# VOWEL NON-VOWEL BASED SPECTRAL WARPING AND TIME SCALE MODIFICATION FOR IMPROVEMENT IN CHILDREN'S ASR

IEEE ICASSP-2021

Hemant Kathania, Avinash Kumar and Mikko Kurimo

# Overview

- Motivation
- A Non-Uniform LPC Based Formant Modification
- Experimental Setup and Speech Corpora
- Results and Discussion
- Conclusion

# Motivation

- ASR system is affected by several factor like inter-speaker variability such as age, gender, accent, speaking-rate,and formant frequencies of the speakers.
- To impart robustness towards this variability techniques like fMLLR and VTLN are used.
- Formant frequencies F1, F2, and F3 are higher in children's speech compared to adults' speech due to the shorter vocal tract length.
- Motivated by this issue, a non-uniform linear predictive coding (LPC) based formant modification technique is proposed by considering whether the given frame of speech is voiced/unvoiced.

# A Non-Uniform LPC Based Formant Modification

- The proposed approach a segmenting module, which segments the speech data into vowel and non-vowel like regions.
- The vowel like regions are first detected by using a recently reported method [1].
- After speech segmentation, Formant modification is carried out to the LP spectrum using warping.
- The pole-zero value of filter is chosen to be different for vowel and non-vowel like regions.

# Cont'd

All-pass filter D(z) to warp the LPC spectrum[2].

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

Where $\alpha$ is a warping factor in the range of $-1 < \alpha < 1$.

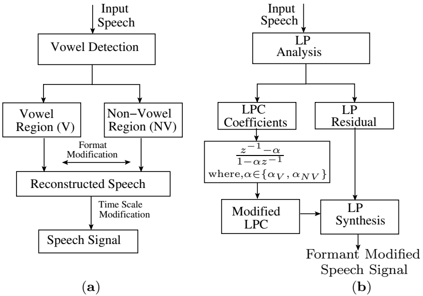
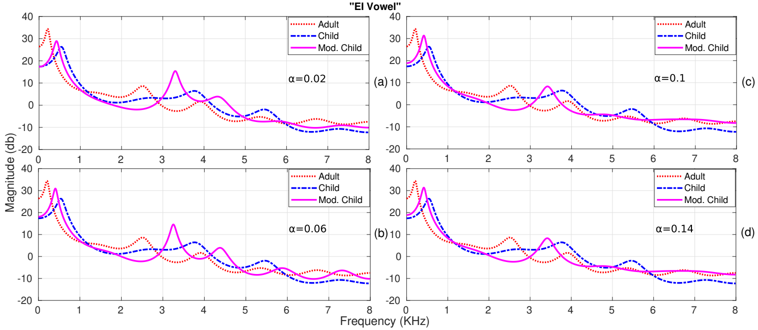


Figure 1: simplified block diagram of an ASR system

# Cont'd

Spectral smoothing of a voiced frame (vowel /EI/) of speech having fundamental frequency affected by the proposed approach is shown.



# Speech Corpora

Table 1: Speech corpora details for WSJCAM0 and PFSTAR used in ASR .

| Corpus | WSJCAM0 | | PF-STAR | |
|---|---|---|---|---|
| Language | British English | | British English | |
| Purpose | Training | Testing | Training | Testing |
| Speaker kind | Adult | Adult | Child | Child |
| No. of speakers (male & female) | 92 | 20 | 122 | 60 |
| Age group | $>$ 18 years | $>$ 18 years | 4-14 years | 4-14 years |
| No. of words | 132,778 | 5,608 | 46,974 | 5,067 |
| Duration (hrs.) | 15.5 | 0.6 | 8.3 | 1.1 |

# Experiments Setup

- The Kaldi speech recognition toolkit used to develope a children's ASR.
- The 40-channel Mel-filterbank were used to compute 13-dimensional base MFCC features.
- For normalization, cepstral feature-space maximum likelihood linear regression (fMLLR) was used.
- DNN-HMM based acoustics model was explored [3] with 8 hidden layer and 1024 hidden nodes.
- Bigram language model (LM) was used.

# Results and Discussion

We also compared with diffrent existing methods, Synchronized overlap-add fixed synthesis (SOLAFS) [4] and Real-time iterative spectrogram inversion with look-ahead (RTISI-LA). [5]

Table 2: Results on proposed method and comparison with TSM algorithms RTISILA and SOLAFS.

| Acoustic model | WER (in %) | | | |
|---|---|---|---|---|
| | | TSM | | |
| | Baseline | RTISILA | SOLAFS | SW |
| DNN | 19.76 | 16.96 | 15.00 | **14.37** |

Table 3: Effect of combining the proposed method with TSM methods.

| Acoustic model | WER (in %) | | |
|---|---|---|---|
| | SW | SW +RTISILA | SW + SOLAFS |
| DNN | 14.37 | 13.39 | 10.58 |

# Cont'd

Table 4: WERs on DNN-based ASR for children's development set. The WERs show the effects of varying $\alpha_V$ and $\alpha_{NV}$.

| $\alpha_V$ / $\alpha_{NV}$ | **WER (in %)** | | | | | |
|---|---|---|---|---|---|---|
| | 0.4 | 0.6 | 0.8 | 1.0 | 1.2 | 1.4 |
| 0.4 | 21.24 | 20.86 | 20.52 | 20.13 | 20.42 | 20.73 |
| 0.6 | 21.09 | 20.72 | 20.27 | 19.90 | 19. 76 | 20.22 |
| 0.8 | 20.66 | 20.39 | 19.89 | 18.73 | **18.53** | 18.96 |
| 1.0 | 21.03 | 20.62 | 20.14 | 19.82 | 19.66 | 20.19 |
| 1.2 | 21.37 | 21.15 | 20.77 | 20.33 | 20.18 | 22.12 |
| Baseline | 21.83 | | | | | |

# Cont'd

Table 5: Results on combined proposed method with RTISILA and SLOAFS and effect of vowel and non-vowel based parameter selection.

| Acoustic | WER (in %) | | | | | |
|---|---|---|---|---|---|---|
| model | without VNV | | | With VNV | | |
| | | SW + | SW+ | SW | SW + | SW+ |
| | SW | RTISILA | SOLAFS | SW | RTISILA | SOLAFS |
| DNN | 14.37 | 13.39 | 10.58 | 13.66 | 13.04 | 10.08 |

Table 6: Results on proposed method on pooled adults and children speech on system training. Effect of vowel and non-vowel based parameter selection.

| Acoustic | | WER (in %) | | | | | |
|---|---|---|---|---|---|---|---|
| model | | without VNV | | | With VNV | | |
| | Baseline | SW | SW + | SW+ | SW | SW + | SW+ |
| | | | RTISILA | SOLAFS | | RTISILA | SOLAFS |
| DNN | 12.26 | 11.25 | 11.14 | 8.89 | 10.86 | 10.57 | 8.51 |

# Conclusion

- The proposed method gives a relative improvement of 31% over a baseline with DNN acoustic model using MFCC acoustic features.
- The proposed + SOLAFS combined system gives a relative improvement of 49% as compared to baseline system.
- In pooled system also found improvement.

[1] A. Kumar, S Shahnawazuddin, and G. Pradhan, “Non-local estimation of speech signal for vowel onset point detection in varied environments,” *Interspeech 2017*, pp. 429–433, 2017.

[2] H. W. Strube, “Linear prediction on a warped frequency scale,” *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[3] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, “Improved feature processing for deep neural networks.,” in *Interspeech*, 2013, pp. 109–113.

[4] D. Henja and B. Musicus, “The solafs time-scale modification algorithm,” , 1991.

[5] X. Zhu, G. T. Beauregard, and L. L. Wyse, “Real-time signal estimation from modified short-time fourier transform magnitude spectra,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

*Thank you*