INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving Children's Speech Recognition through Explicit Pitch Scaling based on Iterative Spectrogram Inversion

W. Ahmad[1], *S. Shahnawazuddin*[2], *H. K. Kathania*[1], *G. Pradhan*[2] *and A. B. Samaddar*[3]

[1] Department of Electronics and Communication Engineering, NIT Sikkim, India.
[2] Department of Electronics and Communication Engineering, NIT Patna, India.
[3] Department of Computer Science and Engineering, NIT Sikkim, India

(waquar, hemant.ece)@nitsikkim.ac.in, (s.syed, gdp)@nitp.ac.in, absamaddar@yahoo.com

Abstract

The task of transcribing children's speech using statistical models trained on adults' speech is very challenging. Large mismatch in the acoustic and linguistic attributes of the training and test data is reported to degrade the performance. In such speech recognition tasks, the differences in pitch (or fundamental frequency) between the two groups of speakers is one among several mismatch factors. To overcome the pitch mismatch, an existing pitch scaling technique based on iterative spectrogram inversion is explored in this work. Explicit pitch scaling is found to improve the recognition of children's speech under mismatched setup. In addition to that, we have also studied the effect of discarding the phase information during spectrum reconstruction. This is motivated by the fact that the dominant acoustic feature extraction techniques make use of the magnitude spectrum only. On evaluating the effectiveness under mismatched testing scenario, the existing as well as the modified pitch scaling techniques result in very similar recognition performances. Furthermore, we have explored the role of pitch scaling on another speech recognition system which is trained on speech data from both adult and child speakers. Pitch scaling is noted to be effective for children's speech recognition in this case as well.

Index Terms: children's speech recognition, pitch mismatch, pitch scaling.

1. Introduction

It is well known that the performance of an automatic speech recognition (ASR) system is affected by speaker, context and environment variability. Any mismatch in the acoustic or linguistic attributes captured by the training and test data degrades the recognition performance. An extreme example of such a mismatched ASR is the task of transcribing children's speech on acoustic models trained using speech data from adult speakers. Even on pooling data from both adult and child speakers, the observed recognition performance for the children's speech is noted to be poorer than that for the adults' matched ASR task. The observed degradations are mainly due to large differences in both the acoustic and linguistic correlates between the speech from the adult and child speakers [1, 2, 3, 4, 5, 6]. Several studies for addressing the acoustic mismatch in the context of children's ASR have been reported [7, 8, 9]. Recently, a number of works have also explored acoustic modeling based on deep neural network (DNN) [10] for improving children's speech recognition [11, 12, 13, 14, 15, 16].

Among the various mismatch factors reported in literature, the differences in the pitch values for adults' and children's speech is a major one. In this paper, we have explored the role of explicitly scaling the pitch of the children's speech so that it matches with that for the adults' speakers. Scaling the pitch value is expected to reduce the pitch-induced acoustic mismatch. In this regard, we have explored an existing pitch scaling technique that iteratively estimates the pitch modified signal from its short-time Fourier transform magnitude spectrum. Explicitly scaling the pitch of the children's speech is found to be highly effective in the context of *children's mismatched ASR*. In this study, the task of recognizing children's speech using acoustic models trained on speech data from adult speakers is referred to as the children's mismatched ASR. Further, we have also studied the scope of discarding the phase information during the signal re-estimation. This is achieved by using discrete cosine transform (DCT) in place of Fourier transform. During our experimental exploration, it was noted that discarding the phase information does not degrade the performance of the mismatched ASR. In this paper, the effectiveness of explicit pitch modification is studied on ASR systems developed using three dominant acoustic modeling techniques including DNN. Furthermore, the DCT-based pitch scaling method is also compared with some of the existing techniques.

The remaining of this paper is organized as follows: In Section 2, the explored method for explicit pitch scaling is presented. In the next Section 3, the experimental evaluations are presented. Finally, the paper is concluded in Section 4.

2. Phase independent pitch scaling

As mentioned earlier, one of the major acoustic mismatch factor in the context of children's mismatched ASR is the differences in the pitch values for adults' and children's speech. Earlier works have reported that the range of pitch for children's speech is from 200 Hz to 350 Hz while that for the adults' lies predominantly in between 100 Hz to 200 Hz. The primary objective of the presented study is to modify the pitch of children's speech so that it matches with that of the adults' speech. As a result, the pitch-induced acoustic mismatch between the training and test will be normalized to a large extent. This, in turn, will improve the recognition of children's speech using acoustic models trained on speech data from adult speakers.

The basic premise of the pitch modification is that stretching or compressing the analog waveform of an audio signal changes the pitch of the signal. On the other hand, in the case of digital audio signals, pitch scaling can be achieved by resampling the given signal. If an audio signal sampled at a frequency f_i is re-sampled such that the new sampling frequency is f_o, on playing back the re-sampled signal at the original sam-

http://dx.doi.org/10.21437/Interspeech.2017-302

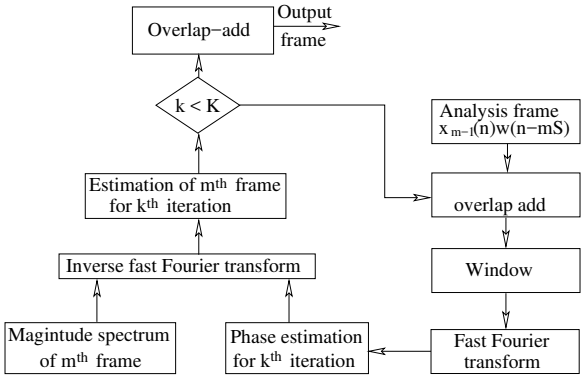

Figure 1: *Block digram depicting the iterative estimation process RTISI-LA for deriving the estimate of a speech signal from its short-time magnitude spectrum.*

pling frequency f_i, a change in the pitch of the signal by a factor of $\frac{\log(f_i/f_o)}{\log(\sqrt[12]{2})}$ semitones is noted [17]. Semitone is the concept in music theory where 12 semitones indicate an octave. At the same time, an undesired change in the length of the audio signal by a factor of f_o/f_i is also observed. This undesired change in the length needs to be optimally compensated. The compensation of length is achieved by a process generally referred to as time-scale modification (TSM). Several algorithms for TSM are available and one of the reported technique which is explored in this study is discussed next.

The approach for TSM explored in this paper is based on the real-time iterative spectrogram inversion with look-ahead (RTISI-LA) algorithm [18, 19]. In order to precisely change the pitch value while effectively retaining other relevant information, the re-sampled signal is reconstructed from its short-time Fourier transform magnitude. Consider an analysis frame of length $\mathcal{L}$ and let the pitch be scaled downwards by a factor of q ($0 < q < 1$). Using a block of $\mathcal{L}' = q\mathcal{L}$ samples per frame, re-sampling is performed in time-domain to get back a frame of length $\mathcal{L}$. Next, short-time Fourier transform magnitude (STFTM) of the obtained frame of speech is computed. The STFTM helps in understanding how an audio signal is perceived in term of its frequency components by combining the imaginary and real part into a single number. This aspect is exploited by the RTISI-LA for reconstructing an audio signal from its STFTM through an iterative process.

The frame by frame signal re-estimation process using the RTISI-LA is depicted in the Figure 1. Let us assume that the first $m-1$ frames of the given audio signal $x(n)$ have already been reconstructed from their respective STFTM while the m^{th} frame is to be synthesized next. To achieve this, a partial analysis frame is created by overlap-adding the estimated results for the frame $m-1$, $m-2$ and $m-3$ of $x(n)$ considering an overlap of 75%. The fourth quarter of this partially filled frame is all zeros. Further, let the partial frame be denoted by $x_{m-1}(n)w(n-mS)$ while the reconstructed frame be $x_m(n)$ where S is step size between adjacent frames. In order to ensure 75% overlap, the value of window length is chosen to be four times that of the step size, i.e., $\mathcal{L} = 4S$. Next, the Fourier transform of the partial frame is calculated after employing a scaled Hamming window. The phase information estimated from the Fourier transform of the partial analysis frame is then combined with the given STFTM for the m^{th} frame. A newer estimate for the m^{th} frame is then obtained by the inverse Fourier transform of the so derived frequency-domain signal.

In this work, we have slightly modified the RTISI-LA algorithm. Instead of taking the Fourier transform, discrete cosine transform (DCT) [20] is used. Discrete cosine transform is extensively used in many signal processing applications [21, 22]. The DCT provides a very compact representation of the data using real coefficients only [20, 23]. For any signal $x(n)$ which is a sequence N coefficients, the DCT is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right]; \qquad (1)$$

where, $k, n = 0, 1, 2, \ldots, N-1$. Similarly, the inverse discrete cosine transform (IDCT) is defined as:

$$x(n) = \frac{2}{N}\sum_{k=0}^{N-1} X(k) \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right]. \qquad (2)$$

Using DCT-IDCT pair, the phase information gets discarded (effectively, the phase is equal to 0 or π). This is motivated by the fact that, in the case of ASR, only the magnitude spectrum is used for computing front-end acoustic features. This phase independent pitch scaling approach is referred to as *DCT-based RTISI-LA* in the remaining of the paper. Interestingly, ignoring the phase did not bring about any perceivable difference in the reconstructed signal when compared to that obtained through RTISI-LA. Furthermore, discarding the phase information during pitch scaling does not lead to significant change in the recognition performance as evident from the experimental studies presented in the following section.

3. Experimental setup and evaluation

In this section, we first describe the datasets used in the experimental evaluations. Next, the experimental results demonstrating the effectiveness of pitch scaling in context of children's mismatched ASR are discussed.

3.1. Adults' and children's speech databases

For the continuous speech recognition task, two different British English speech corpora were employed viz. the WSJCAM0 adults' speech corpus [24] and the PF-STAR children's speech corpus [25]. The WSJ0CAM0 database was split into orthogonal train and test sets (none of the speakers were common in the training and test sets). The train set (TR-AD) comprised of 15.5 hours of speech data from 92 male/female speakers above the age of 18 years. The total number of utterances in the train set was $7,852$ with a total $132,778$ words. The adults' speech test (TS-AD) set consisted of 0.6 hours of speech from 20 speakers with a total of $5,608$ words. In the case of PF-STAR, the age of the child speakers in this corpus lies between 3-14 years. This database was also split into orthogonal train and test sets. The train set (TR-CH) consisted of 8.3 hours of speech data from 122 children. The total number of utterances was equal to 856 with a total of $46,974$ words. A subset of 313 utterances (63 speakers) from TR-CH was used as the development data (DEV-CH) as and when required. The children's speech test set (TS-CH) consisted of 1.1 hours of speech data from 60 children with a total of 5067 words. All the experimental studies reported in this paper are performed on wideband speech data (sampled at 16 kHz rate). Since the PF-STAR database is originally sampled at $22,050$ samples per second, down-sampling was performed for consistency.

3.2. ASR system specifications

In order to evaluate the efficacy of explicit pitch scaling, an ASR system was developed on adults' speech train set (TR-AD) using the Kaldi speech recognition toolkit [26]. For front-end acoustic feature extraction, the speech data was analyzed using overlapping Hamming windows of length 20 ms with 50% overlap. The 13-dimensional base Mel-frequency cepstral coefficients (MFCC) were computed using a 40 channel Mel-filterbank. Next, time-splicing of the base features was done considering a context size of 9. The dimensionality of the resulting time-spliced features was then reduced to 40 using linear discriminant analysis and further de-correlation was done through maximum likelihood linear transform. This was followed by cepstral mean and variance normalization (CMVN). Feature normalization using feature-space maximum-likelihood linear regression (fMLLR) was performed next employing speaker adaptive training [27].

For developing the ASR system, context-dependent hidden Markov model (HMM) was employed. At the same time, three prominent approaches namely Gaussian mixture model (GMM), subspace GMM (GMM) and DNN were explored for generating the observation probabilities for the HMM states. In the GMM-HMM system, cross-word modeling with decision-tree-based state tying was employed and the maximum number of tied-states (senones) was fixed 2000. Each triphone model consisted of a 3-states HMM with 8 diagonal covariance Gaussian components per state. In order to develop the SGMM-HMM-based system, the number of Gaussian in the universal background model was chosen to be 400. The number of leaves and Gaussians in the SGMM were selected to be 9000 and 7000, respectively. For the DNN-HMM-based system, 8 hidden layers with 1024 nodes per layer having *tanh* nonlinearity were employed. An initial learning rate of 0.015 was selected. This was then reduced to 0.002 in 20 epochs. Extra 10 epochs were employed after reducing the learning rate. The minibatch size for training was selected as 512. The fMMLR-normalized time-spliced features were further spliced in time considering a context size of 9 prior to training the DNN.

A domain-specific 1.5k bigram LM trained on the transcripts TS-CH was employed while decoding the children's test set TS-CH. This LM has an OOV rate of 1.20% and perplexity of 95.8 with respect to TS-CH. A lexicon of 1,969 words including the pronunciation variations was employed while decoding the children's test set. For evaluating the matched case recognition performances, the TS-AD test set was used. The MIT-Lincoln 5k Wall Street Journal bi-gram LM which has a perplexity of 95.3 with respect to TS-AD was used. The lexicon employed in the matched case consisted of 5,850 words including the pronunciation variations. The metric used in this study to evaluate the recognition performance of the developed ASR systems is the word error rate (WER).

3.3. Results and discussions

The baseline WERs for TS-CH with respect to the GMM-, SGMM- and DNN-based ASR systems are given in Table 1. Compared to GMM-HMM systems, significant reduction in WERs are noted when SGMM/DNN-based acoustic modeling is used. It is to note that, the WER for the TS-AD set with respect to the DNN-HMM system is 6%. Thus, despite the use of CMVN and fMMLR, the severe pitch mismatch leads to extremely degraded recognition performances for children's test set when compared to adults' matched testing. In order to improve the performance, explicit pitch scaling was performed. The MATLAB implementation for RTISI-LA algorithm has tunable pitch compensation factor (semitone) that can be varied from -12 to 12 (in steps of 1) in order to modify the pitch. Since, the explored approach is minor modification over the RTISI-LA algorithm, semitone is used as the compensation in this case as well. The WERs with varying pitch compensation factor are also enlisted in Table 1. Its evident that pitch scaling results in huge reductions in WERs. Moreover, positive values for the compensation factor does not help which is quite expected. The effect of explicitly modifying the pitch is demonstrated using the histograms shown in Figure 2 and Figure 3. For this analysis, 50000 short-time frames of speech were collected from TR-AD and TR-CH sets and pitch was computed. Only those frames for which the value for pitch was non-zero and finite were considered. For the unscaled case, i.e., Figure 2, more frames lie in the higher pitch region for children's speech when compared to the adults. As argued earlier, on downward scaling, the mismatch in pitch gets reduced as visible from Figure 3.

Table 1: *The WERs for the children's test set with respect to different acoustic models (trained on adults' speech) with variations in the pitch compensation factors for the DCT-based pitch scaling method.*

Compensation factor	WER (in %)		
	GMM	SGMM	DNN
Baseline	32.69	24.67	19.68
-2	22.30	18.23	13.95
-3	20.41	**17.11**	13.49
-4	19.85	17.25	**13.00**
-5	**19.21**	17.13	13.88
-6	21.09	17.81	14.59

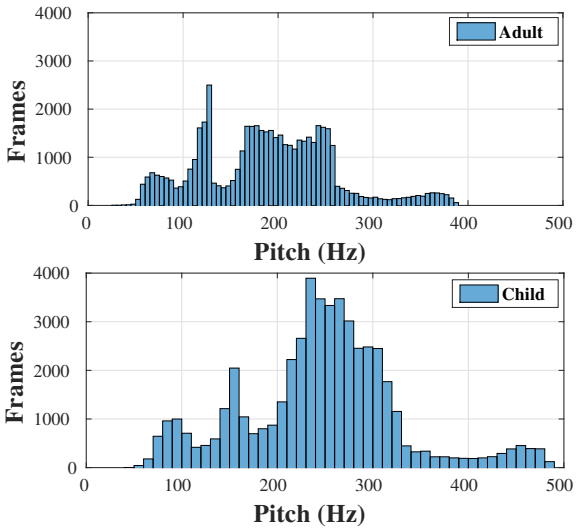

Figure 2: *Histogram depicting the pitch variations for adults' and children's speech (computed using* 50,000 *frames).*

3.4. Determining the optimal value of the pitch compensation factor through maximum likelihood grid search

A speech recognition system can be accessed by adult as well as children speakers of both the genders. Hence, scaling down

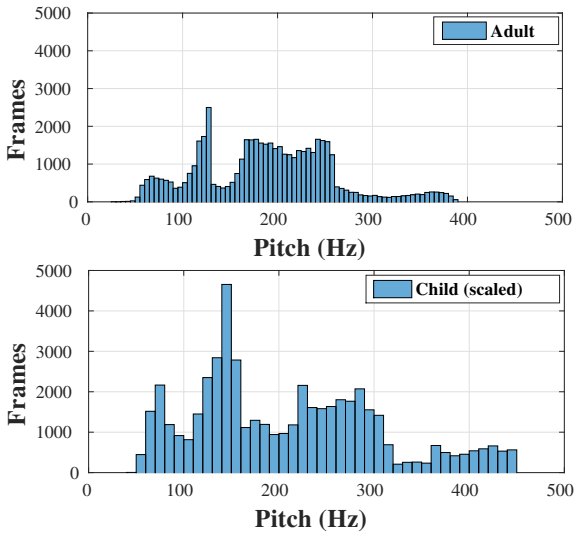

Figure 3: *Histogram depicting the pitch variations after downward scaling in the case of children's speech (computed using* $50,000$ *frames).*

Table 2: *WERs for TS-CH set with respect to DNN-HMM-based ASR system trained on adults' speech employing ML grid search for automatically selecting pitch compensation factor. WERs are also given with respect to an ASR system trained on mix of speech data from adult and child speakers.*

Speech data used for acoustic modeling	**WER (in %)**		
	Baseline	**RTISI-LA**	**DCT-based RTISI-LA**
TR-AD	19.76	12.78	13.11
TR-AD + TR-CH	11.47	9.83	9.93

the pitch of the given data is not the only possibility. At times, the optimal recognition performance may be obtained by upward pitch scaling. To address this issue, the developed system should be able to determine the optimal value of the pitch compensation factor on its own. In order to automatically determine the optimal pitch compensation factor for a given test utterance, a maximum-likelihood (ML) grid search was used. In this case, the given test utterance (without any change in pitch) was first decoded using the developed ASR system to derive the first-pass transcription. Next, the pitch was changed using different pitch compensation factor. The range of values for the pitch compensation factor was varied from -6 to 6 in steps of 1. For each case, the MFCC were computed for the pitch modified data. The feature vectors were then forced-aligned with respect to the acoustic models under the constraints of the first-pass transcription. The pitch compensation factor that resulted in the highest likelihood was chosen to be optimal and corresponding pitch modified utterance was re-decoded.

The WER for TS-CH set with respect to DNN-HMM-based ASR system trained on adults' speech employing ML grid search for automatically selecting pitch compensation factor is given in Table 2. For better contrast, WER for the case when the original RTISI-LA algorithm is used is also enlisted. From

Table 3: *Comparison of the DCT-based RTISI-LA approach with other existing pitch scaling techniques in terms of WERs for children's mismatched testing.*

Speech data used for acoustic modeling	**WER (in %)**			
	Baseline	**HPSS**	**Phase Vocoder**	**DCT-based RTISI-LA**
TR-AD	19.68	15.36	16.42	**13.11**
TR-AD + TR-CH	11.47	10.46	11.03	**9.93**

the results given in Table 1 and Table 2, it is evident that explicit pitch scaling is very effective in the context of children's speech recognition. Furthermore, employing DCT does not degrade the performance in statistical sense. To further build our confidence, another DNN-HMM-based ASR was developed using speech data pooled from both adult and child speakers (TR-AD + TR-CH). Pooling children's speech into training resulted in huge reduction in baseline WER as given in Table 2. This is primarily due to a lower degree of acoustic/linguistic mismatch between the training and test data. Despite that, explicit pitch scaling is found to be very effective and the same can be noted from the WERs given in Table 2. Even among the children themselves, the pitch variation with age of the speaker is more diverse than that for the case of adult speakers. These differences lead to a certain degree of pitch-induced acoustic mismatch and hence explicit pitch scaling helps.

3.5. Comparison with other pitch scaling techniques

From the experimental results presented in the earlier subsections, the modified DCT-based RTISI-LA technique was noted to be as effective as the original one. In this sub-section we compare the DCT-based RTISI-LA technique with two more pitch scaling approaches. The techniques considered for comparison are the one based on harmonic percussive source separation (HPSS) [28] and the phase vocoder method [29]. The earlier discussed ML grid search is employed for optimally selecting the compensation factor for those methods as well. The WERs for the TS-CH test set with respect to the DNN-HMM-based ASR system are given in Table 3. It is to note that both HPSS and phase vocoder result in huge reductions in the WERs. Moreover, the DCT-based pitch scaling method is noted to give the best recognition performance.

4. Conclusion

In this paper, the role of explicitly scaling the pitch of children's speech is studied in the context of children's mismatched ASR. In this regard, an existing pitch scaling technique based on iterative spectrogram inversion (RTISI-LA) is effectively used for reducing the pitch-dependent acoustic mismatch. Furthermore, we have also explored the effect of discarding the phase information during spectrum reconstruction. To achieve the same, DCT has been used in place of Fourier transform in the original RTISI-LA algorithm. Discarding the phase does not produce any unwanted changes in the reconstructed signal. Furthermore, statistically insignificant differences in WERs are noted on discarding the phase information. The phase independent pitch scaling technique is also compared with other approaches for pitch modification and is found to outperform those methods.

5. References

[1] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.

[2] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[3] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, February 2002.

[4] A. Potaminaos and S. Narayanan, "Robust Recognition of Children Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, November 2003.

[5] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[6] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proc. Workshop on Child Computer Interaction*, September 2014.

[7] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition." in *Proc. INTERSPEECH*, 2009, pp. 568–571.

[8] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. INTERSPEECH*, 2009, pp. 1607–1610.

[9] S. Shahnawazuddin and R. Sinha, "Low-memory fast on-line adaptation for acoustically mismatched children's speech recognition," in *Proc. INTERSPEECH*, 2015.

[10] G. Dahl, D. Yu, L. Deng, and A. 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.

[11] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *Proc. Spoken Language Technology Workshop (SLT)*, December 2014, pp. 135–140.

[12] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Proc. INTERSPEECH*, 2014, pp. 1468–1472.

[13] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q. Jiang, T. N. Sainath, A. W. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, 2015, pp. 1611–1615.

[14] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. 1, 2016.

[15] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. INTERSPEECH*, 2016.

[16] S. Shahnawazuddin, K. T. Deepak, G. Pradhan, and R. Sinha, "Enhancing noise and pitch robustness of children's ASR," in *Proc. ICASSP*, 2017.

[17] J. Driedger and M. Müller, "Tsm toolbox: Matlab implementations of time-scale modification algorithms," in *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, 2014, pp. 249–256.

[18] 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.

[19] G. T. Beauregard, X. Zhu, and L. Wyse, "An efficient algorithm for real-time spectrogram inversion," in *Proceedings of the 8th International Conference on Digital Audio Effects*, 2005, pp. 116–118.

[20] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[21] K. R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[22] Y. A. V. Phamila and R. Amutha, "Discrete cosine transform based fusion of multi-focus images for visual sensor networks," *Signal Processing*, vol. 95, pp. 161–170, 2014.

[23] G. Strang, "The discrete cosine transform," *SIAM review*, vol. 41, no. 1, pp. 135–147, 1999.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.

[25] 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.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.

[27] S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.

[28] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[29] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999, pp. 91–94.