一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2020-78,SIP2020-109,SP2020-43(2021-03)

Comparison of End-to-End Models for Joint Speaker and Speech Recognition

Kak SOKY[†‡], Sheng LI[‡], Masato MIMURA[†], Chenhui CHU[†], and Tatsuya KAWAHARA[†]

[†] Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[‡] National Institute of Information Communications and Technology (NICT)
Kyoto, Japan
soky@sap.ist.i.kyoto-u.ac.jp

Abstract In this paper, we investigate the effectiveness of using speaker information on the performance of speaker-imbalanced automatic speech recognition (ASR). We identify major speakers and combine other speakers who have a small size of speech, and make a systematic comparison of three methods that use speaker information for ASR including speaker attribute augmentation (SAug), multi-task learning (MTL), and adversarial learning (AL). We conduct experiments on a large spontaneous speech corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC) and an open Khmer text-to-speech corpus. As a result, we find that the use of speaker clustering information improves ASR performance including new speakers. Moreover, AL achieves better performance and more robustness in the speaker-independent setting compared to the other methods. It reduces errors of the baseline model by 4.32%, 5.46%, and 16.10% for the closed test, open test, and out-of-domain test, respectively.

Key words End-to-End, Speech recognition, Speaker recognition, Khmer language, Low-resource, Speech attribute, Multi-task, Adversarial learning

1 Introduction

In the last decade, automatic speech recognition (ASR) has attracted much more attention due to the advancement of deep learning techniques and computing resources. This advancement also enables the shift from conventional architectures such as the hybrid GMM-HMM [1] and DNN-HMM [2] to the End-to-End (E2E) systems [3, 4, 5, 6, 7]. The ensemble of acoustic model (AC), pronunciation model (PM), and language model (LM) into a single neural network model by the E2E model has solved a complex problem of sequence labelling between speech input and output label, and achieved promising results. The E2E model has also become a gateway to open the door for the researchers to work on low-resource languages that are short of natural language processing resources.

Generally, E2E model requires a large speech corpus, but in many low-resource languages, this assumption does not hold and speakers are imbalanced in that there are often dominant speakers. This is true in many cases in major languages such as meetings and courts, in which there are a limited set of speakers and the number of utterances is imbalanced. Therefore, we need to compensate for the speaker imbalance by either adapting the model or normalising the speaker-dependent features to each speaker.

In this context, many techniques have been proposed. A simple method is speaker embedding that is used to improve the ASR, but it does not explicitly use the supervision of the speaker information [8, 9, 10]. Multi-task learning (MTL) is introduced to unify training of transcribing the speech and identifying the speaker simultaneously by sharing the same speech feature extraction layers [11, 12, 13]. Adversarial learning (AL) adopts a similar architecture to MTL, but learns a speaker-invariant model to make it more generalised to the new speakers by reducing the effects of speaker variability [14, 15, 16, 17, 18]. Most recently, speech attribute augmentation (SAug) is introduced to embed the speaker attribute tags into the training label and generates those tags together with the transcription in a single decoder [19, 20, 21].

Y. Adi et al [22] compared MTL and AL on the Wall Street Journal (WSJ) dataset, showing that MTL and AL did not impact on the word error rate (WER) in this task, but they are promising only on the letter error rate after adding more speaker-labelled data. This suggests that these methods are not effective for a small amount of speech by many speakers.

In this work, we address the speaker-imbalanced problem of a large vocabulary spontaneous speech corpus using the Extraordinary Chambers in the Courts of Cambodia (ECCC). We identify major speakers and cluster other speakers. Then we make a systematic comparison of three approaches that use speaker information: MTL, AL and SAug [20]. We conduct experiments on both in-domain and out-of-domain dataset.

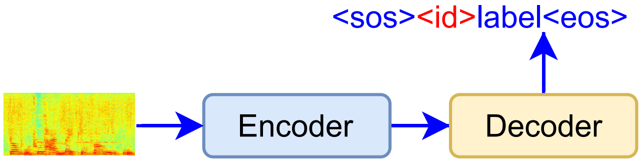

Figure 1: SAug using speaker ID

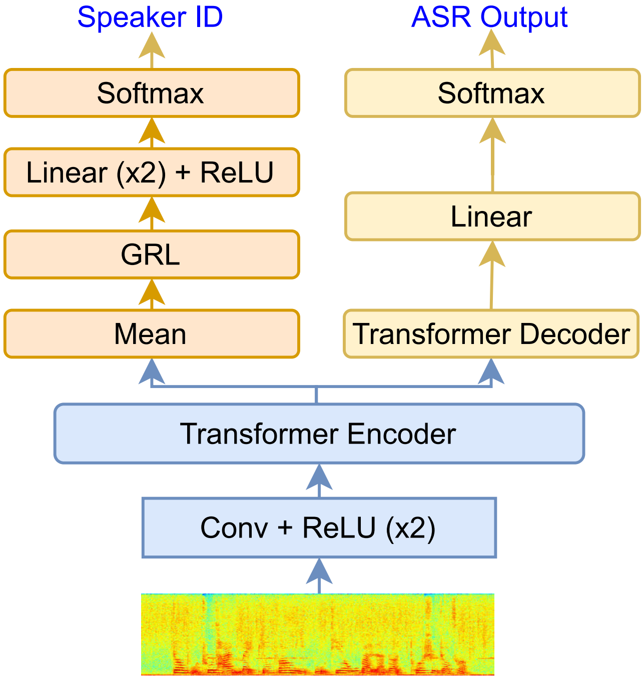

Figure 2: Architecture of MTL and AL. In AL, GRL reverses the gradient at the backpropagation stage, while in MTL, this layer is not used.

2 Speaker and Speech Recognition

Speaker recognition (SRE) and ASR are complementary tasks to each other. When we identify speakers, it is often easy to recognise their speech. Thus speaker adaptation is conducted. A speaker ID can be used in joint training of ASR and SRE, but this is not a common practise when there are a large number of speakers in the dataset. On the other hand, these approaches are expected to be effective for the speaker-imbalanced setting.

In this section, we briefly present the methods used in this work: SAug in Figure 1, MTL and AL in Figure 2.

2.1 Speech Attribute Augmentation (SAug)

The speech attribute is analogous to language ID in a multilingual model. It can be speaker ID, gender, and age. They are placed in front of the token sequence of each utterance. Given a sequence of n acoustic features $X_i = \{x_1, ..., x_n\}$ where each x is a feature vector, a model must produce sequence tokens $Y_i = (s, y_1, ..., y_m)$ where s is a speech attribute and y is a sequence of vocabulary tokens.

The network is trained to output the attribute label at the beginning of decoding, thus we do not have to prepare classifiers for these attribute explicitly.

2.2 Multi-Task Learning (MTL)

Given a sequence of acoustic features X_i, a model must produce a sequence of vocabulary tokens Y_i and a speaker label s separately. This joint recognition is possible when the number of speakers is limited and each speaker has a large amount of data. However, this method is not applied to many speakers with fewer data including unseen speakers, hence we cluster the speakers. Our goal is to make use of these speaker labels to optimise both SRE and ASR losses. Therefore, the loss in MTL is defined as:

$$\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{asr} + \alpha * \mathcal{L}_{sre}, \tag{1}$$

where α is a weight of SRE task.

2.3 Adversarial Learning (AL)

To train an ASR system to be able to recognise a huge number of speakers is not practical. Moreover, the ASR should not solely depend on a specific set of speakers. In this context, AL learns an acoustic representation which is speaker-invariant by incorporating adversarial loss of SRE, which is combined with the loss of ASR. To learn the speaker-invariant, we need to reverse the gradient of backward propagation in gradient reversal layer (GRL) as presented in [14].

Let the parameters θ_{enc}, θ_{asr}, and θ_{sre} denote as the encoder, ASR, and SRE output networks, respectively. The parameters are updated in backward propagation as follows:

$$\theta_{asr} \longleftarrow \theta_{asr} - \epsilon\frac{\partial\mathcal{L}_{asr}}{\partial\theta_{asr}}, \tag{2}$$

$$\theta_{sre} \longleftarrow \theta_{sre} - \epsilon\frac{\partial\mathcal{L}_{sre}}{\partial\theta_{sre}}, \tag{3}$$

$$\theta_{enc} \longleftarrow \theta_{enc} - \epsilon(\frac{\partial\mathcal{L}_{asr}}{\partial\theta_{enc}} - \lambda\frac{\partial\mathcal{L}_{sre}}{\partial\theta_{enc}}), \tag{4}$$

where ϵ is a learning rate and a negative coefficient λ (-λ) is used to reverse the gradient from the speaker classification as shown in Figure 2.

3 Experiments

3.1 Data setup

The ECCC[1] is a trial to prosecute the senior leaders that committed the crimes during the period of Democratic Kampuchea (Khmer Rouge regime) in Cambodia from 1975 to 1979. The trial is subsequently divided into four cases. Presently, the trial is in progress, and only a small part of the resources has been published to the public. For this reason, we choose only the first caseload that started from February 17, 2009 to November 27, 2009. The public hearing has been recorded in a courtroom. Lately, those videos have been published in Youtube[2]. For the proceedings, they have been manually transcribed and published in a digital format at its official website. Each audio has a length in the range of 5 to 150 minutes, and speakers include indicted person, witnesses, judges, clerks, co-prosecutors, lawyers, civil parties, and translators. We have collected 222 recording sessions and then built a large spontaneous speech corpus that consists of 78,944 utterances (about 186 hours) with 28 speakers (6 female) as presented in Table 1.

Due to the imbalanced amount of speech in these speakers, we classify the speakers into a group of 7 (Gr7) and a group of 6 (Gr6). In Gr6, there are a president, an accused, three translators, and other speakers into one, whereas in Gr7, we further classify other speakers into two based on gender. The dataset is split into three sets. For testing, we selected 10 speakers (about 10 hours) with a smaller data size in the corpus as presented in Table 1. Besides the testing, we randomly split into a validation set and training set by 5% and 95%, respectively. The validation set is referred to a closed test (TestClosed) because some speakers are included in the training set but the speaker distribution is matched, while the test set is an speaker open test (TestOpen). We also evaluate with the out-of-domain test set (TestOut) which is a high-quality text-to-speech for Khmer prepared by Google[3]. We present the detail in Table 2.

ID	Gender	Hour	ID	Gender	Hour
Training set					
Pre	M	31	S37	F	3
AC	M	30	S39	M	3
T1	M	34	S42	M	5.5
T2	M	23	S48	F	2
T2	M	17	S49	F	0.5
S08	M	1.5	S55	M	5.5
S14	M	6	S57	F	6
S30	M	1	S58	M	2.5
S35	M	5.5	S80	M	1
Testing set					
S01	M	2	S50	F	1.5
S10	M	1	S53	M	0.1
S19	M	1	S72	M	1.5
S21	M	0.15	S77	M	1.5
S31	M	0.7	S89	F	1.5

Table 1: Data distribution of ECCC: there are five major speakers - a president of the chamber (Pre), an accused person (AC) and three Translators - T1, T2, T3.

Table 2: Data statistics of this work

Data	#utt (#hour)	#character
Train	70,908 (167)	5.67M
TestClosed	3,733 (9)	294K
TestOpen	4,262 (10)	335K
TestOut	2,906 (4)	123K
Total	**81,809 (190)**	**6.43M**

3.2 Baseline E2E-ASR System

We adopt a Transformer-based E2E-ASR system with ESPnet toolkit [23] (e = 6, d = 6, d^{ff} = 1024, d^{head} = 4, d^{att} = 256) for all experiments in this work. The 80-dimensional log Mel-filter bank features which were mean and variance normalised per speaker are extracted with a 10ms frame shift of 25ms windows. Then, we subsample the input features using two-layer time-axis CNN with ReLU activation, d^{att} channels, stride size 2 and kernel size 3. The model was jointly trained with CTC (weight $\alpha = 0.3$) without using any language model. The training was conducted in a single 12GB GPU Titan X (Pascal) with 25000 warmup steps, 32 mini-batch sizes, and 30 epochs.

3.3 SAug System

The configuration of this model is the same as the baseline system except for the output label. We embed speaker IDs as the speech attribute labels to a ground-truth in training. To measure ASR performance, we just remove the beginning attributes from the transcription and calculate the character error rate (CER). We calculate SRE performance with the speaker attribute label.

3.4 MTL and AL Systems

We share the same transformer encoder of ASR and SRE tasks. The ASR decoder is the same as baseline system, but this network takes a mean of the h^{enc} of encoder output and feeds it to 2 fully-connected layers followed by the ReLU activation function and Softmax to generate the speaker label output. The GRL is triggered in the AL system to compute the reversed gradient at the backward propagation phase. In the forward propagation phase, MTL and AL are acted in the same operation.

[1] www.eccc.gov.kh/

[2] www.youtube.com/user/krtribunal/

[3] https://openslr.org/42/

3.5 Results and Discussion

We evaluate the performance of the compared models based on the CER. In Table 3 for SAug, Gr6 outperforms the model using all 28 speakers (Gr28) and Gr7 because the number of female speakers is small. With the Gr6, all compared methods using speaker information improved ASR performance over the baseline. However, SAug is not as effective as MTL and AL. It suggests that it is difficult to train the model in a single decoder. On the other hand, MTL is as effective as AL on the TestClosed because speaker information is matched in training and testing. On the other hand, AL is effective for unseen speakers in TestOpen and TestOut. It reduces the CER of baseline by 4.32%, 5.46%, and 16.10% for the TestClosed, TestOpen, and TestOut, respectively. In TestOut result, we found that most of the errors are attributed to proper nouns (i.e. places, people), which are not covered in the training corpus.

Table 3: ASR performance (CER%)

Model	**Character-based unit**		
	TestClosed	TestOpen	TestOut
Baseline	7.63	11.53	30.00
SAug(Gr28)	7.52	12.37	30.94
SAug(Gr7)	7.42	11.42	26.85
SAug(Gr6)	7.37	11.16	26.20
MTL (Gr6)			
$\alpha = 0.2$	7.31	11.00	25.90
$\alpha = 0.5$	**7.30**	11.08	26.82
$\alpha = 0.7$	7.33	11.18	28.00
AL (Gr6)			
$\alpha = 0.2$	**7.30**	**10.90**	26.95
$\alpha = 0.5$	**7.30**	10.95	**25.17**
$\alpha = 0.7$	7.70	11.50	29.90

Since the speakers in training and TestOpen or TestOut are independent, it is not appropriate to evaluate the SRE performance for them. Therefore, we only calculate the SRE error rate for TestClosed in Table 4. The SRE performance is also improved by SAug and MTL. We also present the confusion matrix of speaker labels for TestOpen and TestOut in Table 5. These results show that the SAug and MTL tend to classify the speakers to most of the trained speakers, while AL classifies all data into the "other" speaker cluster.

Table 4: SRE performance of TestClosed on Gr6

Test	**Speaker error rate (%)**			
	Baseline[4]	**SAug**	**MTL**	**AL**
TestClosed	9.72	**8.81**	9.09	75.16

Table 5: Speaker classification result of Gr6

Model	**#Speaker prediction** ($\alpha = 0.5$)					
	Pre.	**AC**	**T1**	**T2**	**T3**	**other**
TestOpen						
SAug	434	63	296	191	70	3107
MTL	624	58	285	233	69	2993
AL	0	0	0	0	0	4262
TestOut						
SAug	0	0	109	47	3	2747
MTL	9	0	109	138	19	2631
AL	0	0	0	0	0	2906

4 Conclusion

In this work, we address the problem of speaker-imbalanced ASR by identifying major speakers and clustering other speakers who have a small size of speech. The clustering of speakers is efficient and effective for ASR performance. All compared methods using speaker clustering information improve ASR performance in all experiments. More specifically, MTL and AL outperform the SAug and on TestClosed, while AL is robust to new speakers and outperforms others in TestOpen and TestOut sets.

Acknowledgments

We thank Mr. Sei Ueno and Mr. Kouhei Matsuura, from Speech and Audio processing laboratory, Kyoto University for discussing the program implementation and National Institute of Post, Telecoms and ICT (NIPTICT) for giving us the low data of ECCC.

References

[1] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol.77, no.2, pp.257–286, 1989.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech recognition," IEEE Trans. ASLP, 2012.

[3] A. Graves, S. Fernandez, F. Gomez, and J. Shmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proceedings of the 23rd International Conference on Machine Learning, 2006.

[4] A. Graves, "Sequence Transduction with Recurrent Neural Networks," ICML Representation Learning Workshop, Nov. 2012.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,"

[4] The ECCC was built from scratch, and some utterances were missed the speaker information. For that reason, speakers were clustered using x-vector-based.

2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4960–4964, 2016.

[6] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM,” Proc. Interspeech 2017, pp.949–953, 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All You Need,” Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 2017.

[8] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker Adaptation of Neural Network Acoustic Models Using i-vectors,” 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp.55–59, 2013.

[9] T. Ochiai, S. Matsuda, X. Lu, C. Hori, and S. Katagiri, “Speaker Adaptive Training using Deep Neural Networks,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6349–6353, 2014.

[10] W. Chu and R. Chen, “Speaker Cluster-based Speaker Adaptive Training for Deep Neural Network Acoustic Modeling,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5295–5299, 2016.

[11] Z. Tang, L. Li, and D. Wang, “Multi-task Recurrent Model for Speech and Speaker Recognition,” 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.1–4, 2016.

[12] N. Kanda, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Zhou, and T. Yoshioka, “Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers,” Interspeech 2020, pp.36–40, ISCA, October 2020.

[13] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, “Investigation of End-To-End Speaker-Attributed ASR for Continuous Multi-Talker Recordings,” SLT 2021, IEEE, January 2021.

[14] Y. Ganin and V. Lempitsky, “Unsupervised Domain Adaptation by Backpropagation,” Proceedings of the 32nd International Conference on Machine Learning, pp.1180–1189, 07–09 Jul 2015.

[15] Y. Shinohara, “Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition,” Interspeech 2016, pp.2369–2372, 2016.

[16] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, “Speaker Invariant Feature Extraction for Zero-Resource Languages with Adversarial Learning,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2381–2385, 2018.

[17] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B. Juang, “Speaker-Invariant Training Via Adversarial Learning,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5969–5973, 2018.

[18] Z. Meng, J. Li, and Y. Gong, “Adversarial speaker adaptation,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5721–5725, 2019.

[19] L. El Shafey, H. Soltau, and I. Shafran, “Joint Speech Recognition and Speaker Diarization via Sequence Transduction,” Proc. Interspeech 2019, pp.396–400, 2019.

[20] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai, “Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation,” Proc. Interspeech 2019, 2019.

[21] H. Henry Mao, S. Li, J. McAuley, and G. W. Cottrell, “Speech Recognition and Multi-Speaker Diarization of Long Conversations,” Proc. Interspeech 2020, pp.691–695, 2020.

[22] Y. Adi, N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve, “To Reverse the Gradient or Not: an Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3742–3746, 2019.

[23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-End Speech Processing Toolkit,” Proc. Interspeech 2018, pp.2207–2211, 2018.