Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

Multichannel Speech Enhancement of Target Speaker Based on Wakeup Word Mask Estimation with Deep Neural Network

Chol Nam Om
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: ocniit@163.com
Hyok Kwak
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: kh_iit@163.com
Chong Il Kwak
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: kcihrdc@126.com
Song Gum Ho
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: hsghrdc@126.com
Hyon Gyong Jang
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: jhgroom2@126.com

----------**ABSTRACT**----------
In this paper, we address a multichannel speech enhancement method based on wakeup word mask estimation using Deep Neural Network (DNN). It is thought that the wakeup word is an important clue for target speaker. We use a DNN to estimate the wakeup word mask and noise mask and apply them to separate the mixed wakeup word signal into target speaker's speech and background noise. Convolutional Recurrent Neural Network (CRNN) is used to exploit both short and long term time-frequency dependencies of sequences such as speech signals. Generalized Eigen Vector (GEV) beamforming estimates the spatial filter by using the masks to enhance the following speech command of target speaker and reduce undesirable noise. Experiment results show that the proposal provides more robust to noise, so that improves the Signal-to-Noise Ratio (SNR) and speech recognition accuracy.

Keywords - **multichannel speech enhancement, wakeup word, mask estimation, beamforming, deep neural network (DNN).**

Date of Submission: May 17, 2023 Date of Acceptance: June 13, 2023

---

1. INTRODUCTION

Recently, in the field of speech signal processing, the efficiency of speech enhancement is highly raised by combining mask estimation based on Deep Neural Network (DNN) and acoustic beamforming [1], [2].

Speech enhancement is very necessary front-end of various speech applications such as automatic speech recognition, smart speaker and smart elevator. Recently, these smart devices are widely used in our ordinary life but non-stationary noise surrounding us is challenging to improve their performance. So researchers prefer microphone array to capture spatial information and to steer to target speaker [3].

Usually, Time Different of Arrival (TDOA), steering vector, and spatial correlation matrix are used to present the spatial information.

It is shown that minimum variance distortionless response (MVDR) beamformer significantly improves the performance of hearing aids [4]. Different beamformers such as Weighted Delay and Sum (WDAS), MVDR and Generalized Eigenvector (GEV) are compared each other [5], then the GEV is the most superior. Although WDAS is simple, its performance is poor in non-stationary noise surroundings. On the other hand, MVDR is more superior to MVDR but less than GEV. Especially, the GEV doesn't

Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

need any prior information for the transfer function, so it's said that is simpler than MVDR.

Some scholars studied GEV beamforming based on mask estimation with DNN [6]-[9]. In detail, DNN estimates the masks of clean speech and background noise respectively to calculate the spatial correlation matrix and beamformer. Here, the masks aim to separate the clean speech and noise from noisy signal.

Bidirectional Long Short-Term Memory (BLSTM) [2] is well known to perform well to deal with temporal sequences such as speech signal. A LSTM cell can remember the value with long time of sequences. A new method of sign language recognition based on two-stream 3D-CNN and LSTM network is proposed, and the experimental results show that this method can identify Chinese isolated words sign language very well, with an accuracy rate of 98.4% [10]. DNN, Convolutional Neural Network (CNN) and BLSTM are compared and as a result, the BLSTM is better than the others [7], [8]. Moreover, it can be used to reduce and remove not only noise but also reservation.

On the other hand, many speech applications have their specific wakeup word so that they can 'wake up' only when they are called. The wakeup word speech is used as an important clue for the target speaker, so the target speaker's speech mask and noise mask is estimated in the wakeup word region [6]. And then the masks are applied to the following speech command to enhance the target voice separately. The point is that the masks are evaluated once in wakeup word region and then don't have to be updated during the following beamforming procedure. Thus, when the speech command is short, it's more simple and light than those in [7], [8] for beamforming calculation. In addition, a research on the development of a sonar system that can detect the sound target object using beamforming technique is performed [11].

To estimate the wakeup word mask, a Feed-forward Neural Network (FNN) consisting of Fully-Connected (FC) layers and MVDR beamformer are used [6]. Since the simple structure of this FNN, several neighbouring frames are inputted at each time step to deploy long temporal context. But it is well known that the Recurrent Neural Network (RNN) is more suitable to deal with sequences so when we use a RNN for mask estimator, the performance will be further improved.

And some experiment results on small datasets show that the Gated Recurrent Units (GRU) are faster to train and less to diverge than LSTM [12], [13].

The rest of this paper is as follows. In section 2, we propose a method for enhancing the target speaker's command utterance. We explain about training and test data sets for the experiments in Section 3. Section 4 gives evaluation experiments. Finally, conclusions are presented in Section 5.

2. PROPOSED METHOD

Here, we describe a proposal how to enhance the target speaker's command utterance in noisy condition using wakeup word signal. Convolutional Recurrent Neural Network (CRNN) is used to estimate the wakeup word mask and noise mask. And GEV beamformer is applied to enhance the command utterance of target speaker, which is calculated from the estimated masks.

2.1 MULTICHANNEL SPEECH ENHANCEMENT ARCHITECTURE

Voice-controlled systems such as smart speakers and smart phones are automatically activated by wakeup word [14], [15]. The activated system are ready for speech recognition and it needs to remove noise from the following command utterance and enhance speech of the target speaker who calls the system. We use an eight-microphone array for array signal processing.

Fig. 1 shows the architecture of our method. First, wakeup word detection is performed on a channel of microphone array, for instance, on the 1st channel. Next, we estimate the masks for beamforming from the microphone array signals in the region of the detected wakeup word. Wakeup word masks and noise masks are estimated from noisy signals using deep neural networks on every channel and then they are condensed to a wakeup word mask and a noise mask, respectively. The architecture of network is CRNN, which has better performance than DNN with managing temporal signals. Finally, the estimated wakeup word mask and noise mask, we calculate the beamforming vectors and they are applied to the following command utterance to remove or reduce the unwanted noises.

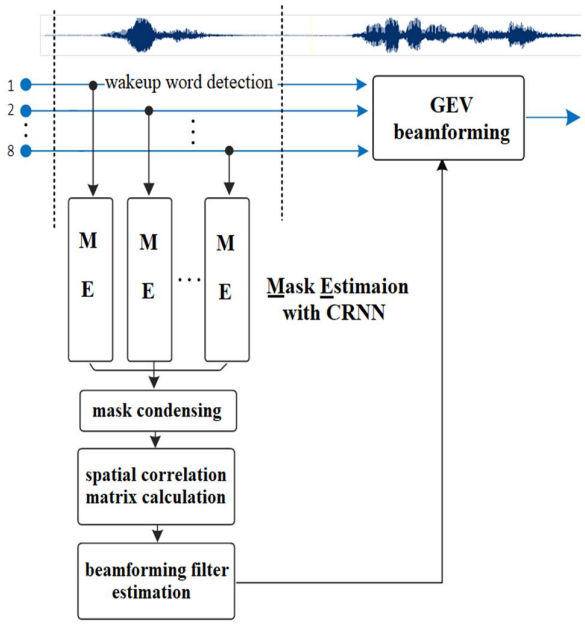

Fig. 1: Architecture of proposed method

Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

2.2 MASK ESTIMATION WITH CRNN

We use the keyword spotting on microcontrollers proposed in [16] for detecting our wakeup word. This approach is based on the depth-wise separable convolutional neural network (DS-CNN) and is suitable on tiny microcontrollers with limited memory. This approach based on DS-CNN is very suitable in resource constrained devices. Korean word "Muagyong" is chosen as our wakeup word. "Muagyong" means ecstasy, a perfect impersonal beatitude, or the state of complete absence of ego. Now, we assume that the region of wakeup word has already been detected by keyword spotting system. We should estimate the wakeup word mask and a noise mask for each channel from microphone array signals.

Fig. 2 shows the proposed network structure for mask estimation. One convolutional layer is at the bottom, which is followed by two recurrent layers. Then, a linear layer is located and finally a FC layer outputs the thresholds in the range of 0~1, which implies whether the signal belongs to either wakeup word or noise at each time-frequency index. The convolutional layer can capture temporal and spectral dependencies in speech features locally [17]. And recurrent layer is beneficial to cope with the long context nature of the sequences. CRNN can exploit both local and global dependencies in time-frequency domain.

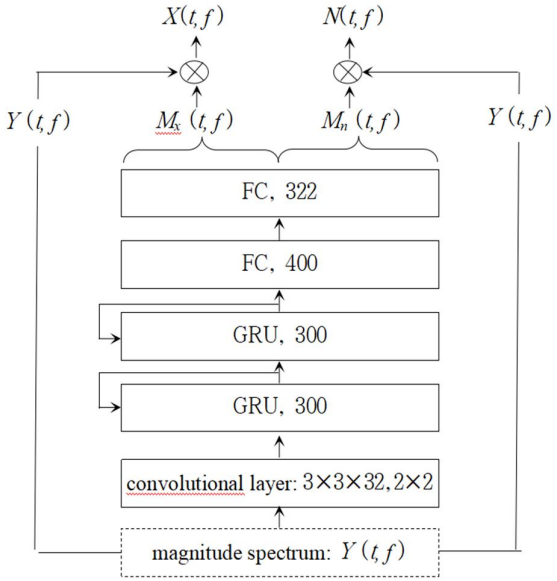

Fig. 2: Architecture of CRNN for mask estimation

The magnitude spectrum is extracted from the noisy signal and then is inputted to the network. Sampling rate is 16Kz, window size is 20ms, overlap size is 10ms and the window is hamming. In recurrent layer, we use GRUs, which is known as better than LSTM in several aspects. Finally a FC layer is laid after double GRUs, whose activation function is Rectified-Linear Unit (ReLU). The output layer has 322(161×2) units, and sigmoid is used for its activation function so that its outputs are able to range from 0 to 1. The first 161 units are corresponding to wakeup word mask and the last 161 units are corresponding to noise mask.

The proposed network is trained so that the result obtained after applying the wakeup mask $M_x(t,f)$ and background noise mask $M_n(t,f)$ to mixture signal, approximate to the clean wakeup word signal $X(t,f)$ and background noise signal $N(t,f)$ respectively. We use the Mean Square Error (MSE) for the loss function as follow:

$$L = \sum_{t=1}^{T}\sum_{f=1}^{F}\left(\left|X(t,f) - M_x(t,f)Y(t,f)\right|^2 + \left|N(t,f) - M_n(t,f)Y(t,f)\right|^2\right) \quad (1)$$

At inference time, it is performed for each channel separately. As a result, we can get 8 wakeup word masks and noise masks respectively. First, we have to condense the several masks to get a single mask for wakeup word and noise, respectively. Although, there are many different ways for mask condensing; max/minimum pooling, average pooling etc. Next, we take the median pooling in equation (2), which occurs less distortion in the condensed result.

$$M_v(t,f) = \underset{j}{median}\, M_v^j(t,f),\ j = \overline{1,8},\ v = \{x,n\} \quad (2)$$

Next, spatial correlation matrix of wakeup word signal and background noise signal by using wakeup word mask $M_x(t,f)$ and background noise mask $M_n(t,f)$ as follows:

$$\mathbf{R}_{vv}(f) = \sum_{t=1}^{L} M_v(t,f)\mathbf{Y}(t,f)\mathbf{Y}^H(t,f),\ f = \overline{1,F}, v = \{x,n\} \quad (3)$$

, where the subscript '*x*' implies clean wakeup word speech and 'n' implies noise in the background. Finally, we get two masks each for wakeup word and noise to estimate the beamformer. We compare the accuracy of two masks estimated by [6] and our proposal in Section 5.

2.3 GEV BEAMFORMER ESTIMATION

Some studies explain about Maximum Signal-to-Noise Ratio (Max-SNR) beamformer, which aims to maximize the Signal-to-Noise Ratio (SNR) truly. Since then, it has been paid attention and presented the investigation in CHIME Challenge [7], [8], which shows that Max-SNR beamforming based on mask estimation with BLSTM performs better than the others.

When we estimate the Max-SNR beamformer, we have to solve Generalized Eigen Value (GEV) problem. So in other words it is called GEV beamformer. J-channel microphone array signal of speech command following the keyword is presented in vector form as follow:

$$\mathbf{Y}(t,f) = \mathbf{X}(t,f) + \mathbf{N}(t,f) \quad (4)$$

, where $\mathbf{Y}(t,f) = [Y_1(t,f), Y_2(t,f), .., Y_J(t,f)]^T$ is noisy signal, $\mathbf{X}(t,f) = \left[X_1(t,f), ..., X_J(t,f)\right]^T$ is clean speech signal and $\mathbf{N}(t,f) = [N_1(t,f), ..., N_J(t,f)]^T$ is noise

Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

signal. Left multiplying both sides of equation (4) by the spatial filter $\mathbf{W}(f)$, we get:

$$\hat{X}(t,f) = \mathbf{W}^H(f)\mathbf{Y}(t,f) = \mathbf{W}^H(f)\mathbf{X}(t,f) + \mathbf{W}^H(f)\mathbf{N}(t,f) \quad (5)$$

The first term in right side implies beamformed output signal, and the second is corresponding to reduced noise. Thus, it is assumed that $\mathbf{X}(t,f)$ and $\mathbf{N}(t,f)$ are uncorrelated, then the output SNR at the f^{th}-frequency bin is written as follow:

$$\mathrm{SNR}_{out}(f) = \frac{\mathbf{W}^H(f)\mathbf{R}_{xx}(f)\mathbf{W}(f)}{\mathbf{W}^H(f)\mathbf{R}_{nn}(f)\mathbf{W}(f)} \quad (6)$$

, which the spatial correlation matrix of clean speech signal and noise signal $\mathbf{R}_{xx}(f)$, $\mathbf{R}_{nn}(f)$ is calculated by equation (3). GEV beamformer can be derived by maximizing the SNR, so that:

$$\hat{\mathbf{W}}(f) = \arg\max_{\mathbf{W}(f)} \frac{\mathbf{W}^H(f)\mathbf{R}_{xx}(f)\mathbf{W}(f)}{\mathbf{W}^H(f)\mathbf{R}_{nn}(f)\mathbf{W}(f)} \quad (7)$$

This optimization problem is equals to the following generalized eigenvalue problem:

$$\mathbf{R}_{xx}(f)\mathbf{W}(f) = \lambda\,\mathbf{R}_{nn}(f)\mathbf{W}(f) \quad (8)$$

Assuming that $\mathbf{R}_{xx}^{-1}(f)$ exists, the principal eigenvector corresponding to $\lambda_{\max}$, the largest eigenvalue of $\mathbf{R}_{xx}^{-1}(f)\mathbf{R}_{nn}(f)$ is the GEV beamforming filter vector. Equation (8) shows that the estimated beamforming filter only depends on the spatial correlation matrix $\mathbf{R}_{xx}(f)$ and $\mathbf{R}_{nn}(f)$.

The suggested beamformer has no need of fixing a reference microphone and acoustic transfer function such as steering vector and TDOA, unlike WDAS and MVDR beamformer. Finally, left multiplying the estimated beamforming filter $\hat{\mathbf{W}}(f)$ to the speech command signal $\mathbf{C}(t,f)$, we can get an enhanced single-channel signal which is expressed as follow:

$$\tilde{C}(t,f) = \hat{\mathbf{W}}^H(f)\mathbf{C}(t,f) \quad (9)$$

, where the SNR at the $\tilde{C}(t,f)$ is $\lambda_{\max}$.

3. DATASET

We use a microphone array "Matrix Voice" with 8 channels for dataset construction and performance evaluation [19]. Loaded with eight microphones, stereo output (headphone and 3W speaker amplifier), plus a ring of 18 RGBW LEDs, WiFi, Bluetooth 4.0 LE, and a dual-core Tensilica Xtensa processor, the Matrix Voice is compatible with all digital voice assistant standards, even "on the edge services like Pocketsphynx and Snowboy," according to Matrix Labs CEO and co-founder Rodolfo Saccoman.

The source of dataset is "Wakeup", which contains 600 wakeup word utterances from different speakers that are recorded using the microphone array, and it's volume is 612kB. As far as concerned, our wakeup word is "Muagyong".

First, we record in several clean rooms, which are roughly 4 by 4.5 meters in dimensions. Speaker are usually 1 or 3 meters away from the microphone. Since the average duration of "Muagyong" is 0.8 s, we limit the length of utterances to 1s.

Next, we simulate different noises such as TV, fan, and even the other speakers' voice, who don't speak "Muagyong".

Next, we mix such noises to clean wakeup speech with different SNRs of -5 dB, 0 dB, 5 dB and 10 dB. This "Wakeup" dataset is used for the training and test set of mask estimation. "Speech Commands" dataset is collection of the utterances that contains the wakeup word and following speech commands, for example, "Muagyong, Turn on the fan". This is used to the evaluation for speech recognition.

It's also divided into two subsets: "speech command-clean" and "speech command-mix". "Speech command-clean" is recorded in clean situation and "speech command-mix" is recorded in noisy situation that we simulate as "Wakeup".

When recording, we use the same microphone array with the "Wakeup". Then, we mix such noises to clean wakeup speech with different SNRs of -5 dB, 0 dB, 5 dB and 10 dB.

4. EXPERIMENTS

We use SDRI to measure how the estimated masks work well. In other words, when we apply the masks to noisy wakeup word signal, wakeup mask has to extract clean wakeup signal and also reduce noises in the background.

In the other hand, the noise mask has to maintain noise signals and remove or reduce the wakeup signal from noisy signal. So After applying the masks, the SDR get higher than before applying. SDRI is expressed as follow:

$$SDR_0 = \frac{1}{F}\sum_{f=1}^{F} 10\log_{10}\left(\frac{\sum_{t=1}^{T} X(t,f)X^*(t,f)}{\sum_{t=1}^{T} N(t,f)N^*(t,f)}\right) \quad (10)$$

$$SDR_1 = \frac{1}{F}\sum_{f=1}^{F} 10\log_{10}\left(\frac{\sum_{t=1}^{T} M(t,f)X(t,f)X^*(t,f)}{\sum_{t=1}^{T} M(t,f)N(t,f)N^*(t,f)}\right) \quad (11)$$

$$SDRI = SDR_1 - SDR_0 \quad (12)$$

, where $X(t,f)$ is target signal which the mask aims to select and $N(t,f)$ is undesirable signal to remove. If SDRI is positive, the mask are considered as correct.

First, we compare our proposal with FNN [6]. Table 1 shows the hyper parameters of two neural networks.

Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

"Wakeup" is used to train and test the networks-the percentage of train and test set is 80%, 20% respectively. Table 2 shows SDRI of two approaches measured in test set. Our proposal achieves the best performance of the other ones. This shows that beamforming performance based wakeup mask estimation is better than usual beamforming method, and the CRNN works better than FNN in modelling temporal/spatial dependency.

Next, we compare different CRNNs with one, two and three recurrent layers, respectively. Table 3 shows their hyper parameters. The evaluation results can be seen in Table 4. Table 4 shows that CRNN-2 which have 2 recurrent layers achieves the best.

Next, the proposal and existing method [6] are compared in the aspect of speech recognition accuracy: word error ratio (WER) each other. DNN-HMM is used for acoustic model, and 3-gram language model. "Speech command" is used to evaluate the speech recognition accuracy. Fig. 3 and Table 5 show the WER of the proposal and existing method [6]. It is known that the WER of our proposal is lower (3.1%) than existing method [6].

Table 1: Parameters of FNN and CRNN

Layer number	FNN	CRNN
1	128 ReLU	3×3×32, 1×2
2	128 ReLU	300 GRU
3	128 ReLU	300 GRU
4	322 sigmoid	400 ReLU
5		322 sigmoid

Table 2: Comparison of SDRI

Architecture	Wakeup word mask [dB]	Noise mask [dB]
FNN	4.1±1.6	3.7±1.2
CRNN	7.3±0.9	6.5±1.1

Table 3: Hyper parameters of CRNN

	CRNN-1	CRNN-2	CRNN-3
cnn	3×3×32, 1×2	3×3×32, 1×2	3×3×32, 1×2
rnn	300 GRU	300 GRU 120 GRU	300 GRU 300 GRU 300 GRU
FC	400ReLU	400ReLU	400ReLU
output	322 sigmoid	322 sigmoid	322 sigmoid

Table 4: Comparison of SDRI of Different CRNNs

Architecture	SDRI [dB]
CRNN-1	4.1±1.6
CRNN-2	7.3±0.9
CRNN-3	7.0±0.8

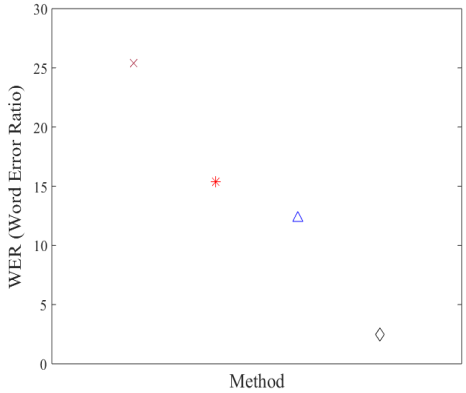

Fig. 3: Comparison of WER in different methods
(× : mixture value, * : FNN+MVDR method,
Δ : proposed method, ◊ : clean value)

Table 5: Comparison of WER

No	Input	WER(%)
1	Mixture	25.4
2	FNN+MVDR [6]	15.4
3	Proposal (CRNN+GEV)	12.3
4	clean	2.5

5. CONCLUSION

We reduce noise by combining and using mask estimation and GEV beamforming method and then enhance the target speaker's command utterance.

Especially, the mask estimation isn't always done, only once in the region of detected wakeup word. And then it is not updated in the following command utterance region so it make possible to effectively enhance the command utterance of target speaker in non-stationary noise situation.

The proposal improves the performance in the both aspects of SDRI and speech recognition accuracy: WER. This tells that the mask estimation accuracy of CRNN is superior to FNN and GEV has more powerful beamforming performance than MVDR. Our proposal can be used to speech enhancement of various smart devices with wakeup word such as smart speaker.

Volume: 15 Issue: 01 Pages: 5754 - 5759 (2023) ISSN: 0975-0290

REFERENCES

[1] B.Y. Xia, and C.C. Bao, Speech enhancement with weighted denoising auto-encoder, Proc. 14th Annual Conf. of the International Speech Communication Association, Lyon, France, 2013, 3411–3415.

[2] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, BLSTM supported GEV beamformer front-end for the 3rd CHIME challenge, Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Scottsdale, AR, 2015, 444-451.

[3] B.D. Van Veen, and K.M. Buckly, Beamforming: a versatile approach to spatial filtering, IEEE Acoustic, Speech and Signal Processing Magazine, 5(2), 1988, 4-24.

[4] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, Multichannel signal enhancement algorithms for assisted listening devices, IEEE Signal Processing Magazine, 32(2), 2015, 18-30.

[5] T. Hori, Z. Chen, H. Erdogan, J.R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech and Language, 46, 2017, 401-418.

[6] Y. Kida, D. Tran, M. Omachi, T. Taniguchi, and Y. Fujita, Speaker selective beamformer with keyword mask estimation, Proc. 2018 IEEE Workshop on Spoken Language Technology, Athens, Greece, 2018, 528-534.

[7] E. Warsitz, and R. Haeb-Umbach, Blind acoustic beamforming based on generalized eigenvalue decomposition, IEEE Transactions on Audio Speech & Language Processing, 15(5), 2007, 1529-1539.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, Proc. 41st IEEE International Conf. on Acoustics, Speech and Signal Processing, Shanghai, PRC, 2016, 196–200.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, A generic neural acoustic beamforming architecture for robust multi-channel speech processing, Computer Speech & Language, 46, 2017, 374-385.

[10] L. Yin, H. Ying, L.D. Kun, L. Rui, and Y.M. Hao, Chinese sign language recognition based on two-stream CNN and LSTM network, International Journal of Advanced Networking and Applications, 14(6), 2023, 5666-5671.

[11] P. Elechi, E. Okowa, and O.P. Illuma, Analysis of a SONAR detecting system using multi-beamforming algorithm, International Journal of Advanced Networking and Applications, 14(5), 2023, 5596-5601.

[12] D. Amodei, S. Ananthanarayan, R. Anubhai, J.L. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, and Q. Cheng, Deep speech 2: End-to-end speech recognition in English and Mandarin, Proc. 33rd International Conf. on Machine Learning, New York, NY, 2016.

[13] Y.B. Zhou, C.M. Xiong, and R. Socher, Regularization techniques for end-to-end speech recognition, Patent, San Francisco, CA, US, US20190130896A1, 2019.

[14] F.Y. Hou, L. Xie, and Z.H. Fu, Investigating neural network based query-by-example keyword spotting approach for personalized wake-up word detection in Mandarin Chinese, Proc. 10th International Symposium on Chinese Spoken Language Processing, Tianjin, PRC, 2017.

[15] G.G.Chen, C. Parada, and G. Heigold, Small-footprint keyword spotting using deep neural networks, Proc. 2014 IEEE International Conf. on Acoustics, Speech and Signal Processing, Florence, Italy, 2014.

[16] Y.D. Zhang, N. Suda, L.Z. Lai, and V. Chandra, Hello Edge: Keyword spotting on microcontrollers, arXiv: 1711.07128, 2017.

[17] T.N. Sainath, and C. Parada, Convolutional neural networks for small-footprint keyword spotting, Proc. 16th Annual Conf. of the International Speech Communication Association, Dresden, Germany, 2015.

[18] A. Krueger, E. Warsitz, and R. Haeb-Umbach, Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation, IEEE Transactions on Audio, Speech and Language Processing, 19(1), 2011, 206–219.

[19] H. Lucy, The MagPi (Raspberry Pi Trading Ltd, 30 Station Road, Cambridge, 2018).