Kathania, Hemant; Kumar, Avinash; Kurimo, Mikko

Vowel non-vowel based spectral warping and time scale modification for improvement in children's ASR

Published in:
IEEE International Conference on Acoustics, Speech and Signal Processing

DOI:
10.1109/ICASSP39728.2021.9414116

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
GNU LGPL

Please cite the original version:
Kathania, H., Kumar, A., & Kurimo, M. (2021). Vowel non-vowel based spectral warping and time scale modification for improvement in children's ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 2021-June, pp. 6983-6987). (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414116

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

Hemant Kathania[1], *Avinash Kumar*[2], *and Mikko Kurimo*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Department of Electronics and Communication, National Institute of Technology Sikkim
hemant.kathania@aalto.fi, avinash_ece@nitsikkim.ac.in, mikko.kurimo@aalto.fi

ABSTRACT

Acoustic differences between children's and adults' speech causes the degradation in the automatic speech recognition system performance when system trained on adults' speech and tested on children's speech. The key acoustic mismatch factors are formant, speaking rate, and pitch. In this paper, we proposed a linear prediction based spectral warping method by using the knowledge of vowel and non-vowel regions in speech signals to mitigate the formant frequencies differences between child and adult speakers. The proposed method gives 31% relative improvement over the baseline system. We have also investigated time scale modification using RTISILA and SOLAFS algorithms and found that our proposed method performs better. Combining the proposed method with RTISILA and SOLAFS results in a further error rate reduction. The final combined system gives 49% relative improvement compared to the baseline system.

Index Terms— Spectral warping, vowel, non-vowel, TSM, children speech recognition.

1. INTRODUCTION

Recent years have seen significant rise in qualitative research in academy and business to use automatic speech recognition (ASR) for the learning of children's language [1, 2]. The efficiency of an ASR system is effected by the many reason such as speaker, context, and environmental variability in real-life applications. The variability of the speaker might cause a mismatch between the trained acoustic models and the recognition of the actual speech, which can lead to severe degradation of performance of recognition. It is referred to as the mismatched ASR when the ASR systems trained on the speech data from the adult speakers are used to test the speech of the children speakers. Several studies have been explored to address the acoustic mismatch in children's ASR [3, 4, 5, 6, 7].

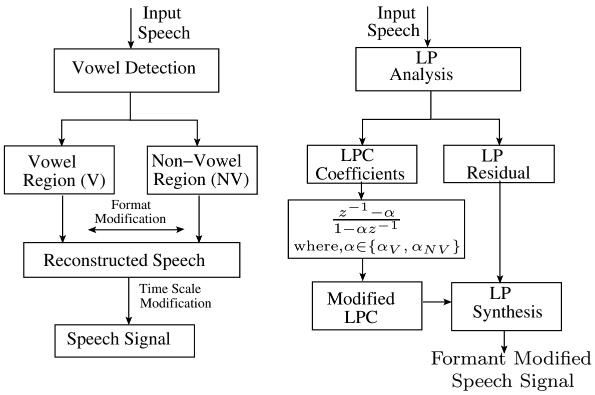
It is well known that, for adult and child speakers, the form of the vocal organs, pitch and speaking rates are substantially different. In compared to adults, major variations in the spectral characteristics of children's voices include higher fundamental and formant frequencies and spectral variability [8, 9]. Formant frequency normalization studied to transforming children speakers vowel formant frequency to adult speaker space in [9]. In fact, the formant frequencies F1, F2, and F3 were found to be highest in children and decreased with increasing age [9]. Motivated by this, we explored vowel, non-vowel region-based formant modification for children's ASR.

The speaker-dependent variations are classified into inter-speaker and the intra-speaker variabilities. The speaking-rate variation among the speakers is a common factor contributing to inter-speaker variability [9, 10]. The earlier works suggested that the average phoneme duration is longer in the case of children [9, 11]. The speaking rate of children is slower than that of adults [12]. The mean speaking rates were reported to be 2.03 syllables/sec and 1.79 syllables/sec for the speech of adults and children, respectively in [13]. The inter-speaker variabilities is the differences in the geometry of vocal organs. The Vocal-tract length variations lead to the scaling of formant frequencies [8]. A few earlier works studied formant, explicit pitch and speaking-rate adjustment for improving the children's ASR [6, 13].

In this paper, a linear prediction (LP) based spectral warping method is proposed to overcome the difference between children's and adults' speech. To optimize the warping parameter for the proposed algorithm we applied a non-local means (NLM) based vowel and non-vowel marking algorithm. Our study shows that the proposed method improved the performance compared to time-scale modification (TSM) based RTISILA and SOLAFS algorithm in recognition of children speech under mismatched condition. We combined the spectral warping with TSM and found that combined system gives further reduction in word error rate (WER).

2. PROPOSED METHODS

The proposed method block diagram is given in Figure 1 (a). The steps involved in the proposed scheme are as follows:
I) First, the vowel and non-vowel regions are identified in speech signals by using a recently reported method [14] .
II) The LP-based spectral warping is performed based on the knowledge of vowel and non-vowel regions in speech signal



(a) (b)

Fig. 1. The block diagram representing:- (a) The proposed spectral warping and speaking rate method based on vowel and non-vowel. (b) A non-uniform LPC based spectral warping.

as shown in figure 1 (b). The pole-zero value of the first-order filter in the equation 4 is selected as the vowel and non-vowel regions (α_V, α_{NV}).
III) The signal is reconstructed after spectral warping and then TSM applied to normalize the speaking rate. Finally the modified speech signal given to ASR.

2.1. Vowel non-vowel detection

In a given speech signal, vowels are the most prominent regions due to their three attributes: larger amplitude, periodicity and longer duration [15, 16, 17]. The prevailing attributes of the excitation source as well as that of the vocal-tract filter response are reflected at those instants. In various speech-based applications, accurate identification of vowels has been successfully used. Motivated by the importance of vowel non-vowel segmentation, we explored a recently reported method for vowel detection for the task of children's ASR [14]. In this method, an approximation of the speech signal at each sample is obtained using non-local means (NLM) estimation. Then the cumulative sum of the magnitude spectrum is used as the front-end feature. Next, the feature is smoothed by processing through a moving average filter over a $50ms$ window. The vowel evidence is obtained from the feature by convolving it with a first-order Gaussian differentiator (FOGD) having a window length of $100ms$ and standard deviation one-sixth of the window.

2.2. Spectral warping

The spectral structure of children's speech is modified by warping the LP spectrum. The warped LP spectrum (denoted by $S_\alpha(f)$) is obtained by modifying the original LP spectrum (denoted by $S(f)$) computed from children's speech using warping function $w_\alpha(f)$, where α is the warping factor:

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

An estimate of the present speech sample $s(n)$ is obtained in the LP analysis as a linear combination of the past P speech sample values as follows:

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

By Z-transforming Eqn. (2), the following equation is obtained

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

Where $\hat{S}(z)$, S(z) and z^{-1} represents the Z-transforms of the prediction, speech signal and unit delay filter respectively, and a_k are the LP coefficients. The unit delay filters are replaced by an all-pass filter D(z) to warp the LP spectrum. The frequency scale warping is carried out using a first order filter [18, 19] given by

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

Here α is the warping factor, whose value is in the range of $-1 < \alpha < 1$. The warped frequency scale matches the psychoacoustic frequency scale with a proper selection of α [20]. By using of the warping function D(z), the spectral resonances (formants) of the LP spectrum can be shifted systematically. The positive values of α, help the spectra to shift towards adult spectra. The warped LP coefficients $(a_k^{'}s)$ can be used with the residual (s(n) - $\hat{s}(n)$) to synthesize the speech signal [6, 21]. Using the knowledge of vowel and non-vowel regions, the formants corresponding to each frame is processed through a first-order filter having different zero and pole values for the vowel and non-vowel regions.

The effectiveness of the spectral warping method is illustrated in Figure 2 by showing LP spectra computed from "EI" vowel utterance spoken by a child and adult speaker (blue dot-dashed and red dashed lines) together with the modified LP spectra of the child's vowel (magenta solid line) computed by the proposed method. The four different zero and pole values used to derive the formant modification are = 0.2, 0.06, 0.1, and 0.14, and the correspondingly modified child spectra are shown in Figure 2 (a), (b),(c) and (d).

The figures demonstrate that the formants of the child speaker are higher compared to those of the adult speaker. Most importantly, the spectra show that the proposed LP-based warping method has moved the spectra (formant) of the child speaker to be closer to those of the adult speaker. Hence, it is expected that the features derived from the speech signals synthesised using the modified LP spectrum reduce the acoustic mismatch between adult and children speech.

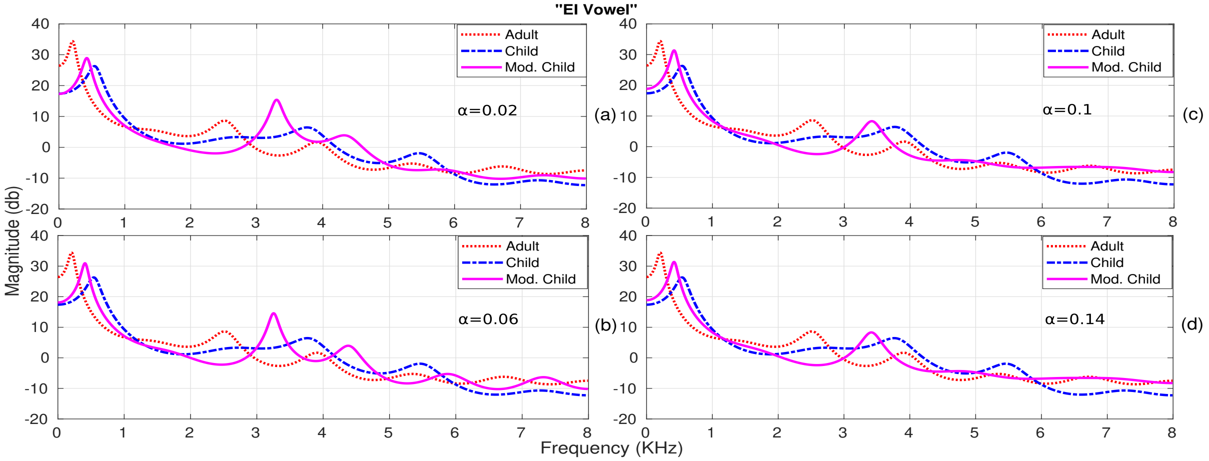


Fig. 2. LP spectra computed from frames of the vowels /EI/ showing variation in spectral frequencies. The red dahsed and blue dot-dashed curves were computed from speech utterances of an adult and child speaker, respectively. The magenta solid curves show spectra after applying the spectral warping method for the utterances of the child speaker when (a) $\alpha_V = 0.02$, (b) $\alpha_V = 0.06$, (c) $\alpha_V = 0.1$ and (d) $\alpha_V = 0.14$.

2.3. Time-scale modification

Speaking rate is a major mismatch factor between the children's and adults' speech. Because children's speech has a lower speaking rate, ASR becomes difficult specifically under mismatched condition. To resolve this issue we investigated two types of time-scale modification techniques. The real-time iterative spectrogram inversion with look-ahead (RTISI-LA) algorithm [22, 23, 24] constructs a high-quality time-domain signal from its short-time magnitude spectrum with varying parameter s from scale 0.65 to 1.85 with step size 0.10. We also investigated synchronized overlap-add fixed synthesis (SOLAFS) [25] based time scale modification. The main idea is that with fixed synthesis rate the windows are overlap-added into the output. In order to maximize the similarly between the output and the new window in the overlap region, the starting positions of the windows $x_m[n]$ are adjusted during analysis. The speaking rate modified varying parameter k varied from 0.6 to 1.8 with step size of 0.10.

3. SPEECH CORPUS AND BASELINE SYSTEM

Adult speech data used in this work was obtained from WSJ-CAM0 [26]. Children's speech data was obtained from the PF-STAR corpus [27] to simulate a mismatched ASR task. Both the WSJCAM0 and PF-STAR corpora are British English speech databases. The train set of WSJCAM0 has a total 15.5 hours of speech data with 92 adult (male and female) speakers. The total duration of children speech data for testing is 1.1 hours. The age of the child speakers in this corpus varies between 4-13 years from 60 different speakers. Development children set consist 2.5 hours of data from 62 different speakers with age range of 6-14 years. The analyses were performed using wideband speech (sampled at 16 kHz).

A Kaldi toolkit recipe was used to train the system [28]. This utilizes conventional MFCC features using 40 channel Mel-filterbank with a frame size of $25ms$ and frame shift of $10ms$ to train GMM and DNN-based acoustic models [29]. For normalization, cepstral feature-space maximum likelihood linear regression (fMLLR) was used. The fMLLR transformations for the training and test data were generated using the speaker adaptive training [30]. LDA-MLLT+SAT based GMM alignment labels were used to train DNN acoustic models. To decode the children speech test set, a domain-specific bigram language model (LM) was used. This bigram LM was trained on the transcripts of the speech data of PF-STAR excluding the test set. The baseline for DNN-based acoustic models are given in Table 1.

4. RESULT AND DISCUSSION

The performance of the baseline ASR system (i.e., the system trained with adults' speech and tested with children's speech) is reported in Table 1. From Table 1 it can be noted that the reported WER values for baseline system are quite poor. This is due to the acoustic differences between training adult and testing children data. So to alleviate the formant frequency differences and to improve the system performance, the proposed spectral warping method is applied. The spectral warping algorithm has a tunable parameter α that was varied from 0.05 to 0.25 in order to modify spectra to change the formant frequencies, and the best performance at $= 0.1$ is reported in the Table 1.

From Table 1, it can be noted that the proposed method gives a relative reduction of 30% in the WER compared to baseline system. In Table 1, we have also compared our proposed method with time scale modification based speaking rate adaptation (SRA) using RTISILA and SOLAFS algo-

Table 1. 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 2. 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

rithm and found that the proposed method outperforms these two methods. For RTISILA and SOLAFS algorithms the parameter is optimized using the development set and observed that the best value is 1.35 and 1.65 respectively.

Further to enhance the system performance we combined proposed method with TSM algorithms. The combinations are proposed +RTISILA and proposed + SOLAFS and their results are reported in Table 2. The combined system gives improvement as compared to proposed method, it seems combinations are giving complimentary information to ASR. The best combination is proposed + SOLAFS and it gives 49% relative improvement compared to baseline system.

4.1. Selection of the tunable parameters

To optimize the proposed spectral warping based formant modification algorithm parameter α, we used vowel and non-vowel region of speech signal as discussed in Section 2.1. The WERs of development set of children speech on adult data trained DNN based ASR system on α for vowel (α_V) and α for non-vowel (α_{NV}) is given in Table 3. Best combination of α is decided by lower WER as highlighted in the table. The best values for α_V is 1.2 and for α_{NV} is 0.8 chosen using the results given Table 3. The optimized vowel non-vowel based α parameter was applied to our best combination as given in Table 4. We found that vowel non-vowel based algorithm to select the α values gives further improvement in the system performance.

To further validate the effectiveness of the proposed spectral warping method, another DNN-based ASR system was trained by pooling speech data from both the adults and children train sets. Such an ASR system reduces the degree of acoustic and linguistic mismatch by utilizing also children speech in the training. The baseline WER of the pooled system is given in Table 5. From Table 5, it can be seen that the proposed method and the combinations with the other techniques also reduce WER in the pooled system. A relative reduction of 30% in WER is noted compared to the baseline.

Table 3. WERs on DNN-based ASR for children's development set. The WERs show the effects of varying α_V and α_{NV}.

α_V α_{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					

Table 4. Results on combined proposed method with RTISILA and SLOAFS and effect of vowel and non-vowel based parameter selection.

Acoustic model	WER (in %)					
	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 5. Results on proposed method on pooled adults and children speech on system training. Effect of vowel and non-vowel based parameter selection.

Acoustic model		WER (in %)					
		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

5. CONCLUSION

In this paper, we have proposed and studied a spectral warping method based on vowel and non-vowel region to demonstrate its effectiveness in the context of children speech recognition using acoustic models trained on adults speech. The proposed method gives a relative improvement of 31% over a baseline with DNN acoustic model using MFCC acoustic features. We have also compared the proposed method with TSM algorithms RTISILA and SOLAFS and found that the proposed method performs better. By combining the proposed method with RTISILA and SOLAFS, showed a further reduction in WER. Proposed + SOLAFS combined system gives a relative improvement of 49% as compared to baseline system. A pooled system is also developed by pooling together speech data from both adult and children speakers and even in this case the proposed system manages to improve the performance.

6. REFERENCES

[1] Victor Zue, Stephanie Seneff, Joseph Polifroni, Helen Meng, and James Glass, "Multilingual human-computer interactions: From information access to language learning," in *Poc. ICSLP*, 1996, vol. 4, pp. 2207–2210.

[2] Martin Russell, Catherine Brown, Adrian Skilling, Rob Series, Julie Wallace, Bill Bonham, and Paul Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Proc. ICSLP*, 1996, vol. 1, pp. 176–179.

[3] Harald Singer and Shigeki Sagayama, "Pitch dependent phone modelling for hmm based speech recognition," in *ICASSP*, 1992, vol. 1, pp. 273–276.

[4] Xu Shao and Ben Milner, "Pitch prediction from mfcc vectors for speech reconstruction," in *Proc. ICASSP*, 2004, vol. 1, pp. I–97.

[5] Daniel C Burnett and Mark Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, 1996, vol. 2, pp. 1145–1148.

[6] Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, and Mikko Kurimo, "Study of formant modification for children asr," in *Proc. ICASSP*, 2020, pp. 7429–7433.

[7] Ishwar Chandra Yadav, Avinash Kumar, Syed Shahnawazuddin, and Gayadhar Pradhan, "Non-uniform spectral smoothing for robust children's speech recognition.," in *Interspeech*, 2018, pp. 1601–1605.

[8] Raymond D Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of speech and hearing Research*, vol. 19, no. 3, pp. 421–447, 1976.

[9] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[10] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[11] Martin Russell and Shona D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[12] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos, "A review of asr technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009, pp. 1–8.

[13] Shweta Ghai, "Addressing pitch mismatch for children's automatic speech recognition," *Unpublished Ph. D. thesis, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India*, 2011.

[14] Avinash Kumar, S Shahnawazuddin, and Gayadhar Pradhan, "Non-local estimation of speech signal for vowel onset point detection in varied environments," *Interspeech 2017*, pp. 429–433, August 2017.

[15] K. N. Stevens, *Acoustic Phonetics*, The MIT Press Cambridge, Massachusetts, London, England, 2000.

[16] Gayadhar Pradhan, Avinash Kumar, and Syed Shahnawazuddin, "Excitation source features for improving the detection of vowel onset and offset points in a speech sequence.," in *INTERSPEECH*, August 2017, pp. 1884–1888.

[17] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.

[18] Hans Werner Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[19] Unto K Laine, Matti Karjalainen, and Toomas Altosaar, "Warped linear prediction (wlp) in speech and audio processing," in *Proc. of ICASSP*, 1994, vol. 3, pp. III–349.

[20] Julius O Smith and Jonathan S Abel, "Bark and erb bilinear transforms," *IEEE Transactions on speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[21] John Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[22] Hemant Kathania, Mittul Singh, Tamás Grósz, and Mikko Kurimo, "Data augmentation using prosody and false starts to recognize non-native children's speech," in *Proc. INTERSPEECH 2020*, 2020, p. To appear.

[23] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. Tarun Sai, "Creating speaker independent asr system through prosody modification based data augmentation," *Pattern Recognition Letters*, vol. 131, pp. 213 – 218, 2020.

[24] 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.

[25] D. Henja and B. Musicus, "The solafs time-scale modification algorithm," 1991.

[26] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, May 1995, vol. 1, pp. 81–84.

[27] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.

[28] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.

[29] George Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[30] Shakti P Rath, Daniel Povey, Karel Veselỳ, and Jan Cernockỳ, "Improved feature processing for deep neural networks.," in *Interspeech*, 2013, pp. 109–113.