Development of End – to – End Encoder - Decoder Model Applying Voice Recognition System in Different Channels

Syed Jamalullah. R, L.Mary Gladence, V. Rajalakshmi

Abstract — the dissimilarity in recognizing the word sequence and their ground truth in different channels can be absorbed by implementing Automatic Speech Recognition which is the standard evaluation metric and is encountered with the phenomena of Word Error Rate for various measures. In the model of 1ch, the track is trained without any preprocessing and study on multichannel end-to-end Automatic Speech Recognition envisaged that the function can be integrated into (Deep Neural network) – based system and lead to multiple experimental results. More so, when the Word Error Rate (WER) is not directly differentiable, it is pertinent to adopt Encoder – Decoder gradient objective function which has been clear in CHiME-4 system. In this study, we examine that the sequence level evaluation metric is a fair choice for optimizing Encoder – Decoder model for which many training algorithms is designed to reduce sequence level error. The study incorporates the scoring of multiple hypotheses in decoding stage for improving the decoding result to optimum. By this, the mismatch between the objectives is resulted in a feasible form to the maxim. Hence, the study finds the result of voice recognition which is most effective for adaptation.

Keywords— Multichannel system, Word Error Rate (WER), Automatic Speech Recognition (ASR).

I. INTRODUCTION

Encoder – Decoder [1,10] were categorized in two phases, training and evaluation. The differences between these two phases were a conditional probability. In Encoder – Decoder model, Neural Networks (NN) helps by representing sequence of input signals into sequence of output signals. There are two issues encountered in this model. The first is the deficit of a common method for developing the model on evaluation metric such as Word Error Rate (WER). In second, the gap between the prefixed tokens used in the traditional training and evaluation stages. The framework of Encoder – Decoder model [1][2] outlines a multichannel End-to-End Automatic Speech Recognition architecture and also assimilates multichannel speech enhancement components which converts multichannel speech signal to text.

In this study, we came across a multipath adaptation scheme for the multichannel End-to-End (ME2E) Automatic Speech Recognition architecture [1][2] where input data are evaluated by its effectiveness and transferred through an unprocessed noisy speech path compared to single path adaptation scheme, where data's are transferred through a speech enhanced path. Also, implementation of speaker adaptation is recommended in multichannel ASR system.

In the review of previous studies related to ASR, we came across development of ASR system in a single channel setup [2] without voice enhancement. In more pragmatic situations, voice inputs to ASR system were polluted by reverberation and background noises. So, it is important to study the serviceability of End-to-End architecture in a multichannel system.

On the other view, the improvement in the performance of noisy ASR problems can be developed by C H i M E – 4 techniques [2][3] which implements

(i)Beamforming method for processing multichannel signals

(ii) Usage of short language model such as LSTM – based RNN language model.

(iii) Implementation of speaker adaptation techniques.

In our approach, we examine that Voice Recognition System will be more effective for adaptation, when we escalates sequence level evaluation metric to reduce sequence level errors.Ease of Use

II.OVERVIEW OF MULTICHANNEL END-TO-END ASR ARCHITECTURE

In this proposed study, architecture of (ME2E) ASR system is described which assimilates and combines the entire module into single network architecture. This architecture includes, attention based Encoder-Decoder which stands as an ASR part and Neural beam former stands as an enhancement part [2][3]. To connect these components, feature extraction function is used. In the meanwhile of short-time Fourier transformation (STFT) feature [2][3], the sequence recorded at Cth channel is brought to existence for evaluation [2][3][4].

Let $X^c = \{ X_t^c \in C^f \mid t=1,\ldots,T \}$ **……(1)**

Where X_t^c = F – Dimension STFT feature

t = Input time

T = Input sequence length

C = No. of Channels

Revised Version Manuscript Received on 10 September, 2019.

Syed Jamalullah.R , Research Scholar, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India
(Email: syedjamalullahr@gmail.com)
L.MaryGladence, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.
(Email: marygladence.it@sathyabama.ac.in)
V.Rajalakshmi, School of Computing ,Sathyabama Institute of Science and Technology, Chennai, India.
(Email: rajalakshmi.it@sathyabama.ac.in)

Given a multichannel noisy speech inputs $=\{X^c\}^c_{c=1}$ ..... **(2)**

And a posteriori probabilities for the output is considered in this way, which follows

$P(Y \mid \{X^c_{c=1;}\ \wedge_{allFeat}) = \prod_n P(Y_n \mid \{X^c\}^c_{c=1}; Y_1 : n\text{-}1; \wedge_{allFeat})$ ..... **(3)**

$\hat{S}$ = Neural Beam Former $(\{X_c\}^c_{c=1}, \wedge$**bm**$)$

$\hat{O}$ = Feature (Connectionist between Neural Beam Former and the Encoder-Decoder network)

H = Encoder $(\wedge_{enco})$

C_n= Attention $(\wedge_{atten})$

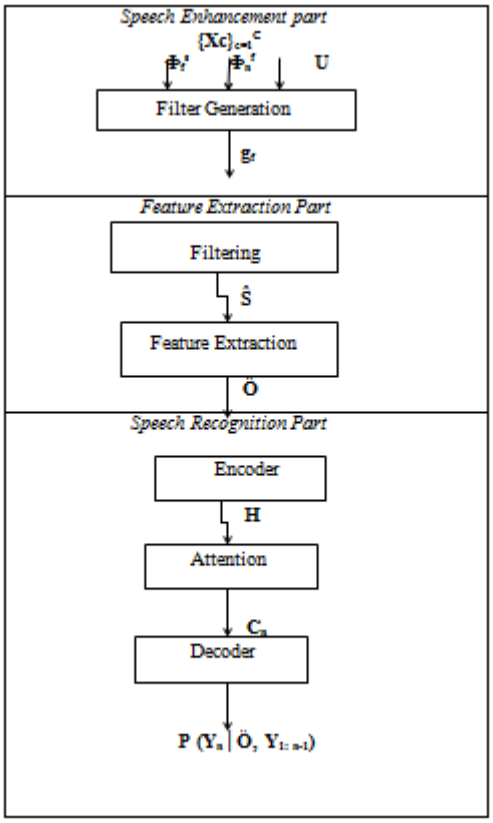

Fig 1: ME2E ASR Architecture

The above architecture is partitioned into three parts (i) Speech enhancement part, which relies on Neural Beam former (ii) Feature extraction part (iii) Speech recognition part, which relies on Encoder-Decoder system.

Initially, Neural Beam Former [3][4] evaluates the filter (gf) through the estimation of three statistics. (i) The cross channel matrix for speech ΦfS (ii) Matrix for noise ΦfN (iii) The reference microphone vector "U" and through linear filtering mechanism [4], it assimilates the multichannel noisy speech {Xc}cc=1 into single channel speech signal (Ŝ). The next step is Features which changes the enhanced STFT feature sequence Ŝ to Ô. Moreover, Encoder transfers the sequence Ô to H. Attention combines all encoder output H to Cn. At last, Decoder, update hidden state and provides output sequence Y based on recursive system.

III.OVERVIEW OF MULTICHANNEL ADAPTATION SCHEME

A. Multichannel Adapting Multipath Scheme

The multichannel End to End ASR system is completely based on neural networks [5][6]. In this system, the data's are transmitted through an unprocessed noisy speech path. This system comes under multipath adaptation scheme. On the other hand, data's are transmitted through a speech enhanced path in single path adaptation scheme. The act of attention based Encoder-Decoder networks remains strong against noisy speech and becomes a powerful ASR back end component [5][6][7].

B. Decoding Procedure Using CHiME–4 Corpus

The CHiME – 4 corpuses is a well-known multichannel noisy ASR standard system. It consists of exaggerated and real speech data recorded using a tablet device with 6-channel microphones [3][4][5] in four different situations. (1) Cafe (CAF), (2) Street Junctions (STR), (3) Public Transportation (BUS) and Pedestrian Area (PED). These data's were grouped together in three subsets (i) Training Set (ii) Development Set (iii) Evaluation Set. The conditions for evaluation follow traditional method. The main difference is we can add additional corpuses for the betterment. The LSTM (Long Short Term Memory) is an artificial Recurrent Neural Network (RNN) which process single data points (such as images) and also entire amount of data (such as video). This system is applicable to tasks in ASR system for decoding procedure.

IV.EVALUATION METRIC WITH ADAPTATION TECHNIQUES& RESULTS

A. Factual Data

For contrivance data's in different scenarios, Stochastic Gradient Descent (SGD) [6][7] system is used. By using this system, Word Error Rate and Character Error Rate were concise with and without using External Language Model, to which 43.3 of WER and 22.4 of CER is the solution for not using External Language Model. Later on, after applying External Language model, it shows a better result reducing WER from 43.3 to 29.7 and CER 22.4 to 17.9.

TABLE1: WORD ERROR RATE AND CHARACTER ERROR RATE OF SPEECH INDEPENDENT SYSTEM OF FACTUAL DATA

	WER	CER
External Language Model Used	29.7	17.9
External Language Model not used	43.3	22.4

B. Speech Adapting System

In this system, five different assignments of the adaptation model parameters [8] were used. (i) Whole network (∧adapt = ∧allFeat), Neural Beam former (∧adapt =∧bm), Encoder (∧adapt = ∧enco), Attention mechanism (∧adapt =∧atten) and Decoder (∧adapt = ∧deco). From this adaptation technique, we came across factual data's

TABLE2: WORD ERROR RATE OF SPEECH ADAPTATION SYSTEM OF FACTUAL DATA

∧adapt	Single Path	Multi Path
{∧allFeat}	26.2	25.4
{∧bm}	27.1	N / A
{∧enco}	25.0	24. 3
{∧atten}	27.4	26.1
{∧deco}	25.9	24.8

C. Environment Adapting System

Neural Beam formers are more effective for environment adapting system than speech adaptation because a noise characteristic doesn't depend on speech but rather speech environment [8][9].

TABLE3: WORD ERROR RATE OF ENVIRONMENT ADAPTED SYSTEM FOR FACTUAL DATA

∧adapt	Scheme	WER
{∧bm}	Single path	27.6
{∧enco}	Multi path	26.8

From the above table, we noted that, the performance of environment adapted system is much better than speech adapting system.

V. CONCLUSION

It is concluded that speech recognition in enhancing the End-to-End Encoder-Decoder model with speech recognition system in attention-based-Encoder-Decoder system and CHiME – 4 corpuses proved the effectiveness of speech recognition executed among multichannel End-to-End Automatic Speech Recognition. It has led to the feasibility of using multichannel reduced Word Error Rate. More so, the Encoder (∧enco) is considered better among the channels.

REFERENCES

1. Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, NavdeepJaitly, Andrew Senior, Vincent Vanhoucke,PatrickNguyen,TaraNSainath,etal., "Deepneural networks for acoustic modeling in speech recognition: The sharedviewsoffourresearchgroups," IEEESignalProcessing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
2. Jan Chorowski, DzmitryBahdanau, DmitriySerdyuk, Kyunghyun Cho, and YoshuaBengio, "Attention-based modelsforspeechrecognition," inAdvancesinNeuralInformation Processing Systems (NIPS), 2015, pp. 577–585.
3. DzmitryBahdanau, Jan Chorowski, DmitriySerdyuk, Philemon Brakel, and YoshuaBengio, "End-to-end attentionbased large vocabulary speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4945–4949.
4. Alex Graves and NavdeepJaitly, "Towards end-to-end speech recognition with recurrent neural networks," in International Conference on Machine Learning (ICML), 2014, pp. 1764– 1772.
5. TsubasaOchiai, Shinji Watanabe, Takaaki Hori, and John R Hershey, "Multichannel end-to-end speech recognition," in International Conference on Machine Learing (ICML), 2017.
6. H.Sak,A.Senior,K.Rao,O. ̇Irsoy,A.Graves,F.Beaufays, and J. Schalkwyk, "Learning acoustic frame labelingforspeechrecognitionwithrecurrentneuralnetwork s," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 4280– 4284.
7. M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition," Proceedings of International Conference on Spoken Language Processing, INTERSPEECH, vol. 5, no. 3, pp. 2406–2409, 2006.
8. T. Hori, Y. Kubo, and A. Nakamura, "Real-time onepass decoding with recurrent neural network language model for speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 6364–6368.
9. D. Bahdanau, D. Serdyuk, P. Brakel, N. R. Ke, J. Chorowski, A. Courville, and Y. Bengio, "Task loss estimation for sequence prediction," International Conference on Learning Representation Workshop, pp. 1–13, 2016.
10. Y. BevishJinila , K. Komathy (2013)," A privacy preserving authentication framework for safety messages in vanet", 4th International Conference on Sustainable Energy and Intelligent System (SEISCON 2013), December 12-14, 2013, pp. 456-461, IET.