Journal of Cardiovascular Disease Research

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

CNN and MFCC based Speech Net: Children Speech Recognition model

C Ramadevi, Kommu Anusha, Pavankumar Thummeti

Department of Computer Science and Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

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. Finally, the simulations revealed that the proposed MFCC-CNN resulted in superior performance as compared to GMM model.

Key words: 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,

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

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.

2. Literature survey

Cole et. al [9] presented initial work towards development of a children's speech recognition system for use within an interactive reading and comprehension training system. They first describe the Colorado Literacy Tutor project and two corpora collected for children's speech recognition research. Next, baseline speech recognition experiments are performed to illustrate the degree of acoustic mismatch for children in grades K through 5. It is shown that an 11.2% relative reduction in word error rate can be achieved through vocal tract normalization applied to children's speech. Finally, they describe our baseline system for automatic recognition of spontaneously spoken story summaries. It is shown that a word error rate of 42.6% is achieved on the presented children's story summarization task after using unsupervised MAPLR adaptation and VTLN to compensate for interspeaker acoustic variability. Povey et. al [10] described the design of Kaldi, a free and open-source speech recognition toolkit. The toolkit currently supports modeling of context-dependent phones of arbitrary context lengths, and all commonly used techniques that can be estimated using maximum likelihood. It also supports the recently proposed SGMMs. Development of Kaldi is continuing and they are working on using large language models in the FST framework, lattice generation and discriminative training.

Dehak et. al [11] presents an extension of our previous work which proposes a new speaker representation for speaker verification. In this modeling, a new low-dimensional speaker- and channel-dependent space is defined using a simple factor analysis. This space is named the total variability space because it models both speaker and channel variabilities. Two speaker verification systems are proposed which use this new representation. The first system is a support vector machine-based system that uses the cosine kernel to estimate the similarity between the input data. The second system directly uses the cosine similarity as the final decision score. They tested three channel compensation techniques in the total variability space, which are within-class covariance normalization (WCCN), linear discriminate analysis (LDA), and nuisance attribute projection (NAP). Peddinti et. al [12] proposed a time delay neural network architecture which models long term temporal dependencies with training times comparable to standard feed-forward DNNs. The network uses sub-sampling to reduce computation during training. On the Switchboard task they show a relative improvement of 6% over the baseline DNN model.

Tong et. al [13] proposed to improve children's automatic speech recognition performance with transfer learning technique. They compared two transfer learning approaches in enhancing children's speech recognition performance with adults' data. The first method is to perform acoustic model adaptation on the pre-trained adult model. The second is to train acoustic model with deep neural network based multi-task learning approach: the adults' and children's acoustic characteristics are learnt jointly in the shared hidden layers, while the output layers are optimized with different speaker groups. Dubagunta et. al [14] 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

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

features like MFCCs. Our studies also showed that augmenting children's 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. Wu et. al [15] demonstrate the efficacy of TDNN-F for the task of automatically recognizing child speech. They build a TDNN-F system that outperforms its alternatives in datasets with various sizes. They explore the impacts of vocal track length normalization (VTLN) and data augmentation on the performance of TDNN-F system. Though effective with traditional models like GMM-HMM, VTLN has no significantly impacts on TDNN-F. When trained with an extremely small dataset, data augmentation helps improve the performance of TDNN-F on test data in the same channel condition as the training set Qin et. al [16] proposed the advancement of the attention mechanism in joint CTC-attention-based speech recognition. In the first phase, they adopted a high-level feature-based joint model from our previous work. The only difference is that they do not use multi-lingual pre-training for DNN. In the second phase, they introduced a new attention type for our end-to-end models. They add extra connections for the second to the last layer of the encoder and then apply multi-head attentions. Unlike other normal attention types, this attention is scored over multi-level outputs of the RNN and therefore brings extra long-term dependencies on the attention. Experiments on TIMIT show that all of our models perform better than all referenced methods and prove the robustness of this method. Further experiments on WSJ and LibriSpeech show that this attention mechanism could achieve the best performance among all end-to-end methods without data augmentation, and it is only slightly worse than the state-of-the-art performance. Shivakumar and Georgiou et. al [17] presents a systematic and an extensive analysis of the proposed transfer learning technique considering the key factors affecting children's speech recognition from prior literature. Evaluations are presented on (i) comparisons of earlier GMM-HMM and the newer DNN Models, (ii) effectiveness of standard adaptation techniques versus transfer learning, (iii) various adaptation configurations in tackling the variabilities present in children speech, in terms of (a) acoustic spectral variability, and (b) pronunciation variability and linguistic constraints. This Analysis spans over (i) number of DNN model parameters (for adaptation), (ii) amount of adaptation data, (iii) ages of children, (iv) age dependent-independent adaptation. Finally, they provided Recommendations on (i) the favorable strategies over various aforementioned - analyzed parameters, and (ii) potential future research directions and relevant challenges/problems persisting in DNN based ASR for children's speech. Wang et. al [18] evaluated transformer-based acoustic models (AMs) for hybrid speech recognition. Several modeling choices are discussed in this work, including various positional embedding methods and an iterated loss to enable training deep transformers. They also present a preliminary study of using limited right context in transformer models, which makes it possible for streaming applications. They demonstrated that on the widely used Librispeech benchmark, our transformer-based AM outperforms the best published hybrid result by 19% to 26% relative when the standard n-gram language model (LM) is used. Combined with neural network LM for rescoring. Hassan et. al [19] provides the two pipelined deep learning architectures that achieve minimum character error rate (CER) and word error rate (WER) on common voice (CV) benchmark. By setting up different experimental configurations and modifications, we are successful in achieving minimum WER, that is, reduced from 43% to 18.08% at a lower validation cost. As this work focused on South Asian's English accents so, there is a little bit of increase in WER and CER for English speakers. The system will be further scalable towards targeting other South Asian languages like Bengali, Urdu, Hindi, and others with more robust datasets and higher accuracy and the training of both pipelines parallelly. Bhardwaj et. al [20] presented a systematic literature review of children's speech recognition systems studied from 2009 to 2020. The data and information gathered from studies on feature extraction, auditory modeling, datasets, different languages, and surroundings can be used to construct an ASR system for children. This also benefits researchers in conducting new studies to improve the

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

recognition of children's speech. Additionally, data acquired from research publications helps in determining the research area trends and research gaps in this field. From the studies, it was found that research into the recognition of children's speech and its variations are in a very small group of studies. Recognition accuracy was also lagging compared to adult speech recognition. After conducting the SLR it was found that most of the papers (62%) were published by the conferences, while Interspeech published 36.17% of the conference research papers. In the case of journals, the majority of the research papers (22.2%) were published by the Computer Speech & Language journal, while WOCCI is one of the popular workshops where 45.45% of the workshop papers were published.

3. Proposed system

This work is used to recognize the speech signal that is children or adult. Initially, dataset is trained with CNN model. Then, random test speech signal is considered, which is pre- processed to remove the different arti facts. The speech dependent spectral features are extracted using MFCC. Furthermore, CNN model is tested with the MFCC features and identifies the class of speech. Finally, either children or adult class is classified. The performance metrics also calculate to prove the effecting of system.

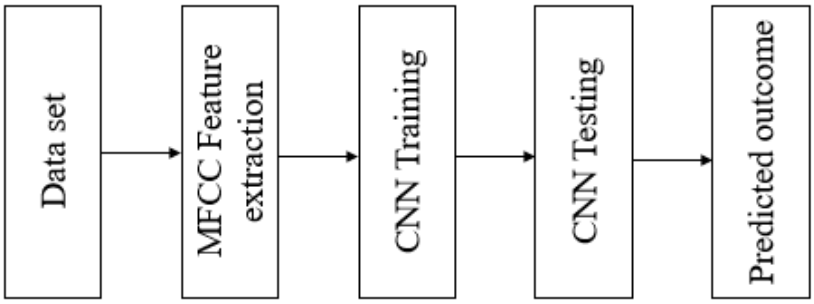


Fig.1: Block diagram of proposed system.

3.1 Dataset

The dataset contains two classes namely children, adult. Here, children class contains 62 speech samples. Similarly, adults class contains 61 speech samples, Respectively

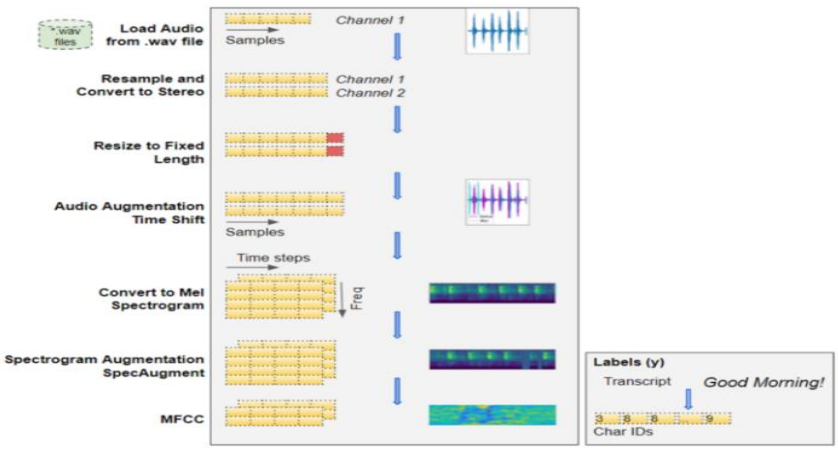


Fig. 2: Speech preprocessing with MFCC.

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

3.2 Preprocessing

In the sound classification article, I explain, step-by-step, the transforms that are used to process audio data for deep learning models. With human speech as well we follow a similar approach. There are several Python libraries that provide the functionality to do this, with librosa being one of the most popular.

Load Audio Files

- Start with input data that consists of audio files of the spoken speech in an audio format such as ".wav" or ".mp3".
- Read the audio data from the file and load it into a 2D Numpy array. This array consists of a sequence of numbers, each representing a measurement of the intensity or amplitude of the sound at a particular moment in time. The number of such measurements is determined by the sampling rate. For instance, if the sampling rate was 44.1kHz, the Numpy array will have a single row of 44,100 numbers for 1 second of audio.
- Audio can have one or two channels, known as mono or stereo, in common parlance. With two-channel audio, we would have another similar sequence of amplitude numbers for the second channel. In other words, our Numpy array will be 3D, with a depth of 2.

Convert to uniform dimensions: sample rate, channels, and duration

- They might have a lot of variation in our audio data items. Clips might be sampled at different rates, or have a different number of channels. The clips will most likely have different durations. As explained above this means that the dimensions of each audio item will be different.
- Since our deep learning models expect all our input items to have a similar size, we now perform some data cleaning steps to standardize the dimensions of our audio data. We resample the audio so that every item has the same sampling rate. We convert all items to the same number of channels. All items also have to be converted to the same audio duration. This involves padding the shorter sequences or truncating the longer sequences.
- If the quality of the audio was poor, we might enhance it by applying a noise-removal algorithm to eliminate background noise so that we can focus on the spoken audio.

Data Augmentation of raw audio

- They could apply some data augmentation techniques to add more variety to our input data and help the model learn to generalize to a wider range of inputs. We could Time Shift our audio left or right randomly by a small percentage, or change the Pitch or the Speed of the audio by a small amount.

Mel Spectrograms

- This raw audio is now converted to Mel Spectrograms. A Spectrogram captures the nature of the audio as an image by decomposing it into the set of frequencies that are included in it.

3.3 MFCC

MFCC stands for Mel Frequency Cepstral Coefficient. MFCC features are widely used in speech recognition problems. Speech is dictated by the way in which we use our oral anatomy to create each sound. Therefore, one way to uniquely identify a sound (independent of the speaker) is to create a mathematical representation that encodes the physical mechanics of spoken language. MFCC features are one approach to encoding this information.

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

The road map of the MFCC technique is given below.

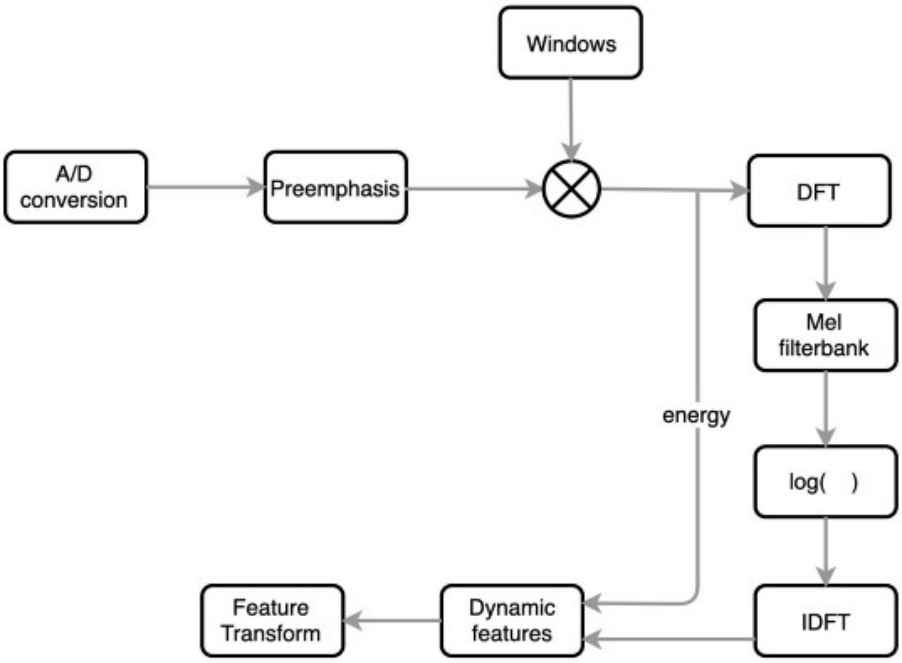


Fig. 3: MFCC block diagram.

A/D Conversion: In this step, convert audio signal from analog to digital format with a sampling frequency of 8kHz or 16kHz.

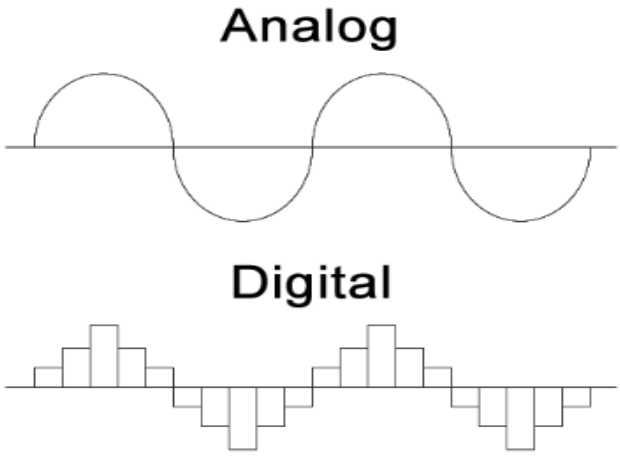


Fig. 4: Illustration of analog to digital conversion.

Preemphasis: Preemphasis increases the magnitude of energy in the higher frequency. When we look at the frequency domain of the audio signal for the voiced segments like vowels, it is observed that the energy at a higher frequency is much lesser than the energy in lower frequencies. Boosting the energy in higher frequencies will improve the phone detection accuracy thereby improving the performance of the model.

Windowing: The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phones in the speech. But in the given audio signal there will be many phones, so we will break the audio signal into different segments with each segment having 25ms width and with the signal at 10ms apart as shown in the below figure. On average a person speaks three words

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

per second with 4 phones and each phone will have three states resulting in 36 states per second or 28ms per state which is close to our 25ms window. From each segment, we will extract 39 features. Moreover, while breaking the signal, if we directly chop it off at the edges of the signal, the sudden fall in amplitude at the edges will produce noise in the high-frequency domain. So instead of a rectangular window, we will use Hamming/Hanning windows to chop the signal which won't produce the noise in the high-frequency region.

DFT (Discrete Fourier Transform): We will convert the signal from the time domain to the frequency domain by applying the dft transform. For audio signals, analyzing in the frequency domain is easier than in the time domain.

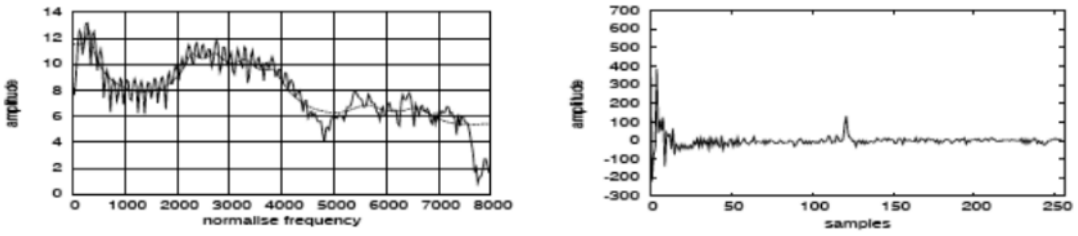
Mel-Filter Bank: The way our ears will perceive the sound is different from how the machines will perceive the sound. Our ears have higher resolution at a lower frequency than at a higher frequency. So, if we hear sound at 200 Hz and 300 Hz we can differentiate it easily when compared to the sounds at 1500 Hz and 1600 Hz even though both had a difference of 100 Hz between them. Whereas for the machine the resolution is the same at all the frequencies. It is noticed that modeling the human hearing property at the feature extraction stage will improve the performance of the model. So, we will use the mel scale to map the actual frequency to the frequency that human beings will perceive. The formula for the mapping is given below.

$$mel(f) = 1127 \ln(1 + \frac{f}{700})$$

Applying Log: Humans are less sensitive to change in audio signal energy at higher energy compared to lower energy. Log function also has a similar property, at a low value of input x gradient of log function will be higher but at high value of input gradient value is less. So, we apply log to the output of Mel-filter to mimic the human hearing system.

IDFT: In this step, we are doing the inverse transform of the output from the previous step. Before knowing why, we have to do inverse transform we have to first understand how the sound produced by human beings.

The sound is actually produced by the glottis which is a valve that controls airflow in and out of the respiratory passages. The vibration of the air in the glottis produces the sound. The vibrations will occur in harmonics and the smallest frequency that is produced is called the fundamental frequency and all the remaining frequencies are multiples of the fundamental frequency. The vibrations that are produced will be passed into the vocal cavity. The vocal cavity selectively amplifies and damp frequencies based on the position of the tongue and other articulators. Each sound produced will have its unique position of the tongue and other articulators. Note that the periods in the time domain and frequency domain are inverted after the transformations. So, the frequency domain's fundamental frequency with the lowest frequency will have the highest frequency in the time domain. The below figure shows the signal sample before and after the idft operation.



ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

The peak frequency at the rightmost in figure(c) is the fundamental frequency and it will provide information about the pitch and frequencies at the rightmost will provide information about the phones. We will discard the fundamental frequency as it is not providing any information about phones. The MFCC model takes the first 12 coefficients of the signal after applying the idft operations. Along with the 12 coefficients, it will take the energy of the signal sample as the feature. It will help in identifying the phones.

3.4 CNN

According to the facts, training and testing of DL-CNN involves in allowing every source image via a succession of convolution layers by a kernel or filter, rectified linear unit (ReLU), max pooling, fully connected layer and utilize SoftMax layer with classification layer to categorize the objects with probabilistic values ranging from [0,1]. Figure 1 discloses the architecture of DL-CNN that is utilized in proposed methodology for CBIR system for enhanced feature representation of word image over conventional retrieval systems. Convolution layer as depicted in Figure 5 is the primary layer to extract the features from a source image and maintains the relationship between pixels by learning the features of image by employing tiny blocks of source data. It's a mathematical function which considers two inputs like source image $I(x, y, d)$ where x and y denotes the spatial coordinates i.e., number of rows and columns. d is denoted as dimension of an image (here $d = 3$, since the source image is RGB) and a filter or kernel with similar size of input image and can be denoted as $F(k_x, k_y, d)$.

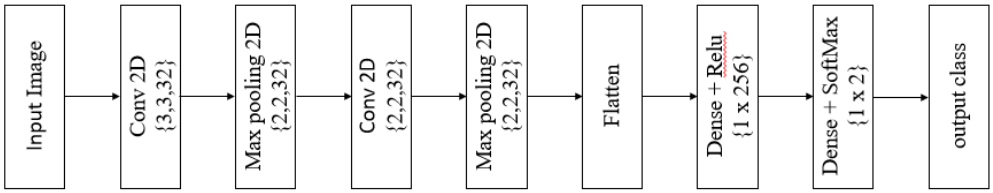


Fig. 5: Representation of convolution layer process.

The output obtained from convolution process of input image and filter has a size of $C\left((x - k_x + 1), (y - k_y + 1), 1\right)$, which is referred as feature map. An example of convolution procedure is demonstrated in Figure 6. Let us assume an input image with a size of 5×5 and the filter having the size of 3×3. The feature map of input image is obtained by multiplying the input image values with the filter values as given in Figure 3.

1	1	1	0	0
0	0	1	1	1
1	1	0	0	1
0	0	0	1	1
1	1	1	0	0

5x5 image

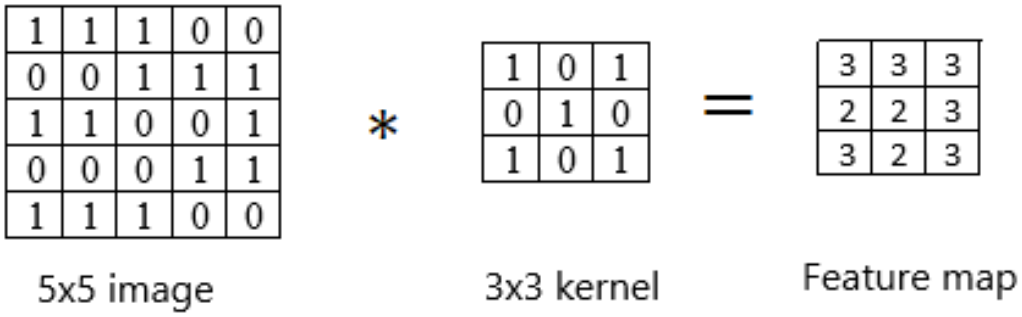
$*$

1	0	1
0	1	0
1	0	1

3x3 kernel

(a)

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019



(b)

Fig. 6: Example of convolution layer process.

(a) an image with size $\mathbf{5 \times 5}$ is convolving with $\mathbf{3 \times 3}$ kernel

(b) Convolved feature map

ReLU layer

Networks those utilizes the rectifier operation for the hidden layers are cited as rectified linear unit (ReLU). This ReLU function $\mathcal{G}(\cdot)$ is a simple computation that returns the value given as input directly if the value of input is greater than zero else returns zero. This can be represented as mathematically using the function $max(\cdot)$ over the set of 0 and the input x as follows:

$$\mathcal{G}(x) = \max\{0, x\}$$

Max pooing layer

This layer mitigates the number of parameters when there are larger size images. This can be called as subsampling or down sampling that mitigates the dimensionality of every feature map by preserving the important information. Max pooling considers the maximum element form the rectified feature map.

5. Results

This section gives the detailed analysis of simulation results implemented using "python environment". Further, the performance of proposed method is compared with existing methods using same dataset.

Fig. 7: Upload Children's & Adult Speech Dataset.

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

Fig. 8: Dataset loaded.

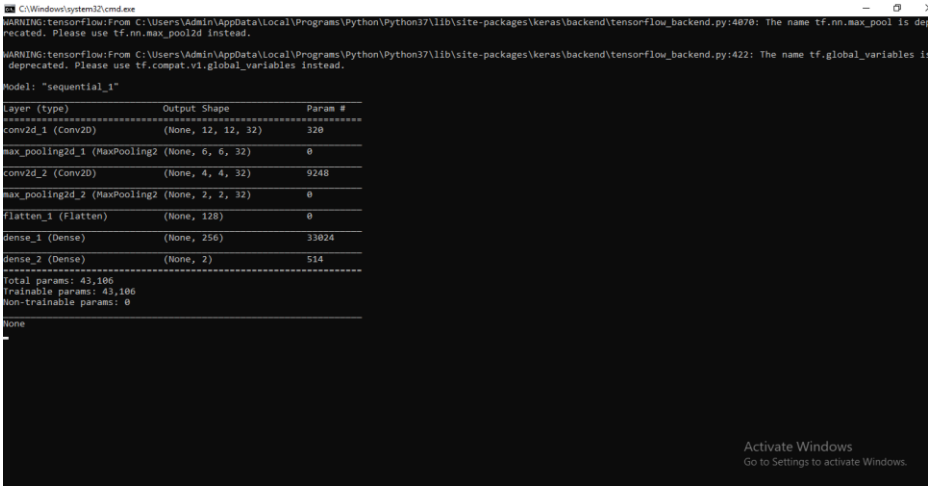


Fig. 9: Layers Description.

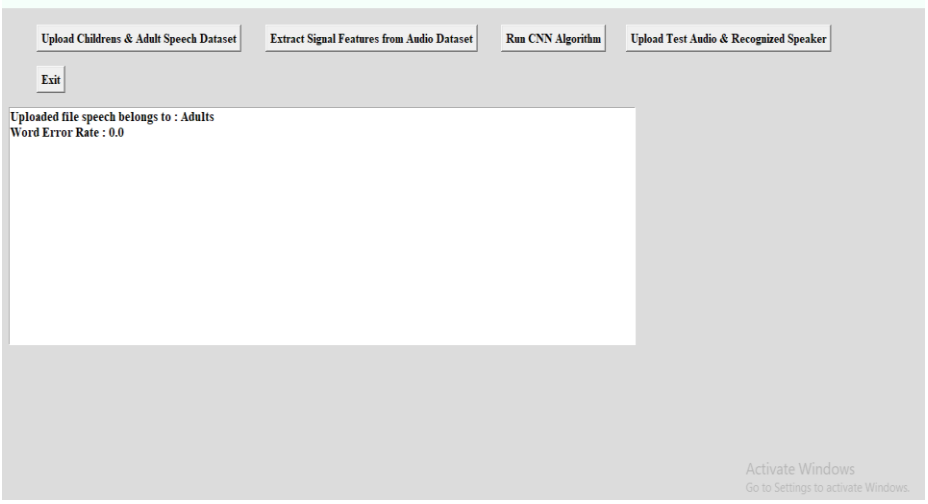


Fig. 10: Predicted outcome.

6. Conclusion

This work 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

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

studies also showed that augmenting children's data with adult speech data could improve the system further. An analysis of the trained CNNs revealed that the CNNs learn to model formant in formation invariant to the acoustic differences in children and adult speech.

REFERENCES

[1] 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] R Cole, P Hosom, and B Pellom. "University of Colorado prompted and read childrens speech corpus". Tech. rep. Technical Report TR-CSLR-2006-02, University of Colorado, 2006.

[10] D. Povey. "The Kaldi Speech Recognition Toolkit". In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 788–798, 2011.

[12] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Proc. Interspeech, 2015.

[13] R. Tong, L. Wang and Bin Ma. "Transfer learning for children's speech recognition",2017 International Conference on Asian Language Processing (IALP), 2017, pages=36-39

[14] S. P. Dubagunta, S. Hande Kabil and M. Magimai.-Doss, "Improving Children Speech Recognition through Feature Learning from Raw Speech Signal", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5736-5740, doi: 10.1109/ICASSP.2019.8682826.

[15] F. Wu, L. P. Garc´ıa-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network", in Proc. Interspeech, 2019.

[16] C. X. Qin, W. L. Zhang, and D. Qu, "A new joint CTC-attention based speech recognition model with multi-level multi-head attention", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2019, no. 1, p. 18, 2019.

ISSN:0975 -3583,0976-2833 VOL10, ISSUE 04, 2019

[17] P. G. Shivakumar and P. Georgiou. "Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations". Comput Speech Lang. 2020 Sep; 63:101077. doi: 10.1016/j.csl.2020.101077. Epub 2020 Feb 18. PMID: 32372847; PMCID: PMC7199459.

[18] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig, and M. L. Seltzer, "Transformer-based acoustic modeling for hybrid speech recognition", in Proc. ICASSP, 2020.

[19] M. A. Hassan, A. Rehmat, M. U. Ghani Khan, M. H. Yousaf, "Improvement in Automatic Speech Recognition of South Asian Accent Using Transfer Learning of DeepSpeech2", Mathematical Problems in Engineering, vol. 2022, Article ID 6825555, 12 pages, 2022. https://doi.org/10.1155/2022/6825555

[20] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq and H. Hamam. "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review". Appl. Sci. 2022, 12, 4419. https://doi.org/10.3390/app12094419