IMPROVING CHILDREN SPEECH RECOGNITION THROUGH FEATURE LEARNING FROM RAW SPEECH SIGNAL

S. Pavankumar Dubagunta[1,2]*, Selen Hande Kabil*[1,2]*, and Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Children speech recognition based on short-term spectral features is a challenging task. One of the reasons is that children speech has high fundamental frequency that is comparable to formant frequency values. Furthermore, as children grow, their vocal apparatus also undergoes changes. This presents difficulties in extracting standard short-term spectral-based features reliably for speech recognition. In recent years, novel acoustic modeling methods have emerged that learn both the feature and phone classifier in an end-to-end manner from the raw speech signal. Through an investigation on PF-STAR corpus we show that children speech recognition can be improved using end-to-end acoustic modeling methods.

Index Terms— Children speech recognition, acoustic modeling, convolutional neural networks, end-to-end training.

1. INTRODUCTION

Automatic speech recognition (ASR) task focuses on transcribing the linguistic message from speech signals. ASR systems are aimed to handle the variability in data stemming from different resources, such as the acoustic environment (noise, channel conditions), the speakers (speaker variability), the vocabulary (out of vocabulary words), the style (effect of continuous vs isolated speech on the degree of articulation).

Even though significant emphasis has been put on the field of ASR, children speech recognition continues to be a challenging task mainly due to acoustic and linguistic variability in children speech (as compared to adult speech). More precisely, the acoustic and linguistic characteristics of children speech differ as a function of age depending on the anatomical differences in the vocal tract geometry, the ability to control the articulators and prosody, and the scope of linguistic knowledge [1].

On the acoustic side, previous studies demonstrate that children speech exhibits higher fundamental and formant frequencies, and greater spectral variability in comparison to adult speech [1, 2, 3]. The close fundamental and formant frequency values cause difficulties during the feature extraction stage in ASR systems, that aims to decompose speaker dependent information (i.e. fundamental frequency) from the phoneme dependent information (i.e. formants) and retains the latter [1]. In addition, the fact that children speech formant values show greater variability results in more overlaps among phonemic classes for children, as compared to adults, which degrades the performance of children ASR [1, 2, 4]. In order to reduce the acoustic variability (hence, the acoustic mismatch between children and adult acoustic spaces), vocal tract length normalisation (VTLN), speaker normalisation and model adaptation are used [1], while age dependent models are used to limit the acoustic space [5].

On the linguistic side, the degradation in recognition performance is due to pronunciation variability associated with children [6], as they tend to use incorrect pronunciations, made up words and ungrammatical phrases. In order to overcome linguistic variability, focus has been put on pronunciation and language modeling. In [6], a custom dictionary based on children's pronunciation is shown to be helpful for detecting the common pronunciation mistakes of children as a function of age, which implies that potential improvements in the recognition performance can be accomplished by using proper pronunciation modeling.

Another reason why children ASR poses challenges is the lack of large, publicly available corpora for children speech. On large amounts of data, results from the state-of-art children ASR systems are promising [7]. To address data scarcity, in [8], data augmentation is proposed for children ASR using stochastic feature mapping (SFM), to transform out-of-domain adult data for GMM-based and DNN-based acoustic models.

In this paper, our focus is on acoustic modeling for children ASR. Standard short-term spectral feature extraction for speech recognition typically assumes a speech production model, with the aim to capture vocal tract system information by modeling the short-term spectral envelop. These methods have largely emerged from the analysis of "typical" adult speech and, as discussed earlier, can affect acoustic modeling. Recently, approaches have emerged where both the features and the classifier can be learned from raw speech signals in an end-to-end manner [9, 10, 11]. Through an investigation of one such approach, we show that children ASR systems can be improved by automatic feature learning.

The paper is organized as follows. Section 2 provides a background on the end-to-end acoustic modeling method that is being investigated and motivates the present work. Section 3 details the databases and experimental setup. Section 4 presents the results and an analysis on our findings. Finally, we conclude in Section 5.

2. BACKGROUND AND MOTIVATION

In the conventional ASR systems (Fig.1-conventional method), the task of recognizing speech is divided into several subtasks, each of which are optimized independently. In [12, 9], an end-to-end acoustic modeling approach was proposed, where both the features and the

This work was partially funded by the Hasler foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS) and by the Swiss National Science Foundation under the project Sparse and HIerarchical Structures for Speech Modeling (SHISSM). The authors gratefully thank them for their financial support and for a fruitful collaboration. We also thank the CSTR group of University of Edinburgh for the Kaldi recipe. e-mail: (see http://www.idiap.ch/en/people/directory).

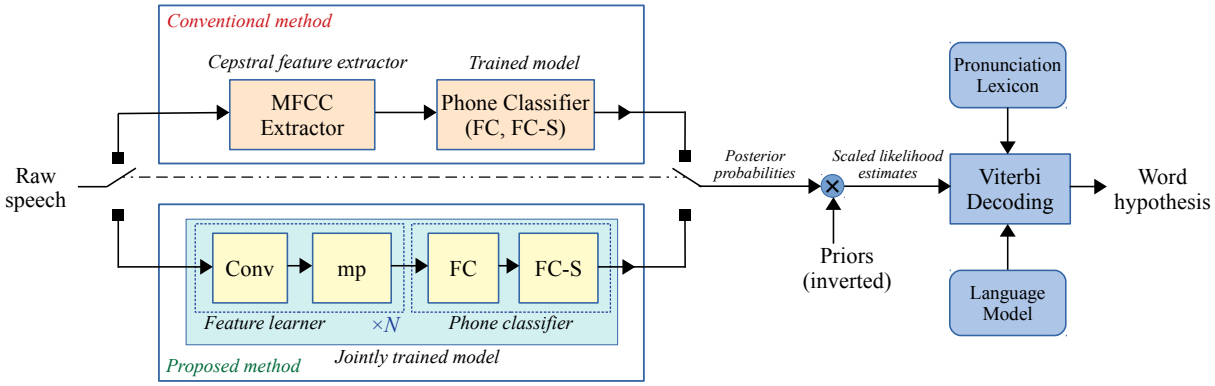

Fig. 1. ASR system flow illustrating the conventional and proposed methods. Conv: convolutional layer with ReLU activations, mp: max-pooling layer, FC: fully connected layer with ReLU activations, FC-S: fully connected layer with softmax activation.

classifier are jointly learned. As shown in Fig.1-proposed method, the CNN based end-to-end acoustic modeling approach is composed of a feature learning stage, that consists of several convolution layers, and a classifier stage, that consists of fully connected (FC) layers (also called a multi-layer perceptron (MLP)) and an output layer.

The hyper parameters of the system include: (i) the window size of the speech input (w_{seq}), (ii) the number of convolution layers N, (iii) for each convolution layer $i \in \{1, \cdots N\}$, kernel width kW_i, kernel shift dW_i, number of filters nf_i and maxpooling size mp_i and (iv) the number of hidden layers in the MLP. All these hyper-parameters in the original work were determined through cross validation. In doing so, the approach also determines the short-term processing applied on the input speech. More precisely, the first convolution layer kernel width (i.e. kW_1) and the kernel shift (i.e. dW_1) are the frame size and frame shift that operate on the signal, respectively. Figure 2 illustrates the first convolution layer processing. Note that the frame rate of the system is determined by the shift of input speech window of size w_{seq}, which was fixed to 10 ms, as done conventionally.

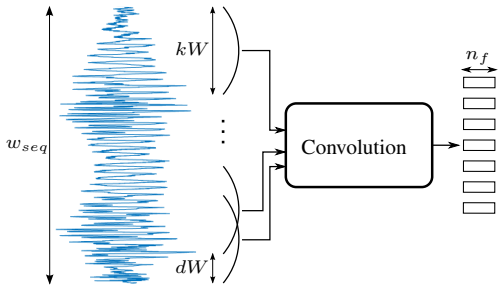

Fig. 2. Illustration of first convolution layer processing.

In [9], it was found that the first convolution models "subsegmental" speech, i.e. speech signal of about 2ms, which is less than one pitch period. Upon analysis of the filters using two different methods, namely, spectral dictionary based interpretation [12] and guided backpropogation based analysis [13], it was found that the CNN learns to model formant frequency information for phone posterior probability estimation. This is interesting, given the fact that the approach does not assume any specific model for the speech signal. Furthermore, it was found that, with fewer number of parameters, this approach is able to yield comparable or better performance than the standard cepstral feature based systems. This paper aims to exploit these two aspects, i.e. automatic feature learning and systems with fewer number of parameters, to improve the performance of children ASR systems.

3. EXPERIMENTAL SETUP

This section first provides a description of the databases and protocols, and then describes the systems developed.

3.1. Datasets

We used PF-STAR dataset [14] for experimenting on children speech and WSJCAM0 [15] for adult speech. Both the datasets contain utterances in British English recorded using two microphones. PF-STAR contains speech from 158 children aged 4 to 14 years and WSJCAM0 is a large vocabulary dataset with 140 speakers. We used BEEP lexicon [16] for PF-STAR ASR. For WSJCAM0, we used the *standard* protocol of using BEEP lexicon added with pronunciations from CMU dictionary for unseen words.

For the experiments on PF-STAR, we used 14.8 hours of data from both the recorded channels, i.e. head mounted microphone (denoted as channel A) and far-field microphone (denoted as channel B), for training the models, as this could partially overcome data scarcity. For the neural network training, we use the `eval/adapt` data of PF-STAR as a cross-validation set. We report results on both the channels A and B of test data separately.

Standard training (train), development (dev) and test sets of WSJCAM0 were used for experimentation. Standard 20k pruned trigram LMs of WSJ corpus were used in decoding WSJCAM0 utterances.

Language model (LM) for PF-STAR was built as follows: one LM was built from the training set with Witten-Bell smoothing and another using normalised text from MGB-3 challenge [17] with Witten-Bell smoothing. The two LMs were linearly interpolated by weights determined based on their perplexities on the PF-STAR cross-validation set (described above), and the resultant model is pruned to remove low probabilities using 10^{-8} as a threshold.

Table 1. CNN architectures. n_f: number of filters, kW: kernel width, dW: kernel shift, mp: max-pooling.

Model	Layer	Conv			mp
		n_f	kW	dW	
CNN3	1	80	30	10	3
	2,3	60	7	1	3
CNN4	1	200	30	5	4
	2,3,4	100	7	1	2
CNN5	1	200	30	5	4
	2	100	9	1	2
	3	100	8	1	2
	4	100	7	1	2
	5	100	6	1	2

3.1.1. GMM-HMM systems

Kaldi toolkit [18] was used to train all the GMM-HMM systems. We followed the standard procedure of training systems: monophone, triphone, LDA+MLLT and LDA+MLLT+fMLLR+SAT. The leaf nodes during context-dependent clustering in all the systems were limited to a maximum of 2500 nodes and the number of Gaussians to 15000. SGMM systems were then trained with 2500 leaf nodes, 9000 substates and 400 mixtures per state.

3.1.2. DNN-HMM systems

Keras [19] with Tensorflow [20] backend was used to train all the neural networks. The feature used was 429 dimensional, consisting of 13 dimensional MFCC with CMVN, with 11-frame splicing and their $\Delta + \Delta\Delta$ coefficients. The DNNs, indicated as DNN1 and DNN3, consisted of one and three hidden layers respectively, with 1024 nodes each and with rectified linear unit (ReLU) activations, followed by an output layer with softmax. The monophone DNNs had monophone states as targets, whereas the triphone systems had clusters from SGMM system as targets. The alignments from the corresponding systems were used to train the systems. The DNN parameters were initialised using Glorot uniform distribution method, the default in Keras. The training was performed using stochastic gradient descent with cross-entropy loss, with 20% dropout on all except the final layer, and the learning rate was halved in the range 10^{-1} to 10^{-6} whenever the cross-validation loss stopped reducing. The posterior probabilities from the neural networks were scaled by priors (computed from the targets used for training) and were used for decoding or forced alignment in Kaldi. During decoding, the HMM state transition probabilities were taken from the corresponding GMM-HMM system from which they were trained. Since monophone system alignments were imperfect, the training of the DNN followed a re-alignment process using the DNN-HMM system. The DNNs were then retrained from random initialization. This process was repeated twice.

3.1.3. CNN-HMM systems

The CNNs were trained using Keras-Tensorflow. Raw speech signals were presented as segments of 250ms with a shift of 10ms. Each segment was mean subtracted (by its scalar mean) and normalised by its standard deviation before feeding to the CNN. The CNN architectures are listed in Table 1. All the CNNs contained a single fully connected hidden layer of 1024 nodes with ReLU activations, followed by an output FC layer with a softmax. The hidden FC layer was trained with a dropout of 20%. The labels of the centre-portion of the segment, determined from the training alignments, were used to train the CNNs. The training procedures were similar to those used for the DNNs.

4. RESULTS AND DISCUSSION

Table 2 shows word error rates (WER) on children speech test set (channels A and B) using models trained from children speech and with added adult speech. We observe that the CNN based systems perform consistently better than or comparable to their GMM/HMM and DNN/HMM counterparts. Also, the SGMM systems benefit from data scarcity and from multi-pass decoding to yield competent results. It is worth mentioning that, to the best of our knowledge, the performance 11.99% WER is the best reported on PF-STAR corpus [21, 22].

Table 2. Comparison of WER on children test data with children models and children+adult models. Bold font indicates the best system w.r.t the test set channel in both monophone and triphone trainings.

Model trained on →		Children data		Added adult data	
Children test set →		A	B	A	B
mono	GMM	17.84	19.27	18.43	20.63
	DNN1	15.67	16.63	15.88	17.69
	DNN3	15.84	17.21	15.62	17.60
	CNN3	**15.09**	**15.63**	15.12	16.72
	CNN4	16.21	16.13	15.68	16.90
	CNN5	17.35	17.00	15.82	17.37
tri	SGMM	13.18	14.64	12.38	14.54
	DNN1	14.65	15.52	14.77	16.28
	DNN3	15.54	16.34	14.37	16.41
	CNN3	13.25	13.87	**11.99**	14.42
	CNN4	14.09	14.40	12.49	14.40
	CNN5	13.43	14.21	12.24	**13.77**

Table 3 shows the impact on WER of adding children data to adult ASR. We observe that adding children speech data reduces the performance.

Table 3. Comparison of WER on adult test data with adult models and adult+children models, showing the effect of adding children data on adult speech recognition.

Model trained on →		Adult data		Added children data	
Adult test set →		dev	test	dev	test
mono	GMM	28.28	28.27	28.84	29.04
	DNN1	15.60	15.69	18.27	18.01
	DNN3	13.12	13.18	14.63	14.37
	CNN3	14.96	14.12	16.91	16.18
	CNN4	13.99	13.68	15.74	15.04
	CNN5	14.32	13.80	16.14	15.43
tri	SGMM	9.10	9.44	9.32	9.56
	DNN1	10.98	10.64	11.53	11.80
	DNN3	9.66	9.29	10.30	10.44
	CNN3	10.83	10.24	12.09	11.44
	CNN4	10.31	9.70	11.51	11.08
	CNN5	9.93	9.53	10.85	10.55

4.1. Analysis based on spectral dictionary interpretation

In [12], a spectral dictionary interpretation was proposed to understand the information modeled by the first convolution layer. This approach has been applied in other studies, such as [23] and [24], to understand the spectral information modeled by the CNNs. In this

approach, the spectral response of the filters to the input speech is calculated in the following manner:

1) s_t^c was taken as the input speech segment. For the sake of simplicity, a window size of 30 ms similar to the one used in standard short term processing is used in our analysis.
2) Successive windows of kW samples (30 samples for all models) interspaced by dW samples (10 samples for CNN3, 5 samples for CNN4 and CNN5 models) are taken from s_t^c.
3) For each of these successive window signals (s_t), the outputs of the filters y_t to the input speech signal $s_t = s_{t-(kW-1)/2} \ldots s_{t+(kW-1)/2}$ are estimated as

$$y_t[m] = \sum_{l=-(kW-1)/2}^{l=+(kW-1)/2} f_m[l].s_{t+l} \quad (1)$$

where f_m denotes the m^{th} filter in first convolution layer and $y_t[m]$ denotes the output of the m^{th} filter at time frame t.

4) The frequency response S_t of the input signal s_t is estimated as

$$S_t = |\sum_{m=1}^{M} y_t[m].\mathcal{F}_m|, \quad (2)$$

where $\mathcal{F}_m$ is the complex Fourier transform of the filter f_m.

5) The spectral response of the 30 ms speech is calculated by summing the frequency responses $S_t \; \forall t \in \{1 \cdots (30ms \cdot sf)/dW\}$ at all the frames and dividing it by the number of frames $(30ms \cdot sf)/dW$. Here sf denotes the sampling frequency.

We used the American English Vowels dataset [3] for our analysis. It consists of recordings of 12 vowels (/ae/, /ah/, /aw/, /eh/, /er/, /ey/, /ih/, /iy/, /oa/, /oo/, /uh/, /uw/), for each of its speakers (50 men, 50 women, 29 boys, 21 girls). Additionally, in [3], the frequency ranges for the F0 and formants were calculated and presented for each utterance. We conducted analysis on five vowels (/er/, /ei/, /ih/, /iy/, /oa/) from four speakers (man [with speaker id m01], woman [w10] , boy [b23], girl [g06]). The subset of phones and speakers were chosen based on the confusion matrix presented in [25]. Figure 3 shows the spectral response of a 30 ms frame from the steady state region of /er/ of the boy speaker [b23]. The formant values tend to match with the range provided in the data set. We observed similar trends across different vowels and speakers.

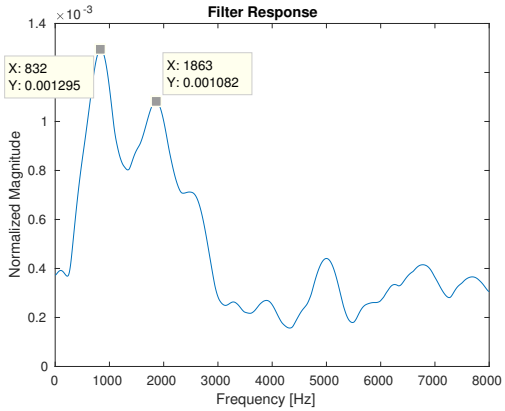

Fig. 3. Average filter response for a speech segment /er/ from CNN3 trained on children speech

4.2. Analysis based on relevance signals

The analysis method described in Sec. 4.1 is limited to the first layer of the CNNs. To gain further insight into what the CNNs learn as a whole, we applied a recently developed *guided backpropagation* based visualization method [13]. Briefly, this visualization technique, given an input signal and the output class, measures how a small variation or perturbation of each input sample value impacts the prediction score. In doing so, the technique tends to measure the importance of each input speech sample for the prediction. This process yields a relevance signal, which can then be analyzed using short-term spectral analysis techniques to understand the information learned by the CNNs. Figure 4 shows the spectrum of the relevance signal computed for the same frame of /er/ of the boy speaker b23 and its envelop based on linear prediction (LP). The formant frequencies computed from the LP envelop (shown) are close to the reference intervals provided in [25].

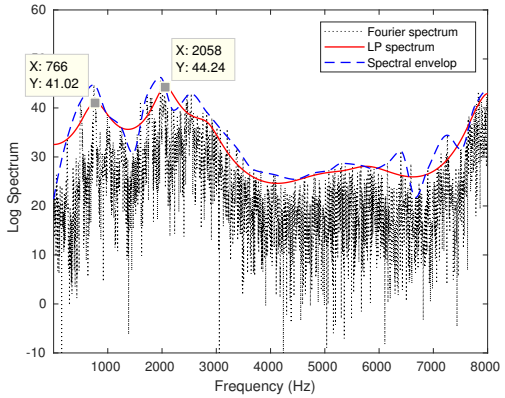

Fig. 4. Spectrum of the relevance signal, its LP spectrum and envelop for the speech segment /er/ from CNN3 trained on children speech

4.3. Adaptation studies

To ascertain that the CNN filters learn representations invariant to differences in adult versus children speech, we conducted an adaptation study, where the adult CNN5 model was adapted in terms of the output layer, freezing the rest of the layers, with the context dependent (CD) states from (a) adult data and (b) children data. It can be observed that, with children CD states, the system yields performances comparable to the CNNs trained from scratch with children speech. The slight drop in performance with adult data CD states could be attributed to the mismatch between children speech and adult speech.

Table 4. Adaptation studies

	A	B
With adult CD states	14.09	16.60
With children CD states	13.36	15.62

5. CONCLUSION

This paper compared the standard cepstral feature based ASR approach and CNN-based end-to-end acoustic modeling approach that jointly learns the relevant features and a phone classifier from raw speech for children speech recognition. Our studies on PF-STAR corpus showed that CNN-based end-to-end acoustic modeling yields better systems than those with the standard features like MFCCs. Our studies also showed that augmenting children data with adult speech data could improve the system further. An analysis of the trained CNNs revealed that the CNNs learn to model formant information invariant to the acoustic differences in children and adult speech.

6. REFERENCES

[1] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children." in *Proceedings of Eurospeech*, 1997.

[2] S. Lee, A. Potamianos, and S. Narayan, "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.

[3] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical Society of America*, vol. 97, pp. 3099–111, 06 1995.

[4] S. Palethorpe, R. Wales, J. Clark, and T. Senserrick, "Vowel classification in children," *The Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3843–3851, 1996.

[5] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adult's speech recognition," in *Proceedings of Italian Computational Linguistics Conference*, 2014.

[6] P. Shivakumar, A. Potamianos, S. Lee, and S. Narayan, "Improving children's speech recognition using acoustic adaptation and pronunciation modeling," in *Proceedings of the Workshop on Child Computer Interaction*, 2014.

[7] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q. Jiang, T. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proceedings of Interspeech*, 2015.

[8] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proceedings of Interspeech*, 2016.

[9] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of Interspeech*, 2013.

[10] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the Speech Front-end With Raw Waveform CLDNNs," in *Proceedings of Interspeech*, 2015.

[11] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," in *Proceedings of Interspeech*, 2015.

[12] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," *Speech Communication*, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.01.004

[13] H. Muckenhirn, V. Abrol, M. Magimai.-Doss, and S. Marcel, "Gradient-based spectral visualization of CNNs using raw waveforms," Idiap Research Institute, Tech. Rep. Idiap-RR-11-2018, Jul 2018. [Online]. Available: http://publications.idiap.ch/downloads/reports/2018/Muckenhirn_Idiap-RR-11-2018.pdf

[14] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *Proceedings of Ninth European Conf. Speech Communication and Technology*, 2005.

[15] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: a british english speech corpus for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, 1995.

[16] "BEEP dictionary," http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, accessed: 01-07-2018.

[17] "MGB challenge lexicon," http://data.cstr.ed.ac.uk/asru/MGB3/data/lm/mgb.normalized.lm, accessed: 01-07-2018.

[18] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[20] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," http://tensorflow.org/, 2015.

[21] M. J. Russell, S. D'Arcy, and L. P. Wong, "Recognition of read and spontaneous children's speech using two new corpora," in *Proceedings of Interspeech*, 2004.

[22] M. J. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, 2007, pp. 108–111.

[23] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proceedings of ICASSP*, 2018.

[24] S. Kabil, H. Muckenhirn, and M. Magimai.-Doss, "On learning to identify genders from raw speech signal using cnns," in *Proceedings of Interspeech*, 2018.

[25] "American vowels database," https://homepages.wmich.edu/~hillenbr/voweldata.html, accessed: 15-07-2018.