AUTOMATIC SPEECH RECOGNITION SYSTEM FOR CHILDREN'S SPEECH

Under the supervision of

Dr. Hemant Kumar Kathania

Assistant Professor

Presented by:

Udara Laxman Kumar

Roll no: Phec220031

Research Scholar

Department of Electronics and Communication Engineering

National Institute of Technology Sikkim

Contents

Introduction

- An Automatic Speech Recognition (ASR) system is a technology that listens to spoken words and converts them into written text.
- The goal of an ASR system is to automatically and accurately convert spoken language into written text using computational algorithms and models.
- An ASR system functions by accurately transcribing spoken language into written text through analysis of acoustic features and language modeling.
- Applications are :
  1. Education.
  2. Voice assistance
  3. Voice search and navigation.

Course-work

Sr.No	Subject Name/Code	Teacher/Faculty	Credit point	Credit Unit
1	Research Methodology (ZZ31101) (3 credits + 1 credit)	Dr. Anindya Biswas	4.0	BC
2	Machine Learning (EC21104)	Dr. Hemant Kumar Kathania	3.0	AB
3	Linear Algebra Stochastic process and Optimization Technique (EC21101)	Dr. Avinash Kumar	3.0	AB
4	Speech Signal Processing and Coding (EC21145)	Dr. Hemant Kumar Kathania	3.0	BB

Table 1: 1^{st} 2^{nd} Semesters Course Work

We had the course work of 4 subjects, total of 13 credits. A score of 8.08 S.G.P.A is achieved.

Literature Survey

Authors	Year	Title	Database	Method	Performance
Barcovschi et al. [1]	2023	A comparative analysis between Conformer-Transducer, Whisper, wav2vec2 for improving the child speech recognition	CMU kids, CSLU kids My_ST	Conformer Transducer Wav2Vev2 Whisper	2.91% WER for pf_star for Wav2vec2
Vishwanath et al.[2]	2022	Spectral Modification Based Data Augmentation for Improving END-TO-END ASR for Children's Speech	Librispeech dataset	LPC analysis, VTLP	16.5% & 6.1% Rel. Imp. on Baseline 3.7 % & 5.1% Rel. reduction WER
Hemant K Kumar et.al. [3]	2022	A formant modification method for improved ASR of children's speech	WSJCAM0, PF_STAR databases	DNN,TDNN, VTLN, SRA	24% & 11% for narrowband 27% & 13% for wideband WER
S.Shanh nawazuddin et al. [4]	2022	Voice Conversion Based Data Augmentation to Improve children's Speech Recognition in Limited Data Scenario	2 British corpus Dataset	GAN-VC Outof Domain augumentation	WER of 7.7 Rel. improvement
Singh et.al[5]	2021	Data Augmentation Using CycleGAN for End-to-End Children ASR	Libri-clean-360, TLT School corpus	Cyclegan-VC	Baseline 27.11% W.E.R (S7) Proposed method W.E.R 23.50 %

Table 2: Literature Survey on related work - I

Authors	Year	Title	Database	Method	Performance
kim et al. [6]	2020	SpecMix : A Mixed Sample Data Augmentation method for Training with T-F Domain Features	VoiceBank, DEMAND	SpecMix speech	2.92 PESQ
Hemant K Kumar et al. [7]	2019	Data Augmentation Using Spectral Warping to robust personalised for Low Resource Children ASR	punjabi children TLT School corpus	Spectral warping, VTLP Prosody modification TDNN	WER of 32.13% -TLT 10.51%- punjabi children
S.Shanh nawazuddin et al. [8]	2017	Enhancing Noise and Pitch Robustness of for impaired speech	WSJCAM0 PF_STAR	SMAC, Pitch sensitivity reduction	6.30% and 12.54% proposed methods

Table 3: Literature Survey on related work - II

Motivation of the Research

- The ASR system, built primarily for adult speech, fails to perform well when tested with children's speech, showing a significant reduction in accuracy [9, 10].
- Children speak and sound differently from adults, and because of this, speech recognition algorithms fail to recognise children's speech when they get their training on how adults speak.
- Mismatch conditions correspond to training the system with adults' speech and testing it with children's speech. Most ASR systems available publicly work well with adults' speech but in the case of children's speech their performance collapses.

Effect of Linear Prediction Order to modify Formant Location

Idea:

- Children's Speech Recognition lags behind adults due to limited publicly available data.
- Neural network models require a large amount of speech data to achieve good performance, whereas for children, a limited amount of speech data is available publicly [11, 12].
- By exploring formant frequency modification using Linear Prediction, showing significant reductions in Word Error Rates [13] across various acoustic models.

- To perform the formant modification[14], the warped LP spectrum is obtained by applying the warping function $w_\alpha(f)$ to the original LP spectrum $S(f)$ computed from children's speech:

$$S_\alpha(f) = S(w_\alpha(f)). \tag{1}$$

- In the conventional analysis of LP, the estimation denoted as $\hat{s}(n)$, is computed using the LP coefficients a_k and the past P samples according to Equation (2).

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k). \tag{2}$$

- By applying the Z-transform to equation(2) is given by:

$$\hat{S}(z) = \left(\sum_{k=1}^{P} a_k z^{-k} \right) S(z), \tag{3}$$

- To introduce the desired formant modifications, we replace the unit delay filter allows us to warp the frequency scale [15] of the LP spectrum.

$$D(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{4}$$

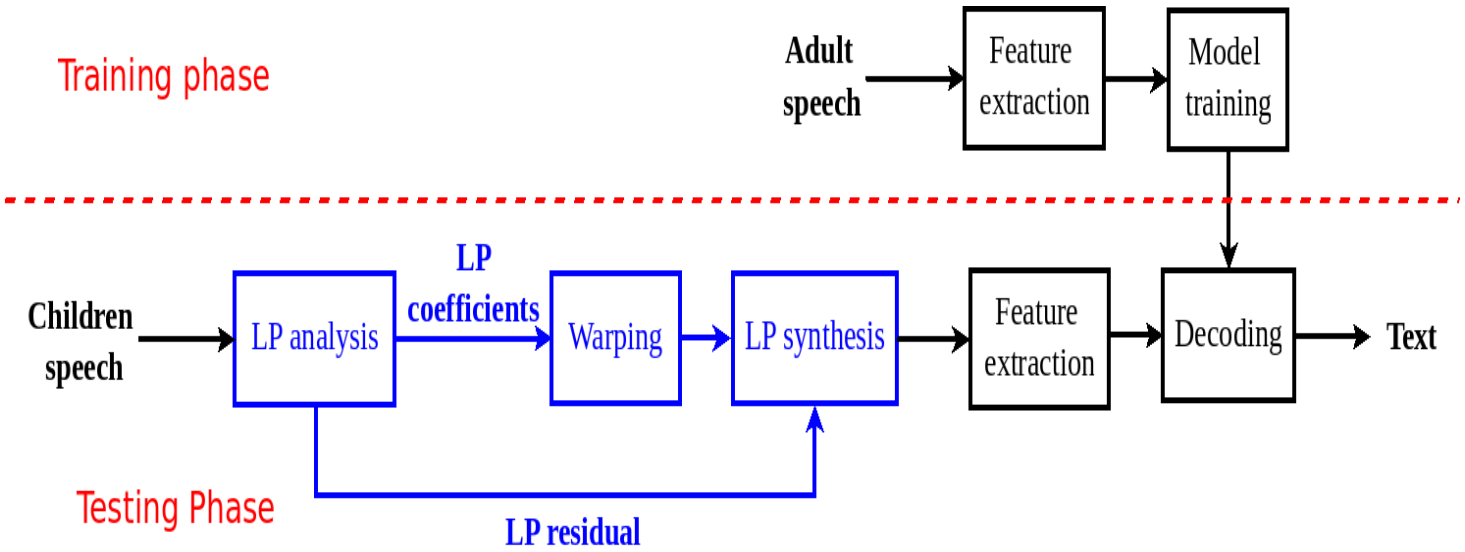

Figure 1: Block diagram of proposed method

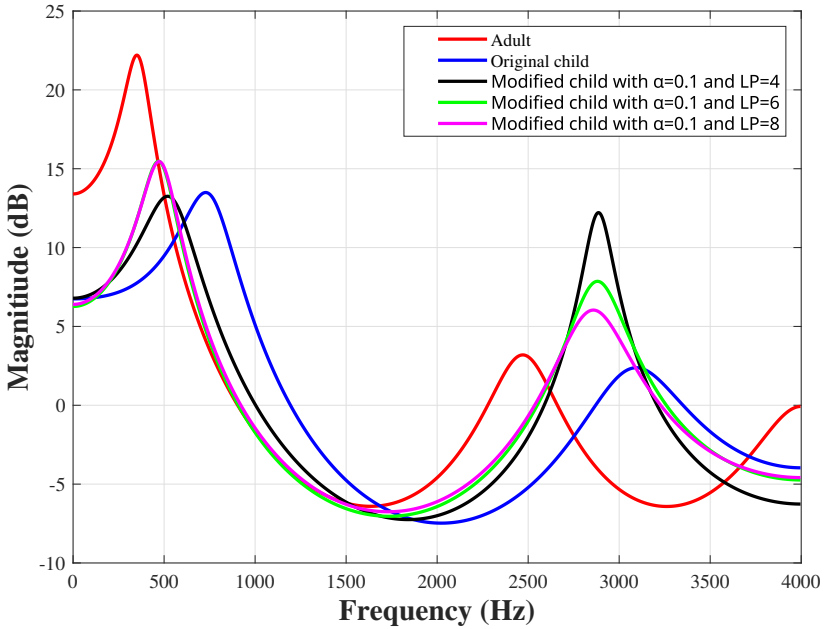

Figure 2: LP spectra computed from narrowband (8kHz) signal showing in formant frequencies.

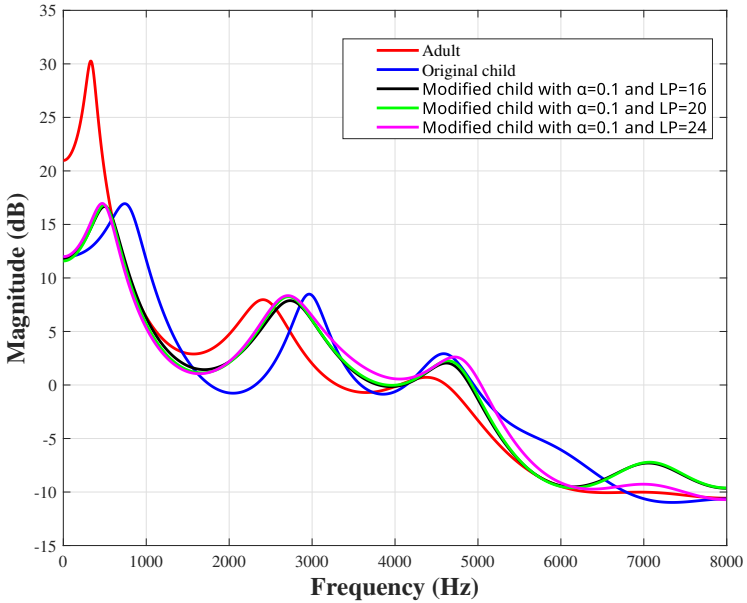

Figure 3: LP spectra computed from wideband (16kHz) signal showing variation in formant frequencies.

Two speech databases were utilized for the experiments - WSJCAM0 and PF-STAR.

Both speech databases are from British English corpora, indicating the linguistic context of the study.

Sl.No	Dataset	Duration	Age
1	WSJCAM0	15.5 Hrs	18-60 years

Table 4: WSJCAM0 Dataset

Sl.No	Dataset	Duration	Age
1	Pf_star	1.1 Hrs	4-14 years

Table 5: Pf_star Dataset

1. **ASR System Building:** The Kaldi setup [16] was employed to construct the Automatic Speech Recognition (ASR) system.

2. **Feature Extraction:** A 13-dimensional base Mel frequency cepstral coefficient (MFCC) including a 20-ms hamming window, 10-ms frame shift, and a 40-channel Mel-filterbank.

3. **Normalization Techniques:**
   - Cepstral mean and variance normalization (CMVN)
   - Cepstral feature-space maximum likelihood linear regression (fMLLR) for de-correlation and normalization [17], respectively.

4. **Acoustic Models Explored:** All models(GMM, DNN, TDNN) were trained using the hidden Markov model (HMM).

5. **DNN-HMM Training:**
   - A 5-hidden layer DNN with 1024 hidden nodes each, tanh nonlinearity.
   - Learning rate reduction from 0.015 to 0.002.

6. **Language Model (LM):** For decoding children's speech, a domain-specific bigram language model (LM) was used, trained on PF-STAR transcripts excluding the test set.

7. **TDNN Acoustic Model Training:** Kaldi setup was used for training the TDNN acoustic model, involving MFCC features, GMM alignment labels, i-vectors [18] for speaker adaptation.

8. **Training Parameters for TDNN:**
   - The initial learning rate was set at 0.0005, further reduced to 0.00005.

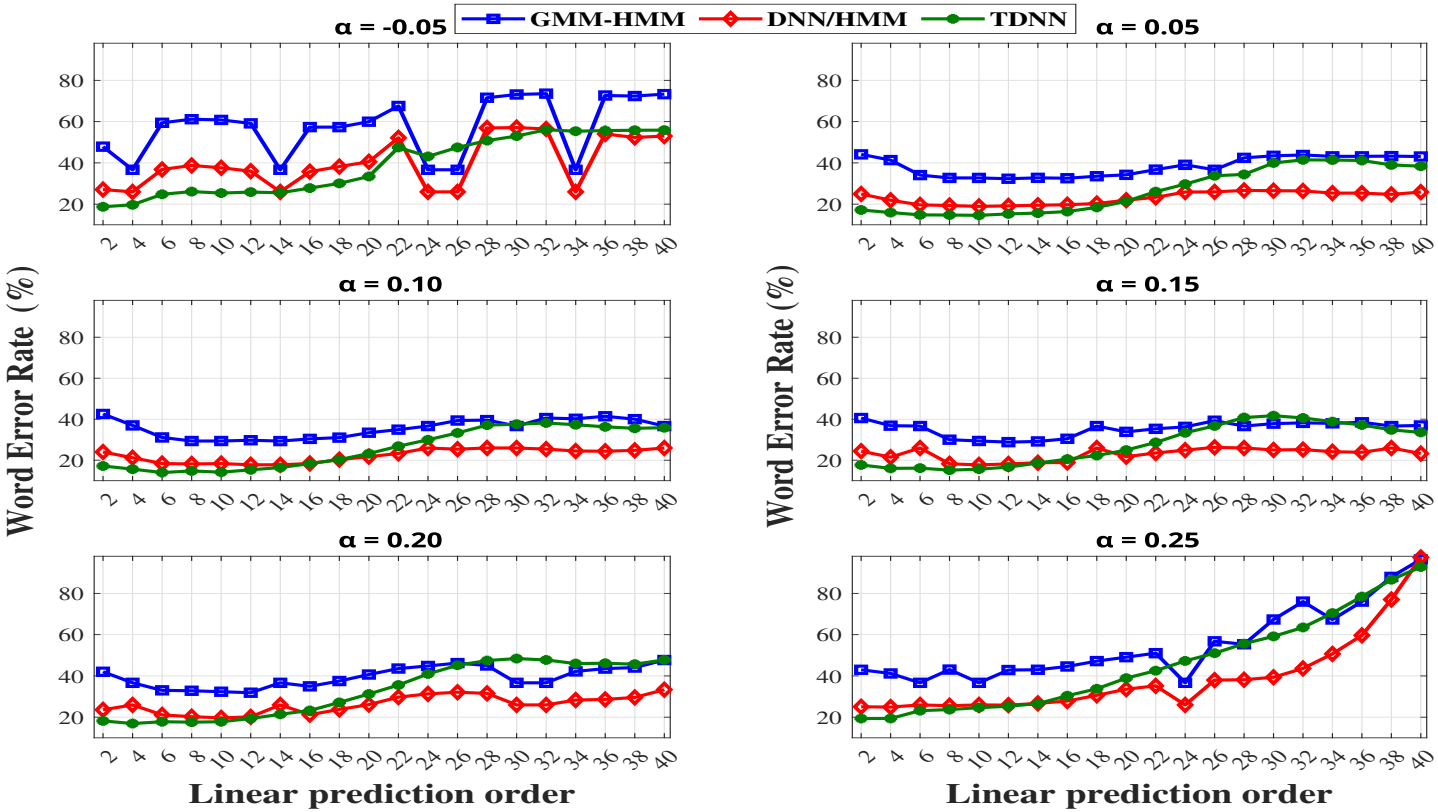

Figure 4: WERs with respect to linear prediction order varied from 2 to 40 with step size 2, and different formant modification factors α varied from -0.05 to 0.25 for children's narrowband (8 kHz) speech. The analysis is done for the GMM-HMM, DNN-HMM, and TDNN acoustic models.

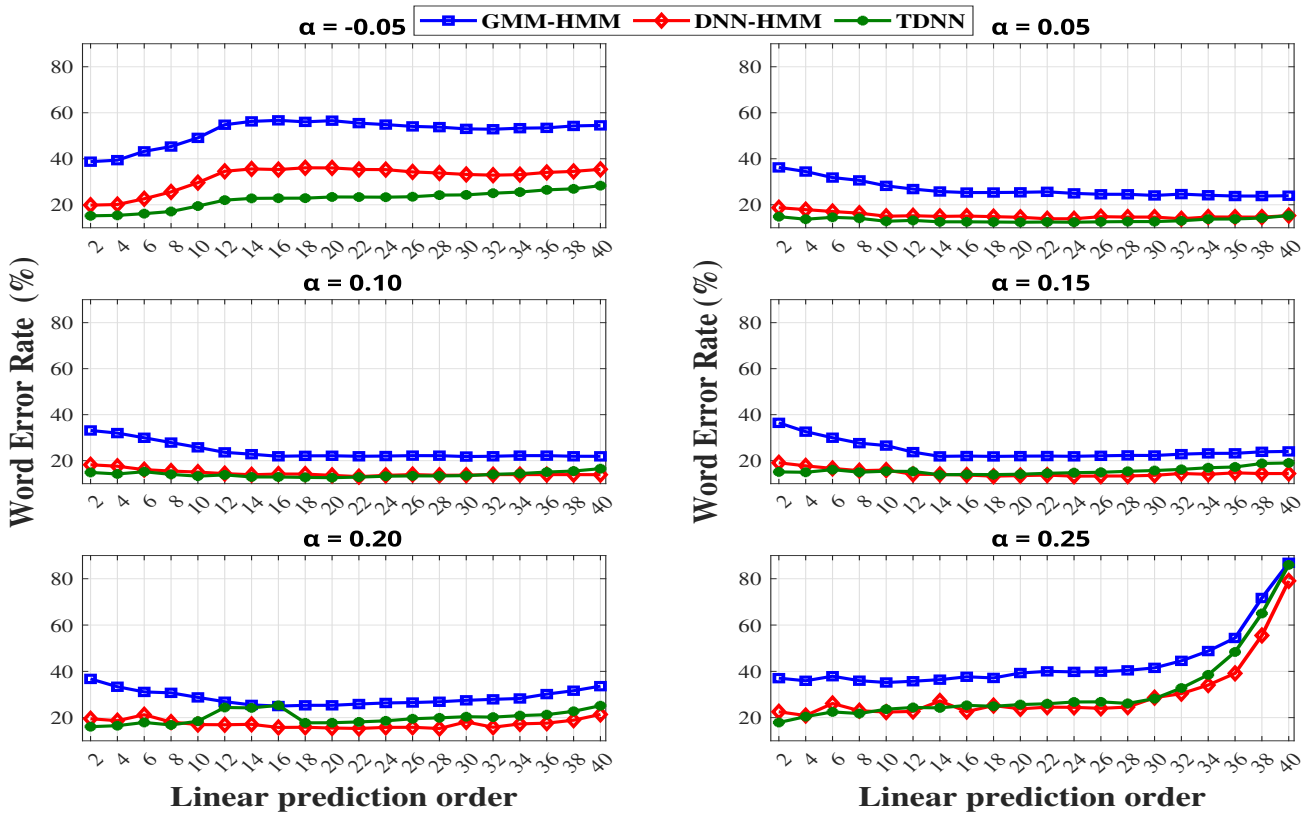

Figure 5: WERs with respect to linear prediction order varied from 2 to 40 with step size 2, and different formant modification factors α varied from -0.05 to 0.25 for children's wideband (16 kHz) speech. The analysis is done for the GMM-HMM, DNN-HMM, and TDNN acoustic models.

Comparison of ASR System WERs

Table 6: Baseline ASR system WERs for the GMM-HMM, DNN-HMM, and TDNN acoustic models.

Speech bandwidth	WER (%)		
	Acoustic model		
	GMM-HMM	DNN-HMM	TDNN
Narrowband	44.65	26.23	15.83
Wideband	32.83	19.58	14.16

Table 7: WERs for the Baseline system, the VTLN and SRA techniques, and the proposed best linear prediction order for the formant modification method. The experiments were conducted using the GMM-HMM, DNN-HMM, and TDNN acoustic models.

Acoustic model	Speech bandwidth	WER (%)			
		Baseline	VTLN	SRA	Proposed
GMM	Narrowband	44.65	35.06	34.23	32.85
	Wideband	32.83	24.30	22.04	20.59
DNN	Narrowband	26.23	23.34	21.98	19.95
	Wideband	19.58	15.17	16.68	14.22
TDNN	Narrowband	15.83	15.19	14.86	13.82
	Wideband	14.16	13.84	13.18	12.19

End-to-End ASR for Low-Resource Children's Speech

- End-to-End ASR integrates feature extraction, acoustic modeling, and language modeling into a single neural network, eliminating the need for separate modules
- These systems work by taking raw audio input, extracting relevant features directly, and predicting text outputs in a single neural network framework.
- End-to-End ASR systems convert spoken language directly into written text using a single model.
- End-to-End ASR models aim to streamline and improve the accuracy of speech recognition tasks.

Sl.No	Dataset	No.of Speakers	Age	Utterances
1	Pf_star	182	4-14 years	958
2	CMU Kids	76	6-11 years	5180
3	CSLU Kids	1100	4-14 years	49977

Table 8: Dataset Description

Sl. No	**Dataset**	**Train**	**Test**	**Model**	**W.E.R**
1	Pf_star	Pf_star	Pf_star	Transformer	17.6%
2	CMU Kids	CMU Kids	CMU Kids	Transformer	11.5%
3	CSLU Kids	CSLU Kids	CSLU Kids	Transformer	1.5%

Table 9: Baselines for different Low-Resource datasets with respective W.E.Rs

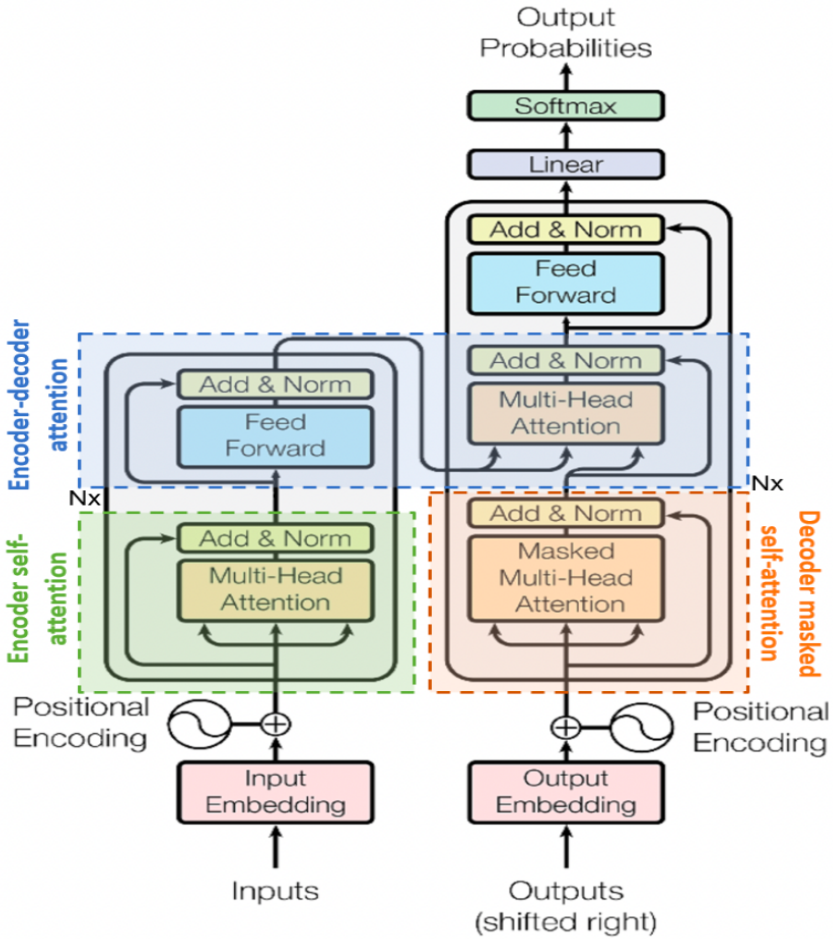

Figure 6: Transformer Model

Reference: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Å., & Polosukhin, I. (2017). **Attention is all you need**. *Advances in neural information processing systems, 30*.

Formant Masking

- To improve speech recognition systems for children, as they differ significantly from adult speech in pitch, pronunciation, and language usage.
- The primary aim was to study the effects of formant masking on the spectral characteristics of speech signals and understand how altering specific formants can impact the sound quality.
- Applied formant masking to modify the spectral characteristics of an audio signal, focusing on the first and third formants.
- Before applying the masking, the LPC spectrum showed clear peaks where the formants were. After masking, these peaks were reduced, changing the sound characteristics of the speech

Formant Masking

Sl. No	Dataset	Train	Test	Model	W.E.R
1	Pf_star	Pf_star	Pf_star	DNN	8.83%
2	Pf_star + fm1	Pf_star	Pf_star	DNN	6.83%

Table 10: Baselines for pf_star dataset without and with first formant masking with respective W.E.Rs

Formant Masking

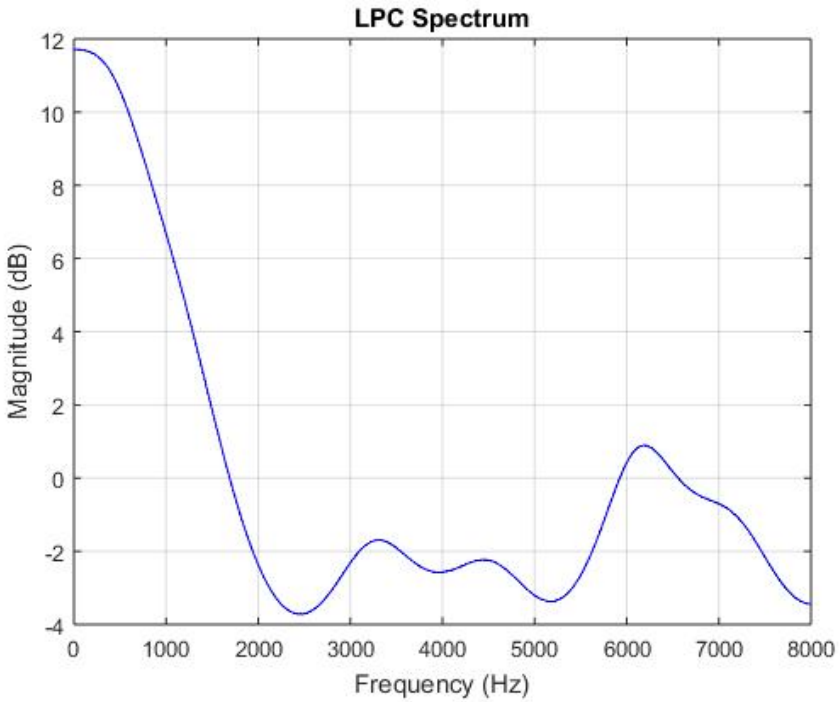

Figure 7: Without Formant Masking

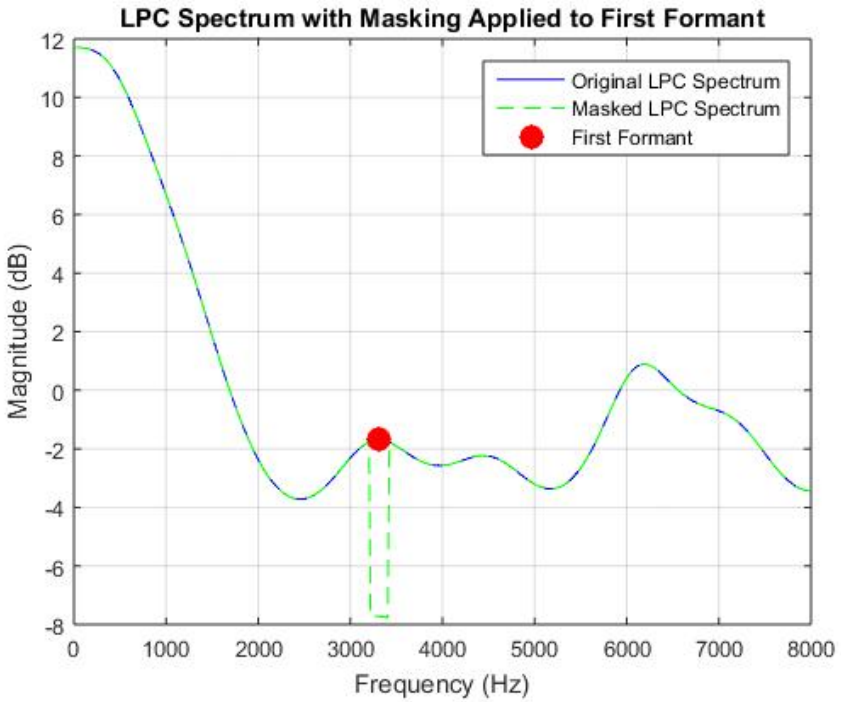

Figure 8: First Formant Masking

Formant Masking

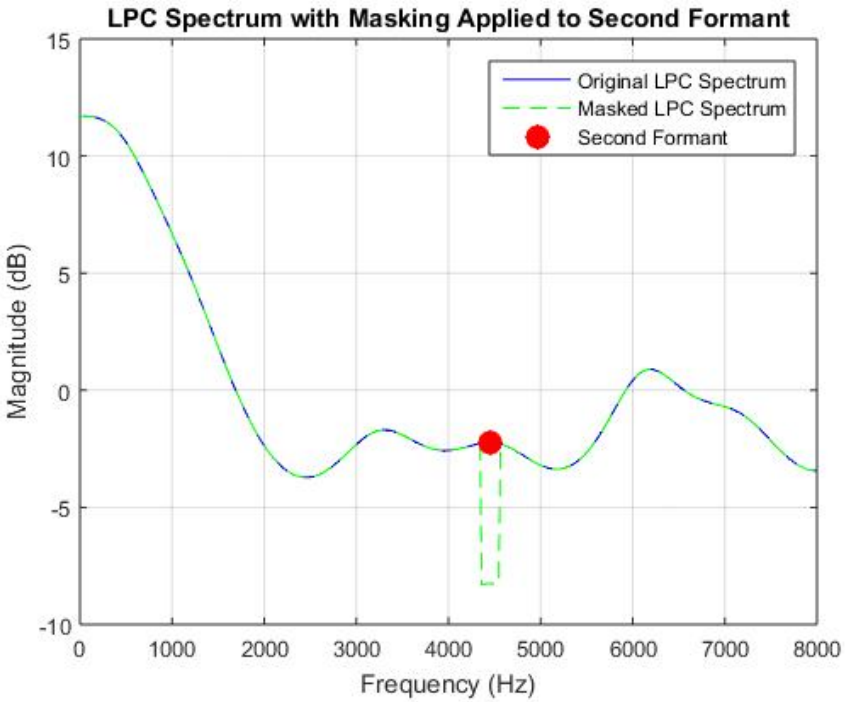

Figure 9: Second Formant Masking

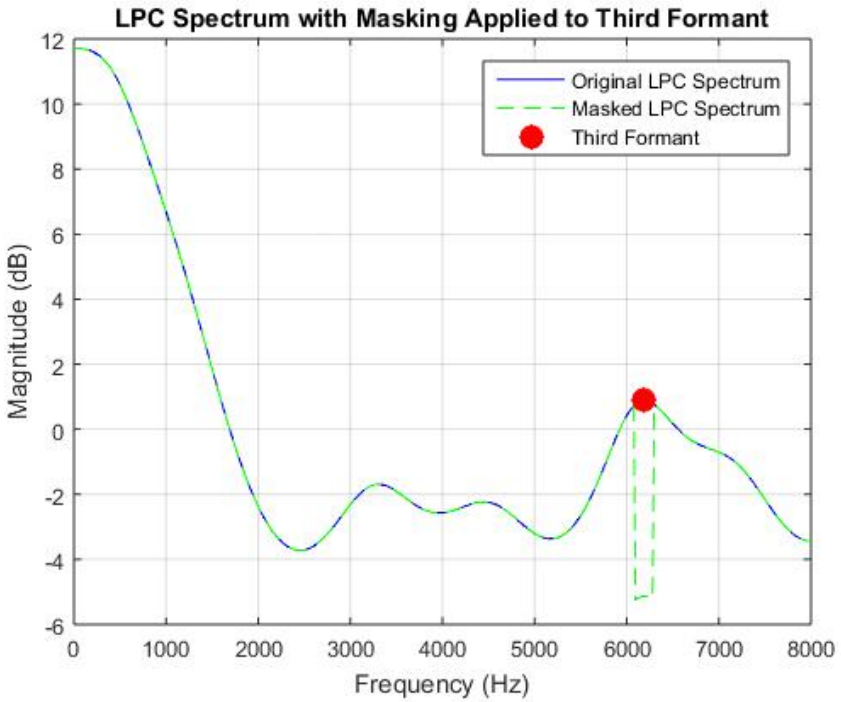

Figure 10: Third Formant Masking

Conclusion

- Adult-trained speech recognition models perform not better on children's speech due to vocal tract length differences, causing an acoustic mismatch between training and testing data.
- The impact of linear prediction order on the proposed formant modification method, finding that adjusting this parameter contributes to improved ASR system performance.
  - The proposed method significantly reduces WER, achieving
  - 13.82% for narrowband speech, 12.19% for wideband speech with the TDNN acoustic model.
- For an End-to-End ASR system for children's speech using a transformer-based model, among other datasets, CSLU Kids achieves the best and most robust performance with a W.E.R of 1.5%.

Future work

- Constructed an ASR system using the Kaldi toolkit, experimenting with models such as GMM-HMM, DNN, and TDNN, which performed well, especially with Low-Resource children's speech data.
- Developed an end-to-end ASR system using the transformer model from the ESPnet toolkit, which adapts to a variety of datasets and improves adaptability
- Investigating formant masking techniques to enhance ASR effectiveness on Low-Resource datasets.
- Determining the potential of formant masking to augment Low-Resource data for both traditional and End-to-End ASR systems.

.

List of Publications

- Kumar, U.L., Kurimo, M. and Kathania, H.K., 2023, November. **Effect of Linear Prediction Order to Modify Formant Locations for Children Speech Recognition**. In *International Conference on Speech and Computer* (pp. 483-493). Cham: Springer Nature Switzerland.

  .

References

[1] Andrei Barcovschi, Rishabh Jain, and Peter Corcoran.

A comparative analysis between conformer-transducer, whisper, and wav2vec2 for improving the child speech recognition.

In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 42–47. IEEE, 2023.

[2] Vishwanath Pratap Singh, Hardik Sailor, Supratik Bhattacharya, and Abhishek Pandey.

Spectral modification based data augmentation for improving end-to-end asr for children's speech.

arXiv preprint arXiv:2203.06600, 2022.

[3] Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, and Mikko Kurimo.

A formant modification method for improved asr of childrenâs speech.

Speech Communication, 136:98–106, 2022.

[4] S Shahnawazuddin, Nagaraj Adiga, Kunal Kumar, Aayushi Poddar, and Waquar Ahmad.

Voice conversion based data augmentation to improve children's speech recognition in limited data scenario.

In *Interspeech*, pages 4382–4386, 2020.

[5] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo.

Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion.

In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824, 2019.

[6] Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu.

Mixspeech: Data augmentation for low-resource automatic speech recognition.

In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7008–7012. IEEE, 2021.

[7] Hemant Kumar Kathania, Viredner Kadyan, Sudarsana Reddy Kadiri, and Mikko Kurimo.

Data augmentation using spectral warping for low resource children asr.

Journal of Signal Processing Systems, 94(12):1507–1513, 2022.

[8] Syed Shahnawazuddin, KT Deepak, Gayadhar Pradhan, and Rohit Sinha.

Enhancing noise and pitch robustness of children's asr.

In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5225–5229. IEEE, 2017.

References

[9] Vivek Bhardwaj, Mohamed Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B. Goud, Ateeq Rehman, Muhammad Shafiq, and Habib Hamam.
Automatic speech recognition (asr) system for children's: A systematic literature review.
Applied Sciences, 04 2022.

[10] A. Potamianos and S. Narayanan.
Robust recognition of children's speech.
IEEE Transactions on Speech and Audio Processing, 11(6):603–616, 2003.

[11] Prashanth Gurunath Shivakumar and Panayiotis Georgiou.
Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations.
Computer Speech and Language, 63:101077, 2020.

[12] S. Shahnawazuddin, N. Adiga, and H. K. Kathania.
Effect of prosody modification on children's ASR.
IEEE Signal Processing Letters, 24(11):1749–1753, 2017.

[13] H. K. Kathania, S. Shahnawazuddin, N. Adiga, and W. Ahmad.
Role of prosodic features on children's speech recognition.
2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5519–5523, 2018.

[14] Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, and Mikko Kurimo.
A formant modification method for improved asr of children's speech.
Speech Communication, 136:98–106, 2022.

[15] Hans Werner Strube.
Linear prediction on a warped frequency scale.
The Journal of the Acoustical Society of America, 68(4):1071–1076, 1980.

References

[16] D. Povey, G. Cheng, Y. Wang, Ke Li, H. Xu, M. Yarmohammadi, and S. Khudanpur.

Semi-orthogonal low-rank matrix factorization for deep neural networks.

In *Proc. INTERSPEECH 2018*, pages 3743–3747. ISCA, 2018.

[17] Shakti P. Rath, Daniel Povey, Karel Veselý, and January Černocký.

Improved feature processing for deep neural networks.

In *Proc. INTERSPEECH*, 2013.

[18] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny.

Speaker adaptation of neural network acoustic models using i-vectors.

In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 55–59. IEEE, 2013.

[19] Dipesh K. Singh, Preet P. Amin, Hardik B. Sailor, and Hemant A. Patil.

Data augmentation using cyclegan for end-to-end children asr.

In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 511–515, 2021.

[20] Syed Shahnawazuddin, Nagaraj Adiga, and Hemant Kumar Kathania.

Effect of prosody modification on children's asr.

IEEE Signal Processing Letters, 24(11):1749–1753, 2017.

Thank You