TRANSCRIBING THE AUDIO STREAM IN CHILDREN

SOMU SATISH KUMAR[1] VEMULA NAGARJUNA[2] SK YEZULLA HUSSAIN[3] GURRAM RAJESH KUMAR[4]
[1]ASST.PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY,
[2] ASST. PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY,
[3]ASST. PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY,
[4] ASST. PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY,
[1,2,3,4] SRI MITTAPALLI COLLEGE OF ENGINEERING

Abstract- Using short-term spectral features to recognize children's speech is a challenging task. One explanation is that the way kids talk is really basic, as when they're still learning how to build frequency values. And when they develop into adults, their vocal apparatus changes too. Because of this, removing traditional short-term spectral features for speech recognition reliably becomes a challenge. Recent years have seen the development of new acoustic modelling algorithms that learn the function and television categorization from the raw speech signal in an end-to-end manner. We show that methods using children's acoustic modelling improve speech recognition using the PF-STAR corpus.

Keywords--Children speech recognition, acoustic modeling, convolutional neural networks, end-to-end training.

I. INTRODUCTION

The task of linguistically transcribing the audio stream is known as automatic speech recognition (ASR). The goal of automatic speech recognition systems is to control data variability caused by several sources, including the acoustic environment (ray, canal conditions), the speakers (speaker variability), the vocabulary (out of words), and the style (effect on the articulation grade of continuous versus isolated speech). The linguistic and acoustic variety of children's voices remains an obstacle to children's speech recognition, even though the area of automatic speech recognition (ASR) has received a lot of attention. Specific auditory and linguistic characteristics of speaking to children include age-related changes in vocal geometry and anatomy, the ability to control articulators and prosody, and the range of language abilities [1]. Young people's speech is more spectral, formational, and fundamental than adults', according to acoustic research [1, 2.3]. Issues arising from values close to the fundamental frequency (i.e., during the feature extraction phase of ASR systems) deconstruct and preserve information derived from the formants of the phoneme.In addition, children's ASR performance is worse than adults' because their speech formant values are more variable, leading to greater overlaps in their phonemic classes[1, 2, 4]. While agenetic models are employed to limit the acoustic area and decrease acoustic variability (thus, acoustic mismatch between children and adult acoustical spaces), methods such as normalising the Voice Length (VTLN), speakers, and adaptation to the model are employed [1]. From a linguistic perspective, children's varied pronunciations—which often include misspelt words and grammatical errors—are associated with deteriorating recognition abilities [6]. In an effort to conquer linguistic variety, pronunciation and language modelling have been the focus. A children's pronunciation-based lexicon was shown to be effective in identifying age-related pronunciation mistakes in children in [6], suggesting that accurate

pronunciation modelling might lead to improvements in recognition ability. Another reason young ASR are encountering challenges is that there isn't a big, publicly accessible corporation that talks to them. Results from cutting-edge children's ASR systems show promise in massive datasets [7]. For children with ASR, the authors of [8] suggest using stochastic feature mapping (S FM) to improve acoustic models based on GMM and DNN, respectively, in order to circumvent data limitations. This research primarily aims to examine ASR acoustic modelling in youngsters. In most cases, a model of speech production is used to extract short-term spectral features for language recognition. The goal is to capture information about the vocal tract system using this model. As previously stated, the findings from studies on "normal" adult speech have basically materialised and have the potential to influence acoustic models. There have been recent developments in end-to-end methods for learning features and the classifier from raw audio signals [9, 10, 11]. We demonstrate that automated feature learning may enhance children's ASR systems via an examination of this sort of method. Everything is laid out correctly on the page. Section 2 provides a synopsis of the end-to-end acoustic modelling methodology that underpins and promotes the present investigation. There is a discussion of the databases and the experimental setup in Section 3. We provide our results and analysis in Section 4. Section 5 concludes the whole thing.

II. RELATEDWORKS

Conventional ASR systems (Fig.1-conventional method) optimise each subtask independently, breaking the task of speech recognition into several smaller tasks. The classifier and attributes are part of an end-to-end approach to acoustic modelling that was laid out in [12, 9]. As illustrated in figure 1, the CNN-based end-to-end acoustic modelling approach is built upon a feature-learning phase that uses multiple convolutional layers, a classification phase that uses fully connected (FC) layers, also called the multi layer perceptron (MLP), and an output layer.

.

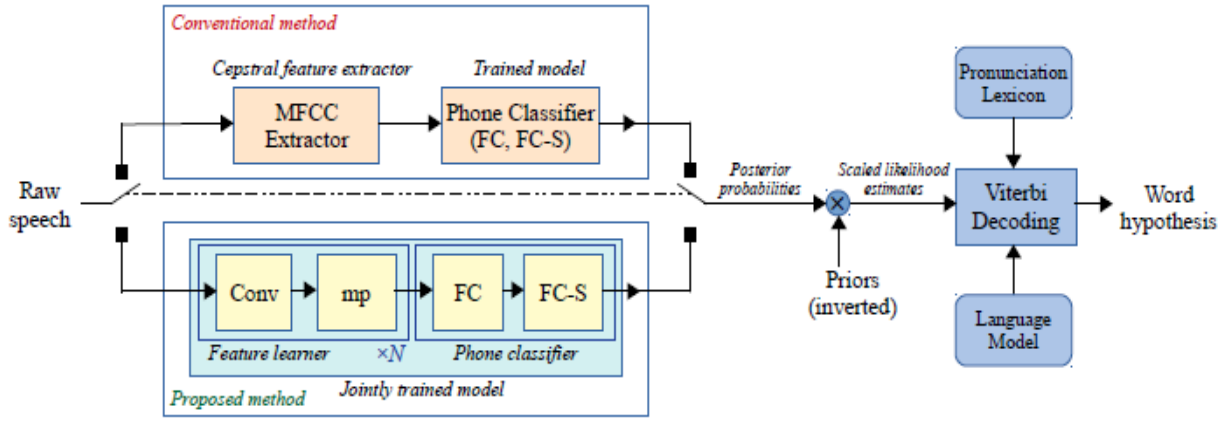

Fig. 1: ASR system flow illustrating the conventional and proposed methods.

This system's hyperparameters consist of the following: (i) the size of the speech input window (wseq), (ii) the total number of convolution layers (N), (iii) for every convolution layer i ∈ f{1..N}, the width of the kernel (kWi), the shift of the kernel (dWi), the number of filters (nfi), the maximum pooling size (mpi), and (iv) the number of hidden layers in the MLP. The original

research detected all of these hyper parameters via cross validation. The input talk's processing speed is likewise affected by this strategy. To be more specific, the frame size and frame shift that operate on the signal are the first kernel layer width (kW1) and the kernel shift (dW1), respectively, of the convolution layer. The processing of the first convolutional layer is shown in Figure 2. This system's frame rate is dictated by the shift of the input speech window of size wseq, which was set to 10 ms in accordance with standard practice.

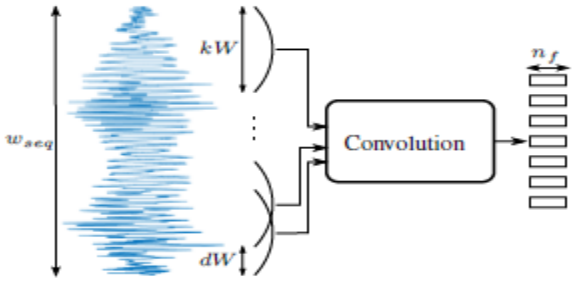

Fig. 2. Illustration of first convolution layer processing.

It was initially discovered in [9] to use convolutional modelling on the "sub-segment," or a 2-millisecond signal smaller than one pitch. Following two separate analyses of the filters—one using a spectral dictionary[12] and the other using back propogation[13]—the CNN learnt to model generating frequency information for post-probability evaluation of the phone. Furthermore, our approach outperforms the conventional cepstral functional system with fewer parameters, or at least provides comparable performance. In order to improve the performance of ASR systems for children, this research will use two features: automated functional learning and a reduced number of system parameters.

III. CONFIGURATION SETTING

This section discusses the databases and protocols first and then the created systems.

Datasets

The PF-STAR [14] and WSJCAM [15] programs were used for the speech tests involving children and adults, respectively. Two microphones were used to record utterances in British English for both datasets. With 140% of the speakers included, PFSTAR is a large vocabulary dataset. It includes 158 kids, ranging in age from four to fourteen years old. The webcam is WSJCAM0. Our team made use of the BEEP [16] lexicon for PF-STAR ASR. With the addition of CMU dictionary pronunciations for unseen words, we have used the standard BEEP lexicon protocol for WSJCAM0.In order to compensate for data shortages, models were trained using data from both the recorded channels (channel A, which included head-mounted microphones) and the far-field microphones (channel B), which were used in conjunction with PF-STAR. Neural network training makes advantage of the PF-STAR evaluation/adaptation data as a cross-validation set. We provide findings separately for the A and B test channels. Training (train), development (dev), and test sets from the standard WSJCAM0 were used for the experiments. We used the standard 20k trigram LMs from the WSJ corpus to decode WSJCAM0 utterances. One LM comes from the Witten-Bell-Smoothing training set and the other from Witten-Bell Smoothing with regular MGB-3 text [17]. This is how the PF-STAR language model was created. The two LMs have been linearly interpolated using weights selected according to

their concerns in the cross-validation set PF-STAR (discussed above) in order to exclude the lower probability using 10-8 as a cutoff.

GMM-HMM systems

All GMM-HMM systems might be trained with the help of Kaldi's toolbox [18]. Utilising standard training system protocols, we provide monophonous, triphonic, and LDA+MLLT variants, in addition to LDA+MLLT+fMLLR+SAT. All systems had a cap of 2,500 sheet nodes and 15,000 Gaussians for context-dependent clustering. The next day, SGMM networks consisting of 9,000 substates, 2500 leaf nodes, and 400 mixes per state were

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

DNN-HMM systems

Training the neural networks with the Tensorflow [20] backend was done using Keras [19]. With 11-screen splicing and related coefficients, the feature used was a 429-size MFCC 13-size CMVN. A softmax output layer with activation using a rectified linear unit (ReLU) followed by one or three hidden layers, each with 1024 nodes, made up the DNNs, which were known as DNN1 and DNN3. For single-phone states, SGMM clusters were designed, while for triphone systems, monophone DNNs were envisioned. It was the alignments from those systems that were used to train the systems. The default parameters for DNNs in Keras were initialised using the Glorot uniform distribution approach. During training, all layers except the final one were subject to a 20% reduction using a cross-entropical loss and stochastic gradient descent. Half of the learning rate fell within the 10-1 to 10-6 range after cross-validation loss stopped dropping. Used to decode or forcefully align neural networks in Kaldi, they are scaled up by priors, which are generated from training objectives. The GMM-HMM system they learnt from during decoding was used to determine the likelihood of the HMM state change. When deciphering Since the alignments produced by the monophone system were subpar, an alignment technique employing the DNN-HMM system was implemented after the DNN training. Afterwards, the DNNs were re-exercised at random. It has already been said twice.

CNN-HMM systems

In order to train the CNNs, Keras-Tensorflow was used. With a 10ms shift, the raw speech signals were shown in 250ms blocks. Prior to feeding it into the CNN, each segment was normalised and removed mean (using its scalar average). The CNN architectures are shown in Table 1. A softmax output FC layer follows each convolutional neural network (CNN) that has one fully connected hidden 1024 node layer. The hidden FC layer was subjected to a 20%

dropout. The segment's centre labels, derived from the training alignments, were used to train CNNs. The DNNs' training procedures were same.

IV. RESULTS AND DISCUSSION

Table 2 shows the word error rates (WER) on the child speech test set (Channels A and B) while employing adult speech as well as speech-trained models. It is worth mentioning that CNN systems consistently outperform GMM/HMM and DNN/HMM. Additionally, SGMM systems are able to provide respectable outcomes by capitalising on data scarcity and decoding multipasses. It should be mentioned that, as far as we are aware, the best-reported PF-STAR corpus is 11.99% WER [21, 22]. Table 3 shows the impact on WER of incorporating data from children into adult ASR. By including data from children's voices, we may see that it reduces performance.

In Table 2, we can see how WER compares to children's models and children+adult models on test data.

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. 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

It was suggested in [12] to understand the data shown on the spectral dictionary's first convolutional layer. In previous studies, the method was used to comprehend the spectrum data represented by the convolutional neural networks (CNNs), as shown in references [23] and [24].

This is how the filters' spectral reactions to the input language are calculated:s_t^cwas used as the section of input speech. Our research makes use of a 30-millisecond frame, which is comparable to that of conventional short-term processing, in order to keep things simple.Interspersed with dW samples (10 samples for CNN3, 5 samples for CNN4 and CNN5 models), successive windows of kW samples (30 samples for all models) are extracted from s_t^c.Every time a new window signal (st) appears, the predicted outputs of the filters applied to the input speech signal (St) = St-(kW-1)/2….St + (kW-1)/2 are

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

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.

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

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

Based on the confusion matrix in [25], the subset of telephones and speakers were chosen. The 30-ms-Frame from the steady-state area of/and/of the boy speaker displays spectral response in Figure 3 (b23). The formant values are often consistent with the range in the data set. In various vowels and speakers we saw comparable tendencies.

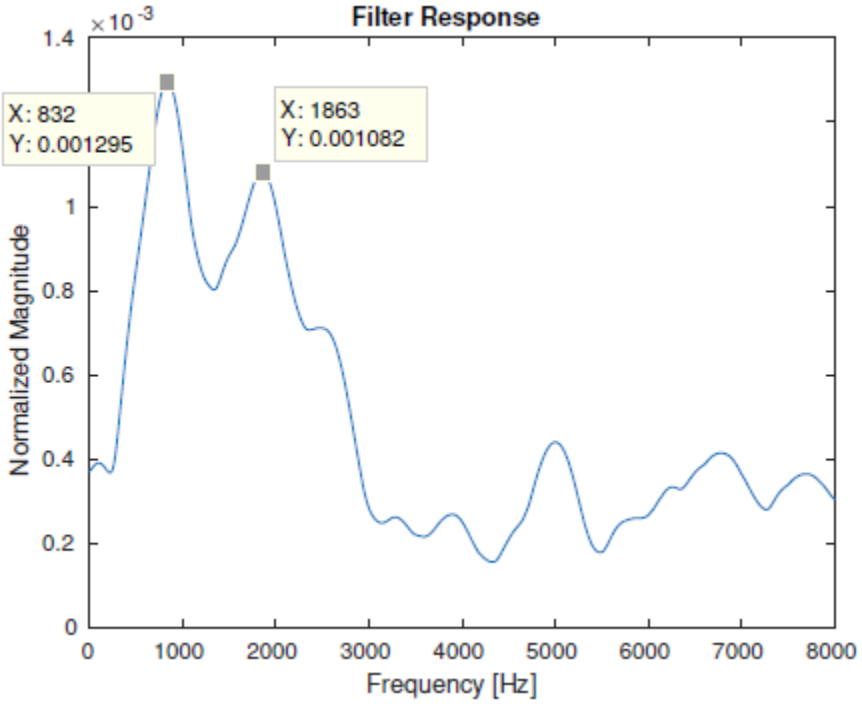

Fig. 3. Average filter response for a speech segment /er/ from CNN3 trained on children speech

V. FUTURE SCOPE AND CONCLUSION

In order to help youngsters learn a new language, this paper compares the conventional cepstral ASR method with a convolutional neural network (CNN) based end-to-end acoustic modelling approach that learns the important features concurrently with the raw language telephone categorisation. Based on our findings from the PF-STAR corpus, systems trained using CNN end-to-end acoustic modelling outperform their more traditional counterparts, such MFCCs. According to our results, the system might be much better if we supplement the input from children with adult voices. Analysing the trained CNNs revealed that they had mastered the art of representing formational information that remains constant across the acoustic variances between children's and adults' speech.

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. T¨uske, R. Schl¨uter, 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, “Endto- 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, “WSJCAMO: 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.peech/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] Anusha, Pureti, T. Sunitha, and Mastan Rao Kale. "Detecting and Analyzing Emotions using Text stream messages." *ECS Transactions* 107.1 (2022): 16913.

[25]Aharonu, Mattakoyya, et al. "Entity linking based graph models for Wikipedia relationships." *Int. J. Eng. Trends Technol* 18.8 (2014): 380-385.