Robust Speech Recognition based on Multi-Objective Learning with GRU Network

Ming Liu*† and Yujun Wang† and Zhaoyu Yan* and Jing Wang* and Xiang Xie*
* Beijing Institute of Technology, Beijing, China
E-mail: liuming0806@163.com, zhaoyu.yan@outlook.com, wangjing@bit.edu.cn, xiexiang@bit.edu.cn
† Xiao Inc, Beijing, China
E-mail: wangyujun@xiaomi.com

Abstract—**This paper proposes a new scheme to execute the task of speech enhancement (SE) for recognition based on multi-objective learning method which uses three objectives in the gated recurrent unit (GRU) network training procedure. The first objective is the main target for the expected SE task by directly mapping the noisy log-power spectrum (LPS) features to clean Mel-frequency cepstral coefficients (MFCC) features. The second one is an auxiliary target to help improving the main one by learning additional information from the back-end acoustic model (AM). The third one is also an auxiliary target achieved by learning some information from mapping noisy LPS to clean LPS. The two auxiliary structures could help the original structure to optimize the network parameters by correcting the errors. This approach imposes more constraints on direct feature mapping and information passing from the acoustic model to the network, enabling the enhanced network to better serve the AM. The experimental results show that the new multi-objective scheme with joint feature mapping and the posterior probability learning method improves the performance of SE. And this scheme significantly lowers the Character Error Rate (CER) of the AM compared to the baseline deep neural network (DNN) network [1].**

I. Introduction

Speech enhancement is a challenging and important research area for speech signal processing applications such as speech communication and speech recognition, whose performance largely depends on the signal quality [1].Traditional speech enhancement methods such as spectral subtraction [2], Wiener filtering [3], amplitude estimation [4] are widely popular due to their low complexity in computation and perform very well in processing stationary noise, but not in unknown non-stationary noise. With recent development of deep learning-based speech processing [5], supervised deep learning approaches have been shown to generate enhanced speech with good qualities [6] based on a mass of known data.Also, DNN with multi-objective learning [7] performs better in predicting LPS characteristics and improving speech quality in SE, and can achieve promising results in challenging real-world speech applications like speech enhancement for speech recognition. But in this study, the focus is on improving speech recognition accuracy. And that DNN is not the only way to improve the accuracy of speech recognition, the gated recurrent unit (GRU) is more suitable for sequence modeling tasks [8].

In this paper, a multi-objective learning method with three targets is proposed to optimize a joint objective function in the task of SE for speech recognition. The objective function includes not only the error of the primary MFCC features, but also the secondary target errors of the posterior probability obtained from the AM and the third target errors of the clean features, such as LPS. The posterior probability uses cross entropy (CE) as the optimization function, while the features use mean square error (MSE) as the optimization function. In the main task, we map the LPS features to the MFCC features. In the auxiliary tasks, information from the acoustic model and the clean LPS features is used to fix the main target to make it better fulfill the requirements of the back-end acoustic model. The proposed method transforms a simple regression task into two tasks, i.e. classification and regression tasks.

The experimental results imply that the multi-objective learning method significantly lowers the CER as compared with the DNN feature mapping baseline when tested on the AM.

II. Multi-objective Learning Speech Enhancement with GRU Network

In [9], recurrent neural network (RNN) is adopted as a mapping function to predict the clean MFCC features from the noisy LPS features. The relationship between the clean and noisy speech features can be well learned because nearly no-assumptions were imposed during the training process. Since the training algorithm uses gradient-based back-propagation through time (BPTT) in the cassical RNN. When the time is long, the residual index that needs to be returned will drop, leading to slow update of the network weight. To solve this problem, long short-term memory (LSTM) [10] network and GRU network have been introduced. GRU parameters are fewer than LSTM and therefore easier to converge while LSTM shows better performance with a large number of data sets. Since small data are used in the experiment, the GRU network structure is adopted to simplify the calculation.

MSE and CE are used to update the weights in the Shared GRU network,

[1] This work is done when Ming Liu was an intern in Speech Group, Xiaomi Corporation, Beijing, China.

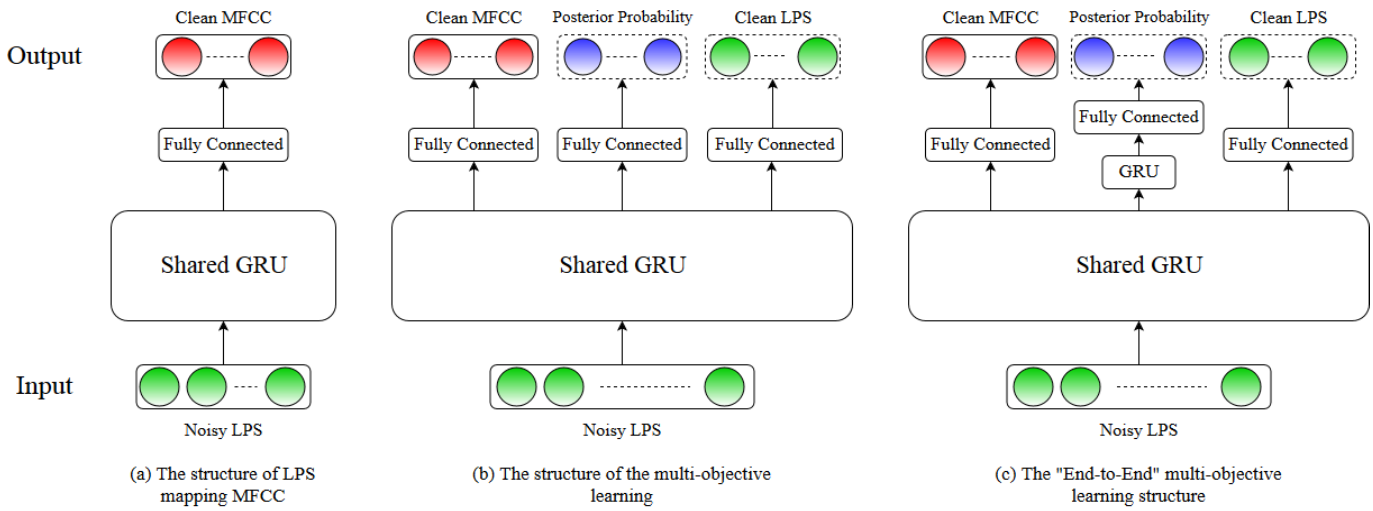


Fig. 1. Three different structures of speech enhancement for recognition. (a) Simple feature mapping enhancement, (b) multi-objective learning simultaneous output enhancement and (c) multi-objective learning similar to End-to-End structure.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} w(y_i - \hat{y}_i)^2 \tag{1}$$

MSE is the mean of the sum of the squares of the corresponding predicted data and original data points. In speech enhancement, the MSE between the target features and the predicted features is always used as the objective function. y_i is the clean MFCC feature, $\hat{y}_i$ is the predicted MFCC feature, N is the total number of data, and $w = 0.5$. It can be seen that the closer the MSE is to zero, the better the model selection and fitting, as well as the data prediction will be.

$$CE = -\sum_{i=1}^{M} z_i \log(p_i) \tag{2}$$

CE is mainly used to measure the difference between two probability distributions. The loss function calculates the cross entropy of the prior information of the training data and the posterior information of the target data to eliminate noise interference. z_i is the posterior probability of the clean data through AM, p_i is the probability of the predicted data, and M is the number of categories.

Multi-objective learning is proposed to jointly predict the primary MFCC features together with the posterior probability obtained by the AM and other continuous features, such as LPS, to enhance learning as follows,

$$\begin{aligned} E &= MSE + CE + MSE^{*} \\ &= \alpha * \frac{1}{N}\sum_{i=1}^{N} w(y_i - \hat{y}_i)^2 \\ &+ \beta * \sum_{i=1}^{M} z_i \log(p_i) \\ &+ \gamma * \frac{1}{N}\sum_{i=1}^{N} w(q_i - \hat{q}_i)^2 \end{aligned} \tag{3}$$

E represents the new target optimization function, which is composed of the MSE calculated by two different features and CE calculated by the posterior probability. q_i is the clean LPS features; $\hat{q}_i$ is the predicted LPS features. In addition, $\alpha > \beta > \gamma > 0$, α, β and γ are the coefficients of the objective functions in the overall optimization function, respectively, used to control the proportion of each target.

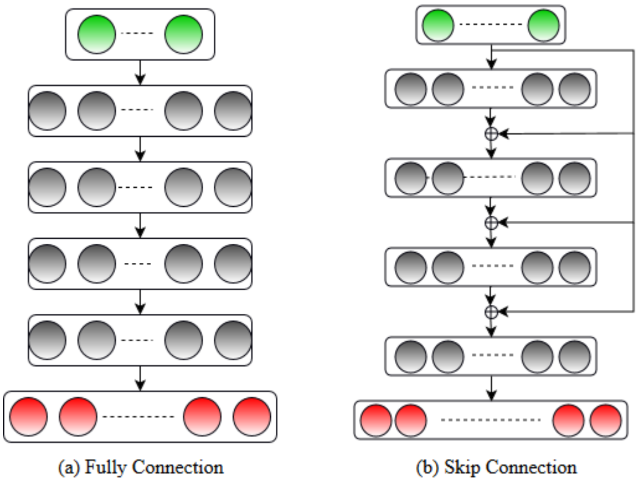
Fig.1 presents the structure of the proposed multi-objective learning based on the simple feature mapping, which is added posterior probability and other features during training. Compared with the former two structures in Fig.1, the output of the posterior probability task is added with the GRU network similar to the End-to-End structure [11], [12]. The method of predicting the posterior probability can promote the clean MFCC. And assisted feature learning can also complement the use of shared GRU in feature information. The output of the shared GRU changes the feature size of the output through the fully connected layer. Overall, multi-objective learning can improve the generalization ability of feature estimation and the matching degree of the acoustic model.

A. *Feature Mapping based speech enhancement*

MFCC is one of the most popular speech features used in speech recognition [13] and speaker recognition [14]. In the front-end enhancement module, the output of the enhancement module usually chooses the MFCC feature because of the input requirements of the acoustic model. The front-end network structure mainly uses the skip connection [15]. Which is implemented on the basis of full connection and can increase the input of each layer and the generalization ability, as is shown in Fig. 2.

B. *Automatic Speech Recognition*

Recently, DNN-LSTM structure has been widely used in speech recognition as acoustic models [16] and shown to be able to preserve long-term temporal information in various



(a) Fully Connection (b) Skip Connection

Fig. 2. Two different network connections.

tasks [17]. In this case, the follow-up experiment will use this acoustic model as the experimental acoustic model. A trigram language model (LM), which was trained with more than 100,000 words in the vocabulary, was used for decoding in the experiments. In order to ensure the normal recognition rate and obtain the required posterior probability, this study adopts XiaoMi TV recognition model as the acoustic model, whose CER is 10.79% in the Kaldi open aishell-1 test set.

C. Multi-objective Learning

The front-end enhancement network only learns the correlation between the feature mapping information and is not associated with the acoustic model. Multi-objective learning can unite both the feature mapping and acoustic model without considering the joint learning of reinforcement and recognition models.

The posterior probability is the probability of correction after obtaining the information about the future results. Which can be determined by all the data about the natural state. Since it makes full use of the prior knowledge and the observed historical time variable information, it is a more reasonable state decision method. Both operations that map noisy LPS features to clean MFCC features and clean LPS features, as well as the posterior probability are jointly predicted, so that some information is removed compared to the independent prediction. In order to retain as much information as possible, we also selected the LPS feature to be trained.

III. Experimental results and Analysis

In this work we conducted on waveforms with 16kHz sampling rate. The training and the test data are pre-processed by adding noise after reverberation, with the reverb data coming from Povey's paper [18]. The reverb contains real and simulation reverb, and the simulation reverb have three types: small, medium and large room. The impulse response reverberation time 60 (RT60) in the room is less than 500 ms. Among 110 noises, 100 noises were used in [19] (The noise types include traffic, machine, home, and bell, etc.) and 10 noises were used in [20] (The noise types include pink, white, street and other noises). We randomly extracted 3000 real room impulse response (RIR) and 100 noises to be added to the training set and the test set named s̈im1.̈ The signal noise ratio (SNR) was randomly generated between 10 and 30 dB. In order to compare the effects of different test sets more intuitively, we extracted the other 1000 real RIR and the rest 10 noises were added to the test sets just like sim1 which is named s̈im2.̈

Network	Reverberation				Noise
	Small	Medium	Large	Real	
Clean	11.07				
None	30.03	33.45	43.67	21.58	18.52
DNN	18.87	22.69	31.43	16.96	14.77
GRU	16.54	19.23	25.19	13.24	13.26
DNŃ	17.11	19.85	27.66	14.38	13.45
GRÚ	15.68	18.47	24.32	12.67	12.78

TABLE I
CERs (%) of four different networks in different environments. DNN represents a network with fully connection, and DNŃ represents a network with skip connection. GRU and GRÚ are the same.

The training set used one hundred thousand short command statements data from the Xiaomi TV Data001 (about 100hrs). The network built on Tensorflow consisted of a single-layer fully-connected output layer and four shared GRU layers with 1024 nodes on each GRU layer. The training of acoustic models and the extraction of feature data are all based on Kaldi implementations. As for feature extraction, the frame length was set to 25 ms with a frame shift of 10 ms.

A. LPS mapping to MFCC

In Table I, we compared the character error rates of the fully connection network and the skip connection network. As well as the DNN baseline and GRU network. Reverberation and noise are added to the test set respectively rather than mixed together. The results show that the skip connection network has improved the enhancement effect and GRU network performs better than DNN. So both GRU and the skip connection networks are used in the subsequent experiments.

B. Joint Prediction of Posterior Probability and LPS

1) The Multi-objective Learning Structure: On the basis of the experimental LPS mapping to the MFCC, we added the posterior probability and the LPS feature branch to the output section as shown in Fig.1. The difference between these two structures is that the output of the GRU is added with a fully-connected layer to change the feature dimension of the output. The output features are compared to the posterior probability of clean speech and the LPS feature of clean speech. During the training process, the information at the moment before output was propagated back to the shared GRU layer to update the GRU layer. In the decoding process, only the MFCC feature was selected as the network output. Finally, the enhanced MFCC features were fed into the back-end acoustic model.

Method	Test	
	sim1	sim2
Clean	11.07	11.07
None	27.83	26.69
FM(DNN)	20.13	22.86
FM(GRU)	19.34	20.56
MOL	17.81	18.84
MOL + ETE	17.75	18.80

TABLE II
CERS (%) COMPARISONS BY PREVIOUS MAPPING BASED NETWORK, MULTI-OBJECTIVE LEARNING NETWORK AND THE END-TO-END NETWORK. FM(DNN) MEANS WE USE LPS MAPPING TO THE MFCC WITH DNN. FM(GRU) MEANS WE USE LPS MAPPING TO THE MFCC WITH GRU. MOL MEANS THE MULTI-OBJECTIVE LEARNING STRUCTURE. MOL + ETE MEANS THE END-TO-END MULTI-OBJECTIVE LEARNING STRUCTURE.

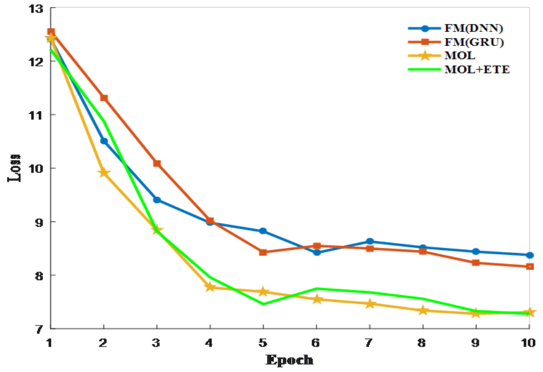


Fig. 3. The loss of different networks

2) The End-to-End Multi-objective Structure: In order to better compare the performance of feature mapping and multi-objective learning, we added another structure: adding several GRU layers between the fully connected layer and the Shared GRU layer, and then training them as a whole. This structure is called an End-to-End multi-objective learning structure. Compared to the previous multi-target structure, the GRU layer is added to ensure that the features are output at the same level, and the posterior probability is finally output. The effect of adding a layer is similar to that of an acoustic model. The newly added network layer and the shared layer are combined to form an end-to-end multi-objective structural system. The dimension of the posterior probability of the new structure output is constrained, to make the probability dimension of the output of the two multi-objective structures be the same.

C. Overall Performance Comparison

Table II shows a comparison of the four different network structures in two different test sets. Compared with the noise speech results, the recognition rate of the benchmark enhanced network has increased by ten percentage points. Compared with the benchmark feature mapping network, the network recognition rate using multi-objective learning has increased by nearly two percentage points, and the CER has dropped to 17.75% on the simulated test set. The loss of training is shown in Fig. 3.

The experimental results show that the multi-objective learning method can significantly improve the recognition rate. At the same time, the End-to-End multi-objective network has the best performance in the four networks. From the results show in Table II, although the test set in the training set is better than the out-of-set test set. However, the out-of-set test set CER has also been improved, indicating that this scheme has better generalization ability.

IV. Conclusion

In this paper, multi-objective learning is proposed to improve GRU training for speech enhancement in recognition. With the posterior probability of acoustic model and the LPS features added to the objective function, this method can better estimate the clean MFCC. The posterior probability carries more information from the back-end acoustic model, and LPS complements the independent prediction characteristics that are lacking in joint prediction. With the above findings, this scheme significantly lowers the CER of the AM compared to the baseline DNN network. In the future, we will further explore whether other identifying information can be added to the proposed scheme.

V. Acknowledgement

The authors would like to thank the Xiaomi Speech team. We also thank Haitong Zhang, Juan Zhao, Yunchao He and Jian Li for their helpful comments and everyone who made Kaldi and tensorflow. This work is partly supported by Natural Science Foundation in China (No. 61571044, No. 11590772, and No.61473041).

References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 416–426, 2013.

[3] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 8, pp. 1285–1295, 2014.

[4] B. Li, Y. Tsao, and K. C. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition." in *Interspeech*, 2013, pp. 3002–3006.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[7] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.

[8] K. Irie, Z. Tuske, T. Alkhouli, R. Schluter, and H. Ney, "Lstm, gru, highway and a bit of attention: an empirical overview for language modeling in speech recognition," RWTH Aachen University Aachen Germany, Tech. Rep., 2016.

[9] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[12] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[13] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on speech and audio processing*, vol. 7, no. 5, pp. 525–532, 1999.

[14] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2006.

[15] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[16] F. A. Gers and E. Schmidhuber, "Lstm recurrent networks learn simple context-free and context-sensitive languages," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1333–1340, 2001.

[17] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neuralhv networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[18] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE*, 2017, pp. 5220–5224.

[19] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[20] A. Varga, "The noisex-92 study on the effect of additive noise on automatic speech recognition," *ical Report, DRA Speech Research Unit*, 1992.