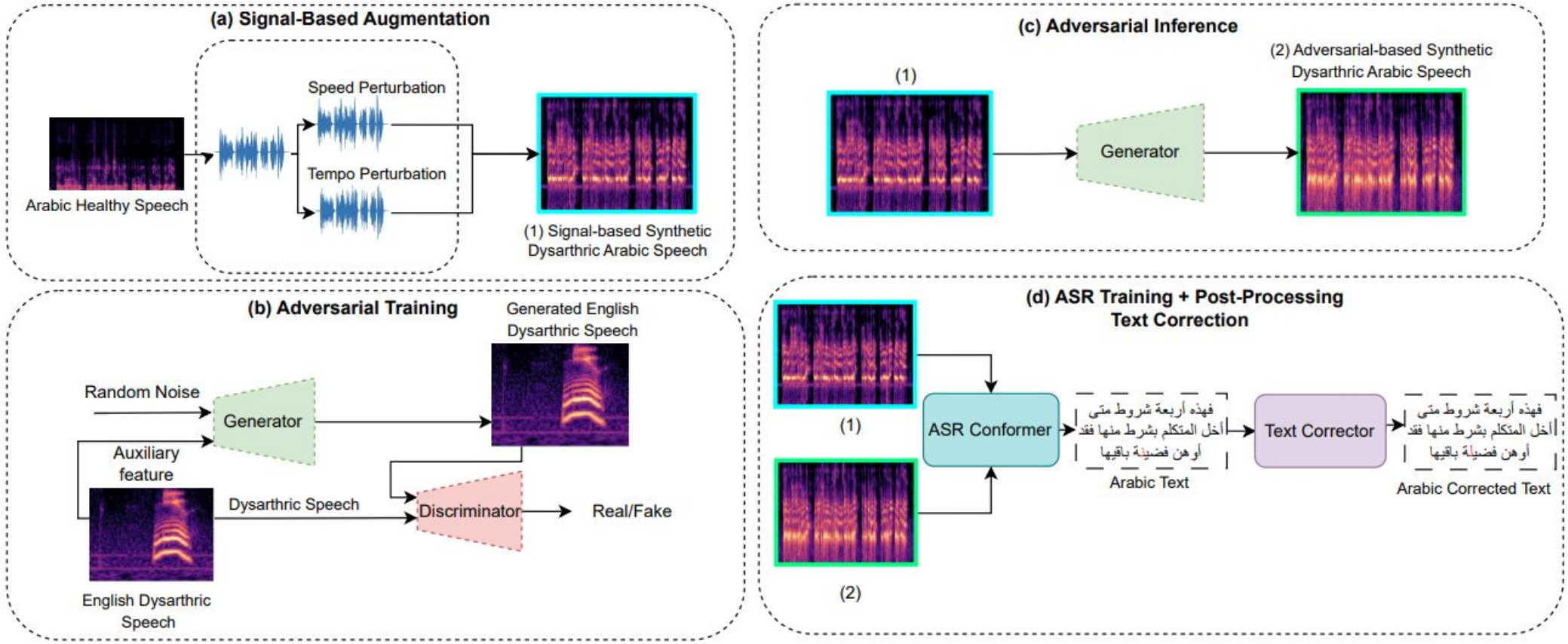
- The authors address Arabic dysarthric automatic speech recognition by proposing a multi-stage augmentation approach. They first modify healthy Arabic speech to simulate dysarthric speech, then employ a Parallel Wave Generative adversarial model trained on English dysarthric data to further enhance the samples.
- Additionally, they fine-tune an Arabic Conformer model and utilize text correction strategies for different dysarthric speech severity levels. The approach significantly improves word error rate (WER) by 81.8% compared to baseline on Arabic synthetic data and 124% on real English dysarthric speech.

Model architecture

- The paper proposes a multi-stage augmentation approach to improve the performance of Arabic dysarthric automatic speech recognition.
- The first stage involves a signal-based approach to generate dysarthric Arabic speech from healthy Arabic speech by modifying its speed and tempo.
- The second stage utilizes a Parallel Wave Generative (PWG) adversarial model trained on an English dysarthric dataset to capture language-independent dysarthric speech patterns and further augment the signal-adjusted speech samples.

Literature Survey

Schematic Diagrams



Fine-tuning:

- The fine-tuning process involves training the Conformer model on the signal-based augmented data, which includes dysarthric Arabic speech generated from healthy Arabic speech by modifying speed and tempo.
- The fine-tuned Conformer model is evaluated based on Word Error Rate (WER) and Character Error Rate (CER) on synthetically generated dysarthric speech from the Arabic common voice speech dataset.

Datasets

The paper uses a combination of these datasets to train and evaluate their proposed augmentation approach for improving Arabic dysarthric speech recognition.

- **Common Voice dataset:** healthy Arabic speech dataset used to generate synthetic dysarthric Arabic samples through signal-based perturbation.
- **MGB-2 dataset:** healthy Arabic speech corpus used to train a Conformer model for Arabic ASR.
- **Torgo dataset:** English dysarthric speech dataset used to familiarize the adversarial generator model with dysarthric speech characteristics beyond speed and tempo.
- **LJspeech dataset:** English dataset used to synthetically generate dysarthric English speech for validation.

Future Work

1. Explore additional variations beyond speed and tempo adjustments in the signal-based approach to address the complexities of dysarthric speech recognition.

2. Investigate the impact of variable recording conditions, uncontrolled body movements, and unbalanced data samples across speakers and diseases on dysarthric speech recognition.

3. Integrate the Parallel Wave GAN with the dataset to reconstruct further dysarthric speech variations introduced through the signal-based component.

Summary of Literature Survey

https://docs.google.com/spreadsheets/d/1qJRPP-TT6xekS5LKJDgvfMRfBI4YqoF6MQSnrsQbzT4/edit#gid=0

Conclusion

From our survey we understood that

- The data scarcity challenges for non-standard speech
- Use of techniques like adaptation, augmentation, multi-stream modelling and direct speech conversion to improve ASR performance and intelligibility for impaired or accented speech

References

[1] Baali, M., Almakky, I., Shehata, S., & Karray, F. (2023). Arabic Dysarthric Speech Recognition Using Adversarial and Signal-Based Augmentation. arXiv preprint arXiv:2306.04368.

[2] “Diagnosis of Disordered Speech using Automatic Speech Recognition,” International Journal of Engineering Research, vol. 8, no. 11, 2020.

[3] Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, Ye Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," arXiv:1904.04169v3 [eess.AS], Oct. 29, 2019.

[4] J. Shor *et al.*, “Personalizing ASR for Dysarthric and Accented Speech with Limited Data,” in *Interspeech 2019*, Sep. 2019, pp. 784–788. doi: 10.21437/Interspeech.2019-1427.

[1]

References

[5] A. Haque, M. Guo, and P. Verma, “Conditional end-to-end audio transforms,” in Proc. Interspeech, 2018.

[6] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in Advances in Neural Information Processing Systems, 2018.

[7] E. Yılmaz, V. Mitra, C. Bartels, and H. Franco, “Articulatory Features for ASR of Pathological Speech.” arXiv, Jul. 28, 2018.Available: http://arxiv.org/abs/1807.10948

[8] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, “Multistage DNN training for automatic recognition of dysarthric speech,” in Proc. INTERSPEECH, Sept. 2017, pp. 2685–2689.

[1]

References

[9] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, “Combining non-pathological data of different language
varieties to improve DNN-HMM performance on pathological speech,” in Proc. INTERSPEECH, Sept. 2016, pp.
218–222.

[10] E. Yılmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, “A Dutch dysarthric speech database for
individualized speech therapy research,” in Proc. LREC, 2016, pp. 792–795.

[11] C. Middag, “Automatic analysis of pathological speech,” Ph.D. dissertation, Ghent University, Belgium, 2012.

[12] Jan Noyes, Clive Frankish (2009), "Speech recognition technology for individuals with disabilities, Augmentative and Alternative Communication," 8:4, 297-303, DOI: 10.1080/07434619212331276333

Thank You