INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Pitch-Adaptive Front-end Features for Robust Children's ASR

S Shahnawazuddin, Abhishek Dey and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, India.
{s.syed,abhishekdey,rsinha}@iitg.ernet.in

Abstract

In the presented work, we explore some of the challenges in recognizing children's speech on automatic speech recognition (ASR) systems developed using adults' speech. In such mismatched ASR tasks, a severely degraded recognition performance is observed due to the gross mismatch in the acoustic attributes between those two groups of speakers. Among the various sources of mismatch, we focus on the large differences in the average pitch values across the adult and child speakers in this work. Earlier studies have shown that the Mel-filterbank employed in the feature extraction is not able to smooth out the pitch harmonics sufficiently in particularly for the high-pitched child speakers. As a result of that, the acoustic features derived for the adult and the child speakers turn out to be significantly mismatched. For addressing this problem, we propose a simple technique based on adaptive-liftering for deriving the pitch-robust features. This enables us to reduce the sensitivity of the acoustic features to the gross variations in pitch across the speakers. The proposed features are found to result in improved performance in the context of deep neural network based ASR system. Further with the use of the existing feature normalization techniques, additional gains are noted.

Index Terms: Children's speech recognition, pitch-adaptive features, DNN.

1. Introduction

The automatic speech recognition (ASR) in the case of the adult speakers has witnessed tremendous improvements over the last few decades. On the other hand, limited efforts have been made towards improving the children's ASR. Only a few works on the children's ASR employing the acoustic modeling based on the deep neural network (DNN) [1] have been reported [2, 3, 4]. The automatic recognition of children's speech is comparatively a much tougher task due to the large differences in both the acoustic and the linguistic correlates between the speech from the adult and child speakers [5, 6, 7, 8, 9]. The ability of a child to produce varying speech sounds properly and accurately improves with the age during the growing phase [10]. Furthermore, children's have smaller vocal organs compared to the adults. Consequently, the speech from the child speakers has a higher fundamental and formant frequencies and greater spectral variability. In addition to that, the overall speaking rate is slower and highly variable in the case of children [7]. The children are reported to have greater values of the mean and the variance for the acoustic correlates of speech than those for the adults. Consequently, the children's speech suffers from a higher degree of inter- and intra-speaker acoustic variability than the adults' speech [7, 11]. From the linguistic perspective, the children are more likely to use *imaginative words*, *ungrammatical phrases* and *incorrect pronunciations* [12].

With the progress made in the speech processing research, several user-specific applications based on speech recognition have been developed that involve human-machine interactions. Some examples of such applications are speech-based information retrieval, speech-based web search and entertainment [13, 12]. In such tasks, the employed ASR system is assessed by both the adults and the children. In general, the ASR systems developed for the adult speakers show a degraded recognition performance when used for transcribing the children's speech. Similar degradation is observed when the ASR systems trained on the children's speech are employed for transcribing the adults' speech data. One way to overcome this problem is to pool a large amount of data from the speakers of all age groups while learning the ASR system parameters [4]. Unfortunately, there is a scarcity of publicly available speech corpus from the child speakers. Alternatively, one can explore ways of improving the recognition of the children's speech on acoustic models trained using the adults' speech. In this paper, the latter approach is explored and is referred to as the *children's mismatched ASR*. The task of recognizing the adults' and the children's speech on the ASR systems trained using their respective domain data is referred to as the *matched ASR*.

As mentioned earlier, the child speakers exhibit higher fundamental frequencies in comparison to the adult speakers. This affects the front-end speech parameterization process resulting in severe pitch-dependent distortions. To address the pitch-induced distortions, we explore pitch-adaptive signal processing for the front-end speech parameterization in this paper. The proposed *adaptive-liftering*-based spectral smoothing approach is observed to enhance the pitch robustness of the acoustic features. The adaptive-liftering-based features are evaluated on a DNN-based ASR system. Furthermore, we have also explored the existing dominant feature-space normalization approaches in the context of the proposed pitch-robust features.

The remaining of this paper is organized as follows: In Section 2, the proposed pitch-adaptive feature extraction approach is discussed. The experimental evaluation of the explored approaches is presented in Section 3. Finally the paper is concluded in Section 4.

2. Pitch-adaptive front-end features

2.1. Motivation

The front-end speech parameterization in the ASR systems involves the short-time analysis of the speech signal. Generally, overlapping Hamming/Hanning windows are used for the analysis of the speech signal. For each frame, the short-time Fourier transform (STFT) is computed and is followed by the Mel-scale warping of the magnitude spectrum. This involves a bank of triangular filters having nonuniform bandwidth with

http://dx.doi.org/10.21437/Interspeech.2016-1020

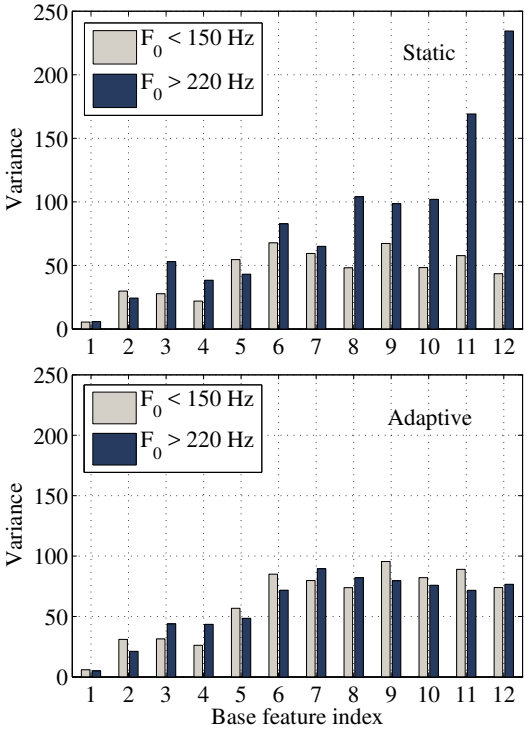

Figure 1: *Variance plots for the base MFCC features (C_1-C_{12}) for the vowel /IY/ corresponding to two broad pitch (F_0) ranges. For the higher F_0 range, the mismatch in the variances of higher-order coefficients is evident for the static MFCCs. A significant reduction in the variance mismatch for those coefficients achieved with the proposed approach can be noted.*

the center frequencies lying on the Mel-scale. By taking the discrete cosine transform (DCT) of the log-energies at the output of the Mel-filterbank, the cepstral representation is obtained. The Mel-frequency cepstral coefficient (MFCC) [14] features are then extracted by the low-time liftering of the resulting cepstra. In this work, the MFCCs extracted following the outlined procedure are referred to as the *default* or the *static* features.

In general, the MFCC features are expected to be free from the effect of the pitch of the speech signal. In [15], it is shown that the MFCCs do get affected for the child (higher pitch) speakers in contrast to that of the adult (lower pitch) speakers. Due to the insufficient smoothing of the pitch harmonics present in the magnitude spectrum of the windowed speech signal, ripples appear in the lower frequency region of the spectrum. This, in turn, leads to an increase in the dynamic range of the higher-order MFCCs in the case of children's speech [16]. To visualize that, the base MFFC feature vectors for nearly 2000 speech frames corresponding to the central portion of the vowel /IY/ extracted from the TIMIT database [17] were collected. The feature vectors were then grouped into two broad pitch (F_0) ranges, i.e., $F_0 < 150$ Hz and $F_0 > 220$ Hz. Next, the variance was computed using all the feature vectors belonging to a particular group. The variance for the two F_0 ranges in the case of static MFCCs is shown in Figure 1 (top pane). An increase in the variance of the higher-order coefficients (C_{10}-C_{12}) is evident from the plots. Since the acoustic modeling approaches employed in the ASR are data dependent, the mismatch in the variances leads to a degradation in the performance.

2.2. Pitch compensation through adaptive-liftering

In order to address the mismatch in the variances, we have explored an adaptive-liftering-based approach for smoothing the spectra. The steps in the proposed scheme are as follows: Using the STFT analysis with a fixed duration Hamming window, the spectral representation of the speech signal is obtained. For each frame, the log-compressed magnitude spectrum is derived. The cepstral representation is then obtained through the inverse discrete Fourier transform (IDFT) of the magnitude spectrum. The steps involved this far are essentially equivalent to the linear filtering. Consequently, the cepstral domain representation retains the periodicity of the speech excitation. For smoothing out the pitch harmonics, a suitable low-time lifter is applied. The liftered cepstrum is then transformed back to the spectral domain using the DFT. Given the smoothed spectrum, the two kinds of front-end features explored in this work are derived following the usual steps. The block diagram for deriving the smoothed spectrum and the corresponding pitch-robust acoustic features is shown in Figure 2.

For determining the duration of the applied low-time lifter L, the average pitch value F for the utterance being analyzed is computed, such that $L = F_s/F$ where F_s is the sampling frequency. A cepstral-domain-based pitch detection algorithm, outlined in [18], is used for the estimation of the pitch in this work. We also performed the pitch estimation using a few other algorithms, viz. TEMPO [19], RAPT [20] and WaveSurfer [21] for checking the consistency. Even though some differences in the frame-specific pitch estimates do exist among these algorithms, the average pitch values turned out to be quite similar in all the cases. Further, to avoid ripples in the derived smoothed spectrum, a liftering window with slanting right-edge is used.

The log-compressed magnitude spectra obtained by the conventional approach employing static signal processing for two different vowels are shown in Figure 3. The degree of spectral smoothening achieved through the proposed approach with variations in the length of the applied lifter window is also shown in Figure 3. The effectiveness of the pitch-adaptive signal processing approach in addressing the mismatch in the variances of the MFCCs is demonstrated in Figure 1 (bottom pane). Compared to the case of the static MFCCs, the mismatches have been considerably reduced. The reduction in the mismatch are expected to result in improvements in the case of the children's mismatched ASR.

3. Experimental evaluation

For the all experimental evaluations presented in this paper, an speaker independent (SI) ASR system is developed using the Kaldi speech recognition toolkit [22] employing the DNN-based acoustic modeling. In the DNN architecture, there are 8 hidden layers with 1024 units employing *tanh* nonlinearities. A soft-max layer representing the log-posterior of the output labels corresponding to the context-dependent hidden Markov model (HMM) states is used as the output layer. An initial learning rate of 0.015 is selected. The learning rate is reduced to 0.002 in 20 epochs. After reducing the learning rate to 0.002, extra 10 epochs are employed. The minibatch size for the DNN training is chosen as 512.

The ASR system is developed on the WSJCAM0 speech corpus [23]. This database consists of 15.5 hours from 92 adult male/female speakers for training. The training set is being referred to as *AdTr* in this paper. There are a total of 7861 utterances with approximately 90 sentences per speaker in the

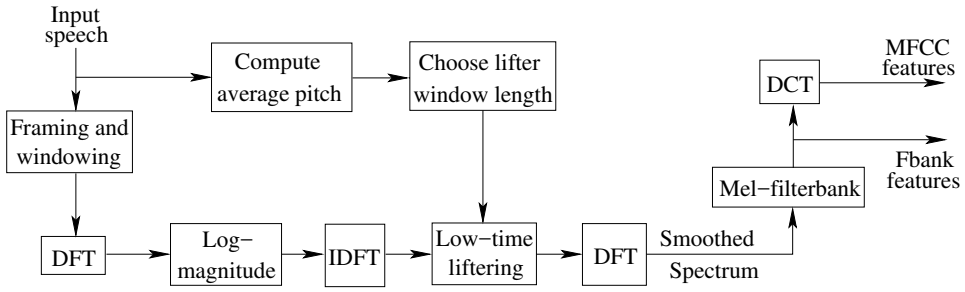

Figure 2: *Block diagram for the extraction of the pitch-robust acoustic features applying adaptive-liftering for spectral smoothening.*

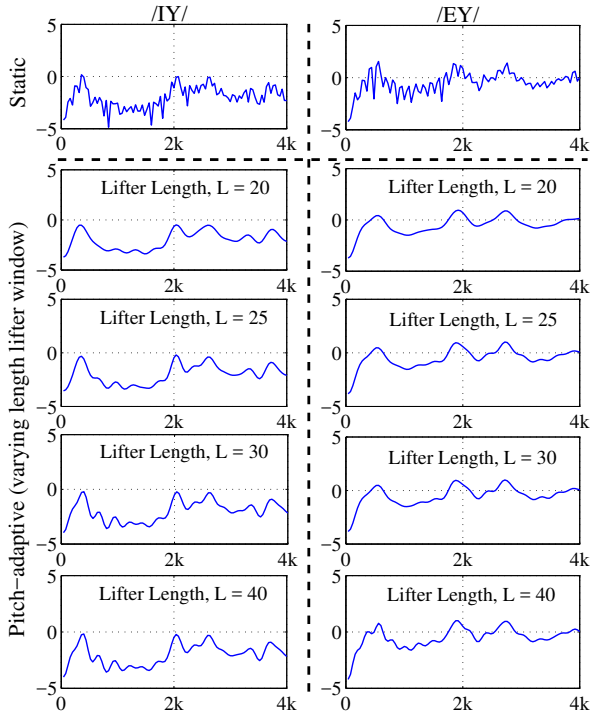

Figure 3: *Demonstration of the spectral smoothing effected by the proposed approach. The left and the right panels show the log-compressed magnitude spectra for a high-pitched (F_0 = 300 Hz) speech frame for the vowels /IY/ and /EY/, respectively, collected from TIMIT. In these panels, the x-axis denotes the frequency values in Hz.*

AdTr set. For the matched case testing, the adults' speech test set (*AdTs*) used. This test set consists of 0.6 hours of speech data from 20 speakers. The number of words in the AdTs set is 5,608. In the mismatched test scenario, the effectiveness of the explored approaches is evaluated using the children's speech test set (*ChTs*) of the PF-STAR British English speech database [24]. This test set contains 1.1 hours of speech data from 60 child speakers with a total of 5067 words. The train set of the PF-STAR (*PfTr*) contains 8.3 hours of data from 122 child speakers. Experimental evaluations are done for the narrowband (sampled at 8 kHz rate) as well as the wideband (16 kHz rate) speech cases.

During the default (or the static) MFCC feature computation, the speech signal is pre-emphasized using a pre-emphasis factor equal to 0.97. For the signal analysis, overlapping Hamming windows of duration 25 ms each having the frame rate of 100 Hz are applied. The 13-dimensional base features (C_0-C_{12}) are computed using 21- and 40-channel Mel-filterbank for

Table 1: *The WERs for adults' speech trained ASR system under acoustically matched and mismatched test conditions. All the reported performances include the fMLLR-based normalization being applied to the MFCC features.*

Speech data used for SI system training	WER (in %)	
	Matched (AdTs)	Mismatched (ChTs)
AdTr (narrowband)	6.20	24.25

the narrowband and the wideband cases, respectively. The base features are time-spliced considering 4 frames to the left and to the right of the current frame. Linear discriminant analysis followed by the maximum likelihood linear transform is applied to derive 39-dimensional acoustic feature vectors. The obtained vectors are further spliced in time taking a context size of 9 before training the DNN-HMM parameters. The frame rate and the frame width are kept the same for the proposed method as well. For the static as well the proposed feature extraction techniques, the MATLAB toolbox called VOICEBOX [25] is used with desired modifications done to the original code.

In the matched case testing, the standard MIT-Lincoln 5k Wall Street Journal bi-gram language model (LM) is used. This LM has a perplexity of 95.3 for the AdTs while there are no out-of-vocabulary (OOV) words. The lexicon used in the matched case has a total of 5,850 words including the pronunciation variations. The use of the MIT-Lincoln LM in decoding ChTs set was not found suitable at all due to the large differences in the word-list and the word counts across the adults' and the children's datasets. Consequently, a domain-specific LM is employed while decoding the ChTs test set to prevent this linguistic mismatch from affecting the results in this study. The domain-specific LM is trained on the transcripts of the speech data in PF-STAR excluding ChTs (i.e., on the transcripts for the ChTr set only). The employed LM has an OOV rate of 1.20% and perplexity of 95.8 for the ChTs set, respectively. A lexicon of 1,969 words including the pronunciation variations is employed. The word error rate (WER) metric is used as a measure of the recognition performance.

3.1. Results and discussion

The WERs for the baseline SI system with respect to the adults' (matched) and the children's (mismatched) test sets are given in Table 1. The given WERs are for the fMLLR-normalized MFCC features case as those are reported to be much superior to the unnormalized ones. The employed fMLLR transform is estimated in the speaker adaptive training mode on a Gaussian-mixture-based ASR system [26]. A highly degraded recognition

Table 2: *Recognition performances for the static and the proposed acoustic features (MFCCs as well as the log-Mel-filterbank energies). Also shown are the percentage relative improvements obtained by the use of the proposed features over their respective static counterparts. For contrast, the PLP features are also included in the static case. For all the studied cases, the acoustic features are normalized using the fMLLR before training the DNN-HMM parameters.*

Data	WER (in %)					% Relative Improvements	
	Static			Adaptive			
Condition	MFCC	Fbank	PLP	MFCC	Fbank	MFCC	Fbank
Narrowband	24.25	23.90	24.00	22.33	22.18	8	7
Wideband	20.38	19.82	20.10	18.14	18.27	11	8

Table 3: *The percentage relative reduction in the WERs obtained over the static case with the use of the proposed MFCC features including the VTLN.*

Data	WER (in %)		%
Condition	Static	Adaptive	Imp.
Narrowband	23.57	21.43	9
Wideband	21.76	19.83	9

Table 4: *The WERs for the proposed MFCC features under the matched test conditions. For evaluating these performances, separate ASR systems are developed for the adult and the child speakers using their respective fMLLR normalized features.*

Testing	WER (in %)	
condition	Static	Adaptive
Adults	6.20	6.28
Children	13.26	13.33

performance can be noted for the mismatched testing case.

For evaluating the effectiveness of the proposed pitch-adaptive features, two separate DNN systems, one using the default/static MFCCs and the other using the adaptive MFCCs, are developed. Table 2 shows the WERs for the two kinds of MFCC features. For contrast, the WERs are also given for the wideband speech case. A significant reduction in the WER is obtained when the pitch-adaptive features are used. For better contrast, we performed the matched case decoding as well. Another DNN-based system is developed using the ChTr train set (narrowband) for this study. The WERs for the matched case testing are given in Table 4. The observed changes in the matched case testing with the use of proposed features happen to be insignificant.

We have also explored the effectiveness of the vocal tract length normalization (VTLN) in the context of the mismatched ASR. For the VTLN, warped features are computed for each of the utterances in the ChTs set by varying the warp factor from 0.88 to 1.12 in steps of 0.02. The differently warped feature sets are aligned against the SI model under the constraint of the first-pass hypothesis. The value of the warp factor resulting in the highest likelihood is chosen to be optimal. During the second-pass decoding, the optimally warped features are employed. The WERs for those experiments are also given in Table 3. Its evident from these studies that the effectiveness of the existing speaker normalization techniques is preserved in the context of the proposed adaptive features.

3.2. Pitch-adaptive Mel-scaled log filterbank features

In the case of the DNNs, some of the recent works have shown that the use of the log-compressed energies at the output of the Mel-filterbank as the acoustic features is superior to the MFCCs [1]. Motivated by those works, we trained a DNN-HMM system using the filterbank features as well. The number of filterbank coefficients is chosen to be 40 as suggested in [1]. The WERs for the baseline DNN system trained using filterbank (Fbank) features are given in Table 2. Moreover, the WERs obtained by using perceptual linear prediction (PLP) features is also enlisted. It is to note that the use of the filterbank/PLP features did not result in significant changes in the recognition performances.

The filterbank features are derived by following the same procedure as that for the MFCCs except that the DCT is avoided in the case of the former. Consequently, we expect that the aforementioned pitch-induced distortions also affect the filterbank features. Motivated by this, the adaptive-liftering approach is applied while deriving the filterbank features. The steps involved in extracting the adaptive-liftering-based filterbank features are shown in Figure 2. The WERs obtained by the use of pitch-adaptive filterbank features is given in Table 2. The proposed filterbank features outperform the static ones. Moreover, the WERs turn out be quite similar for both the kinds of pitch-adaptive front-end features.

4. Conclusion

In this work, we have explored pitch-adaptive signal processing for extracting the front-end features in the context of children's ASR. A novel approach based on adaptive-liftering is presented for smoothing the magnitude spectra prior to the feature computation. The adaptive-liftering-based approach is employed to derive pitch-robust MFCC as well as filterbank features. The effectiveness of the proposed pitch-adaptive features is demonstrated in the DNN-based acoustic modeling paradigm. Both the kinds of pitch-adaptive front-end features are found be highly effective in the case of the children's mismatched ASR. On the other hand, insignificant changes in the WERs are noted for the matched case testing.

5. References

[1] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Commununication*, vol. 49, no. 10-11, pp. 847–860, October 2007.

[12] S. S. Gray, D. Willett, J. Pinto, J. Lu, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in *Proc. INTERSPEECH, Workshop on Child, Computer and Interaction*, 2014.

[13] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010, ch. 4, pp. 61–90.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[15] 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.

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

[17] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[18] P. McLeod, "Fast, Accurate Pitch Detection Tools for Music Analysis," Ph.D. dissertation, University of Otago, Dunedin, New Zealand, May 2008.

[19] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[20] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

[21] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *INTERSPEECH*, 2000, pp. 464–467.

[22] 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.

[23] 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.

[24] 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.

[25] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2005.

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