MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Robotic voice assistant equipped with binaural audio

Bai, Mingsian R.[1], Hsu, Yi-Cheng[2], Lee, Tsung-Han[3] , Chen, Wen-Chuen[4]
Department of Power Mechanical Engineering, National Tsing Hua University
101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013

ABSTRACT

A robotic voice assistant is constructed for entertaining service. The robot is comprised primarily of three functional units: a microphone array, a cloud-based voice assistant and a binaural rendering loudspeaker array. The microphone array is utilized to locate the human user and to extract the speech commands given by the user. The extracted commands are then sent to a convolution neural network built ourselves. The response from the cloud is broadcast at the robot end by using a linear loudspeaker array. Various binaural processing modes are implemented in light of a special inverse filtering approach. The inverse filters are formulated in the time domain, which makes it immune to the noncausal artifacts such as wraparound errors and pre-ringing that are frequently encountered in the frequency-domain formulations. However, the frequency-domain weighting and equalization is still possible in the proposed approach. An industrial personal computer serves as the coordinator of the preceding processing units. With these 3 units working in tandem, the proposed robot is capable of interpreting human commands and responding with immersive binaural audio.

Keywords: Binaural audio, Source localization, Speech command
I-INCE Classification of Subject Number: 74

1. INTRODUCTION

Digital voice assistant (VA) is receiving increased popularity. Despite much research on internet of things (IoT), sound field control in VA application is still relatively untapped in acoustic signal processing. We proposed a system comprised of three units: binaural audio, localization and speech recognition, as depicted in Fig. 1.

[1] msbai@pme.nthu.edu.tw
[2] mark754651@gmail.com
[3] bobolee1239@gmail.com
[4] a0956272621@gmail.com

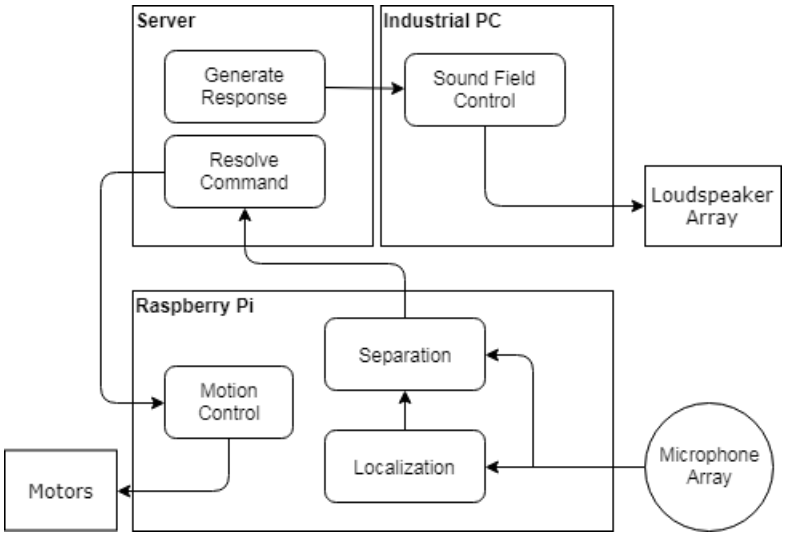

Fig1. Block diagram of the robotic voice assistant

In the acquisition phase of the VA, source localization alongside signal extraction is carried out. To this end, the VA is equipped with a six-microphone uniform circular array (UCA) for single-source localization. We propose a direction of arrival (DOA) estimation algorithm based on time difference of arrival (TDOA) measurement. Instead of conventional TDOA-based algorithms [1-3] that locate the source of interest by finding the intersection of hyperboloid branches, we directly solve a linear system for the DOA, which is appealing for real-time implementation in terms of accuracy and complexity.

As one of the most important ingredients of this work, synthesis of virtual auditory scenes has been an active research topic in audio signal processing. There are two categories of spatial sound field reproduction: multichannel audio and binaural audio. Multichannel audio seeks to recreate globally an immersive sound field, with the aid of multiple loudspeakers [4-6], while binaural audio aims to reproduce the target field at the ears of oftentimes a single user [7, 8]. In the rendering phase of the VA, we focus on binaural audio rendering using a loudspeaker array.

The sound reproduction can be regraded as an inverse problem. In general, two types of deconvolution approaches, the frequency-domain method [9] and the time-domain method [10, 11] can be utilized to design the required inverse filters. In this paper, we build upon the previous research of a time-domain with multichannel inverse filtering. The problem is formulated into an underdetermined system by introducing multiple channels of transducers. The underdetermined problem enables exact model matching without residual errors. The time-domain underdetermined multichannel inverse filtering (TUMIF) approach is employed to design the prefilters for the binaural audio processsor. The TUMIF approach is examplified through the application of cross-talk cancellation (XTC) [12-14].

In the intelligent classification phase of the VA, convolutional neural networks (CNN) are utilized in speech command detection and sound events classification. Speech command recognition can be found in references [15, 16]. In this paper, we apply Very Deep Convolutional Networks for Large-Scale Image Recognition Networks (VGGNet) [17] to classify by using Mel-frequency spectral coefficients

(MFSC) as the features [18]. In the training phase of the networks, batch normalization [19] and learning rate decay, etc., are employed. In addition, the dataset is also augmented to make the system robust to adverse environments with background noise and reverberations. The performance of the inference engine is assessed by using F1-scores.

2. LOCALIZATION

2.1 Coordinate system for the UCA

Under the two-dimensional plane-wave assumption, sound pressure $p(\psi_m)$ received by the *m*th microphone of an *M*-microphone UCA can be expressed as

$$p(\psi_m) = Ae^{-j\boldsymbol{\kappa}\mathbf{r}_m}, \tag{1}$$

where $m = 1,\ldots,M$, $\mathbf{r}_m = r(\cos\psi_m, \sin\psi_m)$, r is the radial distance, and ψ_m is the corresponding angular position of the *m*th microphone. *A* is the amplitude and $\boldsymbol{\kappa} = k(\cos\theta, \sin\theta)$ with *k* being is the wavenumber. The angle θ is the look direction of the incident plane wave. Thus, the array steering vector $\mathbf{a}(\theta)$ of the UCA is given by

$$\mathbf{a}(\theta) = \begin{bmatrix} e^{-jkr\cos(\theta-\psi_1)} & \cdots & e^{-jkr\cos(\theta-\psi_M)} \end{bmatrix}^T \tag{2}$$

2.2 DOA estimation

Taking the first microphone as the reference, we can rewrite the steering vector as

$$\hat{\mathbf{a}}(\theta) = \begin{bmatrix} 1 & \cdots & e^{-jkr[\cos(\theta-\psi_M)-\cos(\theta-\psi_1)]} \end{bmatrix}^T \tag{3}$$

It follows that the TDOAs can be obtained by matching the phase shifts in the elements of the preceding steering vector with those due to the time delays relative to the reference microphone.

$$\begin{bmatrix} 1 & \cdots & e^{-jkr[\cos(\theta-\psi_M)-\cos(\theta-\psi_1)]} \end{bmatrix}^T = \begin{bmatrix} e^{-j\omega\tau_{11}} & \cdots & e^{-j\omega\tau_{M1}} \end{bmatrix}^T, \tag{4}$$

where ω is the corresponding frequency and τ_{m1} denotes the TDOA between the *m*th microphone and the first microphone, and can be estimated by generalized cross correlation phase transformation (GCC-PHAT) [20]. For the *m*th microphone,

$$kr\left[\cos(\theta-\psi_m) - \cos(\theta-\psi_1)\right] = \omega\tau_{m1}$$

$$\Rightarrow r(\cos\psi_m - \cos\psi_1)\cos\theta + r(\sin\psi_m - \sin\psi_1)\sin\theta = c\tau_{m1} \tag{5}$$

Rearranging Eq.(5) for $m = 2,\ldots,M$ into a matrix equation leads to

$$\begin{bmatrix} (\cos\psi_2 - \cos\psi_1) & (\sin\psi_2 - \sin\psi_1) \\ \vdots & \vdots \\ (\cos\psi_M - \cos\psi_1) & (\sin\psi_M - \sin\psi_1) \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = \begin{bmatrix} \frac{c\tau_{21}}{r} \\ \vdots \\ \frac{c\tau_{M1}}{r} \end{bmatrix} \tag{6}$$

Solution of $(\cos\theta, \sin\theta)$ in the equation above gives the information of DOA.

3. BINAURAL AUDIO RENDERING WITH THE TIME-DOMAIN UNDERDETERMINED MULTICHANNEL INVERSE PREFILTERS (TUMIF)

In this chapter, the TUMIF approach is adopted to design prefilters for binaural processing. The problem is formulated into an underdetermined system by introducing multiple channels of control loudspeakers. It follows that there are an infinite number of exact solutions for this problem. The matching model aims to reproduce desired signals at the control points of the ears. Tikhonov regularization (TIKR) [21] is employed to calculate the inverse filters.

3.1 The TUMIF-based binaural audio reproduction

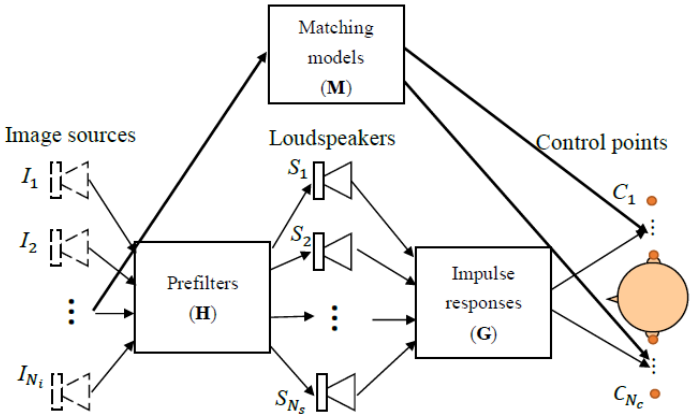

Fig. 2. The block diagram of a general multichannel model-matching problem.

The proposed robust binaural rendering system is depicted in Fig. 2. N_s target sources, N_c control points and N_s loudspeakers are indicated. The design problem can be viewed as a model matching problem:

$$\begin{bmatrix} m_{11}(k) & \cdots & m_{1N_i}(k) \\ \vdots & \ddots & \vdots \\ m_{N_c1}(k) & \cdots & m_{N_cN_i}(k) \end{bmatrix} = \begin{bmatrix} g_{11}(k) & \cdots & g_{1N_s}(k) \\ \vdots & \ddots & \vdots \\ g_{N_c1}(k) & \cdots & g_{N_cN_s}(k) \end{bmatrix} * \begin{bmatrix} h_{11}(k) & \cdots & h_{1N_i}(k) \\ \vdots & \ddots & \vdots \\ h_{N_s1}(k) & \cdots & h_{N_sL_i}(k) \end{bmatrix}, \quad (7)$$

where *m(k)*'s, *g(k)*'s, and *h(k)*'s symbolize the impulse response sequences of the matching model, the room response, the prefilters, and "*" stands for linear convolution. N_i, N_s, and N_c denote the numbers of the target sources, loudspeakers, and control points. Assume that *g(k)*'s, and *h(k)*'s are of lengths L_g and L_h, respectively, so the *m*'s must be of length, $L = \left(L_g + L_h - 1\right)$. In matrix notation, the equation above can be written as

$$\mathbf{M}(k) = \mathbf{G}(k) * \mathbf{H}(k) \quad (8)$$

Assume all three systems are finite-impulse-response (FIR) systems. Equation (8) can be rewritten into the following matrix form:

$$\begin{bmatrix} \mathbf{m}_{11} & \cdots & \mathbf{m}_{1N_i} \\ \vdots & \ddots & \vdots \\ \mathbf{m}_{N_c1} & \cdots & \mathbf{m}_{N_cN_i} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1N_s} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{N_c1} & \cdots & \mathbf{G}_{N_cN_s} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{11} & \cdots & \mathbf{h}_{1N_i} \\ \vdots & \ddots & \vdots \\ \mathbf{h}_{N_s1} & \cdots & \mathbf{h}_{N_sN_i} \end{bmatrix} \tag{9}$$

Or,

$$\mathbf{m} = \mathbf{G}\mathbf{h}, \tag{10}$$

where **m**'s and **h**'s represent the impulse response vectors of the matching models and the prefilters:

$$\mathbf{m}_{ij}(k) = \begin{bmatrix} m_{ij}(0) & m_{ij}(1) & \cdots & m_{ij}(L-1) \end{bmatrix}^T,$$
$$\mathbf{h}_{ij}(k) = \begin{bmatrix} h_{ij}(0) & h_{ij}(1) & \cdots & h_{ij}(L_h-1) \end{bmatrix}^T.$$

The convolution matrix, $\mathbf{G}_{ij}$, associated with the *j*th control loudspeaker and the *i*th control point is defined as

$$\mathbf{G}_{ij} = \begin{bmatrix} g_{ij}(0) & 0 & 0 & 0 \\ g_{ij}(1) & g_{ij}(0) & 0 & \vdots \\ \vdots & g_{ij}(1) & \ddots & 0 \\ g_{ij}(L_g-1) & \vdots & \ddots & g_{ij}(0) \\ 0 & g_{ij}(L_g-1) & \ddots & g_{ij}(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & g_{ij}(L_g-1) \end{bmatrix}_{L\times L_h} \tag{11}$$

Equation (9) represents a general multiple-input-multiple-output (MIMO) model-matching problem

$$\mathbf{M}_{LN_c\times N_i} = \mathbf{G}_{LN_c\times L_hN_s}\mathbf{H}_{L_hN_s\times N_i} \tag{12}$$

The matrices, **M, H,** and **G,** represent the matching model, prefilters, and the room impulse response matrix between the loudspeaker inputs and the control points. It should be noted that **G** is an $LN_c \times L_hN_s$ matrix. For Eq. (5) to be an underdetermined system, the following inequality must be satisfied

$$LN_c < L_hN_s \Rightarrow (L_g + L_h - 1)N_c < L_hN_s \tag{13}$$

Therefore, the length of the prefilters must be selected according to

$$L_h \geq \frac{(L_g-1)N_c}{N_s - N_c}, \quad N_s > N_c \tag{14}$$

3.2 The matching model of cross-talk cancellation

The XTC binaural audio rendering is intended to create a headphone-like listening experience. In this case, the matching model is selected as

$$\mathbf{M}_1 = \begin{bmatrix} \boldsymbol{\delta}_{LL} & \mathbf{0}_{LR} \\ \mathbf{0}_{RL} & \boldsymbol{\delta}_{RR} \end{bmatrix}_{N_c \times 2}, \tag{15}$$

where

$$\boldsymbol{\delta}_{LL} = \begin{bmatrix} \boldsymbol{\delta}_1^T & \boldsymbol{\delta}_2^T & \cdots & \boldsymbol{\delta}_{N_{LL}}^T \end{bmatrix}^T \in R^{LN_{LL}}, \boldsymbol{\delta}_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \in R^L, i = 1, \ldots, N_{LL}$$

$$\boldsymbol{\delta}_{RR} = \begin{bmatrix} \boldsymbol{\delta}_1^T & \boldsymbol{\delta}_2^T & \cdots & \boldsymbol{\delta}_{N_{RR}}^T \end{bmatrix}^T \in R^{LN_{RR}}, \boldsymbol{\delta}_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \in R^L, i = 1, \ldots, N_{RR}$$

$$\mathbf{0}_{RL} = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T \in R^{LN_{RL}}, \mathbf{0}_{LR} = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T \in R^{LN_{LR}}$$

with $\boldsymbol{\delta}_i(k)$ being the unit pulse sequence. N_{LL}, N_{RR}, N_{RL}, and N_{LR} are numbers of control points with subscripts "*L*" and "*R*" denoting "left" and "right." In total, the number of control points on the ipsilateral and the contralateral sides is N_c, i.e., $N_{LL} + N_{RL} = N_{RR} + N_{LR} = N_c$. The matching model aims to minimize the response from the target source to the contralateral ear.

3.3 Inverse filtering with Tikhonov regularization

Tikhonov regularization (TIKR) [20] is utilized to solve the MIMO model matching problem, **M** = **GH**. For an underdetermined and full-rank room response **G**, the minimum-norm least-squares solution is given by

$$\mathbf{H} = \mathbf{G}^T \left[\mathbf{G}\mathbf{G}^T \right]^{-1} \mathbf{M} \tag{16}$$

Although this is an exact solution that gives zero residual errors, some regularization is required to limit the gain of the filters. A Tikhonov regularization (TIKR) method can be used by solving the following optimization problem:

$$\min_{\mathbf{H}} \left(\left\| \mathbf{G}\mathbf{H} - \mathbf{M} \right\|_F^2 + \beta^2 \left\| \mathbf{H} \right\|_F^2 \right), \tag{17}$$

where $\left\| \bullet \right\|_F$ denotes the Frobenius norm and β is a regularization parameter. It can be shown that the optimal solution is

$$\mathbf{H} = \left[\mathbf{G}^T\mathbf{G} + \beta^2 \mathbf{I} \right]^{-1} \mathbf{G}^T\mathbf{M} \tag{18}$$

4. SPEECH COMMAND RECOGNITION

4.1 Speech command dataset

The speech commands dataset provided by Google® [22] is employed in the study. The dataset contains 65000 one-second utterances of 30 short words. We use 10 classes and 10000 records (1000 for each class) as the training set. Pre-processing is needed due to the inconsistence of the dataset. Records are all truncated or padded to 1s. Furthermore, white noise (40 dB) served as the data augmentation helps training results.

4.2 Feature extraction

MFCC has been widely used in automatic speech recognition. By omitting the last discrete cosine transform (DCT) step in Mel Frequency Cepstral Coefficient (MFCC), Mel Frequency Spectrum Coefficient (MFSC) can be obtained in the Mel-frequency bands. In this paper, Mel-spectrogram comprised of MFSC sequences is

employed as the input features for the neural networks. Different frame size and hyperparameters of Fast Fourier Transform must be fine-tuned to achieve better results.

4.3 Network architecture

CNNs have been extensively used in image recognition. We use VGG-like architecture (Fig. 3) that comprises 6 CNNs as the deep learning classifier. The network architecture is advantageous in that two concatenated 3×3 convolutional layers are capable of producing the same results as one 5×5 layer. Therefore, large-size filters such as 11×11 in AlexNet and 7×7 in ZFNet are not required. Furthermore, batch normalization which is commonly used in the VGGNet helps to accelerate the training phase. In addition, dynamic learning rates such as exponentially decaying learning rates or gradient descent with warm restarts [23]) are adopted in this paper.

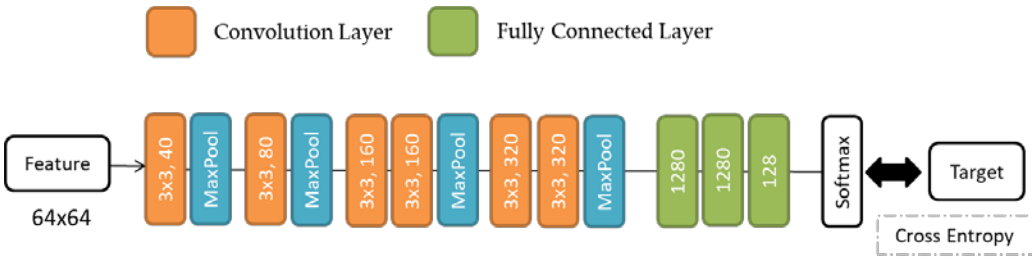

Fig. 3. VGG-like Network Architecture

4.4 Inference

The testing data including 10 classes and 1000 data records (100 for each class) are collected under adverse conditions, e.g., with reverberation, noise, and interference. The performance of the classifier will be assessed by the F1-score, and confusion matrix showing the final results will be discussed in following section.

5. EXPERIMENTS

To validate the DOA estimation algorithm, a six-microphone uniform circular array with radius 3.2 cm is used in the test. A loudspeaker source is sequentially placed at the farfield in 36 positions uniformly spaced around a circle. It turns out that the root-mean-squares error of the localization results obtained using the DOA estimation algorithm in Sec. 2.2 is only 2.64˚.

The experimental arrangement in a listening room is shown in Fig. 4. A six-element linear loudspeaker array with 42 cm aperture length is constructed. The distance between the array and the listener is 80 cm. Two microphones at the vicinity of the ears of an artificial head and torso serve as control points. The impulse response matrices **G** and **M** in relation to the dynamics of the head and torso, the loudspeakers, and the room are measured in advance. The prefilters are calculated on the basis of the aforementioned TUMIF approach.

Fig. 4. Experimental arrangement of the binaural audio rendering system.

The performance of the XTC can be objectively assessed by comparing the channel separations between the ipsilateral and contralateral responses with D_{cs} defined as

$$D_{cs} = \text{ipsilateral FRF (LL, dB) - contralateral FRF (LR, dB).} \quad (14)$$

For simplicity, only the left target source driven with a white noise signal with the quiescent right target source is examined in the experiment. The channel separations achieved with 3 regularization parameters β are shown in Fig. 5. It can be seen from the results that channel separation increases if a small β is used. This suggests that β is a crucial parameter to trade channel separation performance for audio quality.

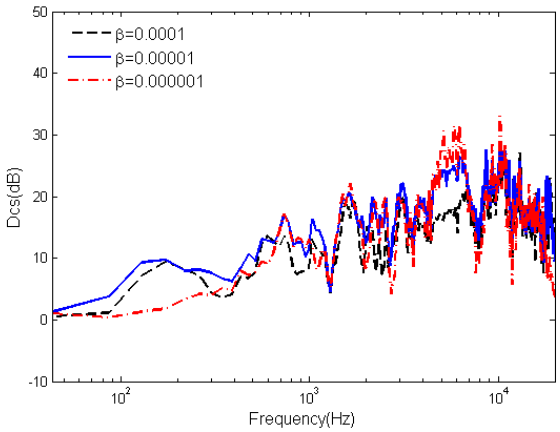

Fig. 5. The measured channel separation of the XTC system from the left target source to the ipsilateral and contralateral control points, with the regularization parameter, $\beta = 10^{-6}$, 10^{-5}, and 10^{-4}.

A listening test is also conducted according to the multiple stimuli with hidden reference and anchor (MUSHRA) procedure for the subjective assessment of audio quality [24]. The distance between the array and the listener is 80 cm. The regularization parameter β used in designing the prefilters is 10^{-6}. The scores of three subjective attributes, sense of widening, coloration, and artifacts, processed by the analysis of variance (ANOVA) are shown in Fig. 6. It can be seen that all subjective

indices are significantly improved with the TUMIF approach.

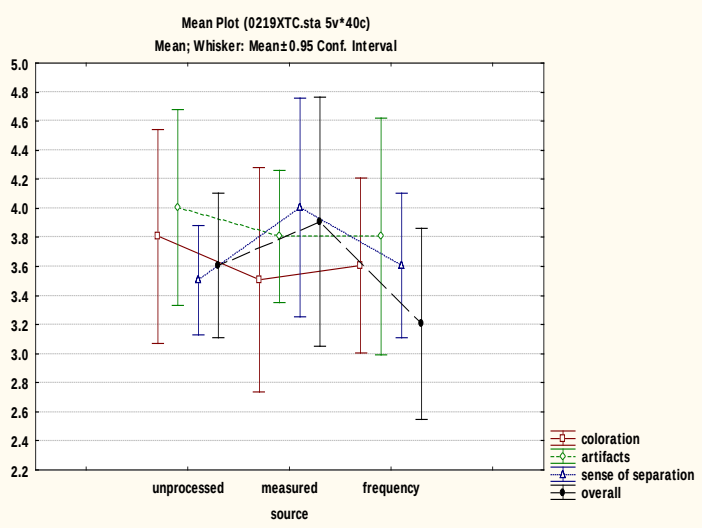

Fig. 6. The ANOVA output of the listening test for XTC.

With the proposed deep learning networks, the F1-score can reach as high as 0.96 in the testing phase. Compared to others, commands “up” and “off” can often be misclassified due to the similarity in the two utterances. As indicated in the confusion matrix (Fig. 7) by a red box, the word “off” is recognized as “up” by a probability of 11%, which is the highest misclassification rate in all classes. The channel mismatch in different voice recorders may have contributed to the high misclassification rate.

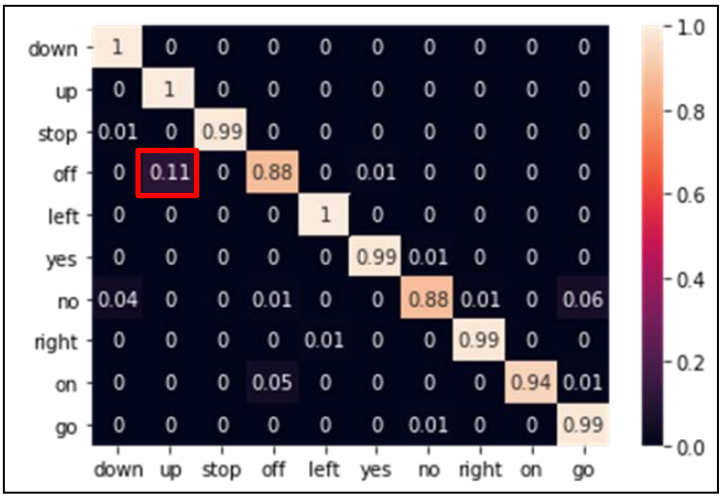

Fig. 7. Confusion Matrix.

6. CONCLUSIONS

A system of entertainment robotic VA is proposed and implemented in this paper. A circular microphone array is used to localize the source and extract its signal. Neural networks serve as a voice command recognizer. The response is rendered by using binaural processing that is implemented on a linear loudspeaker array. By formulating the model-matching problem into an underdetermining system, nearly perfect matching with mild regularization can be achieved. It is crucial to select the regularization

parameter β that strikes a proper balance between match performance and audio quality. Experimental result shows that the time-domain method yields better separation performance and audio quality than the frequency-domain method.

7. ACKNOWLEDGEMENTS

The work was supported by the Ministry of Science and Technology (MOST) in Taiwan, Republic of China, under the project number 107-2221-E-007 -039 -MY3.

8. REFERENCES

1. S. Argentieri, P. Danes, and P. Soueres, *"A survey on sound source localization in robotic: From binaural to array processing methods"*, Comput. Speech Language, vol. 34, no. 1, pp. 87-112 (2015)

2. H. C. So, *"Source localization: Algorithms and analysis"*, Handbook of Position Localization: Theory, Practice and Advances, Chapter 2, edited by S. A. Zekavat and R. M. Buehrer, Wiley-IEEE (2011)

3. K. W. Cheung, H.C. So, W. K. Ma, and Y. T. Chan, *"Least square algorithms for time-of-arrival-based mobile location"*, IEEE Trans. Signal Processing, vol. 52, no. 4, pp.1121-1128, Apr. (2004)

4. T. Sporer, *"Wave field synthesis--generation and reproduction of natural sound environments"*, in Proceedings of the 7th International Conference on Digital Audio Effects, Naples, Italy (2004).

5. M. Kolundzija, C. Faller, and M. Vetterli, *" Designing practical filters for sound field reconstruction"*, in Proceedings of the 127th AES Convention, Audio Engineering Society, New York (2009).

6. M. A. Gerzon, *"Ambisonic in multichannel broadcasting and video"*, J. Audio Eng. Soc. **33**(11), 859-871 (1985).

7. B. B. Bauer, *''Stereophonic earphones and binaural loudspeakers"*, J. Audio Eng. Soc. **9**(2), 148-151 (1961).

8. W. F. Druyvesteyn and J. Garas, *"Personal sound"*, J. Audio Eng. Soc. **45**, 685–701 (1997).

9. O. Kirkeby, P. A. Nelson, and H. Hamada, *"Fast deconvolution of multichannel systems using regularization"*, IEEE Trans. Speech and Audio Processing, **6**(2), 189-195(1998).

10. O. Kirkeby and P. A. Nelson, *"Digital Filter Design for Inversion Problems in Sound Reproduction"*, J. Audio Eng. Soc. **47**(7/8), 583-595 (1999).

11. M. Miyoshi and Y. Kaneda, *"Inverse filtering of room acoustics"*, IEEE Trans. Acoust., Speech, Signal Process. **36**(2),145-152(1988).

12. M. R. Schroeder and B. S. Atal, *''Computer simulation of sound transmission in rooms"*, Proc. IEEE **51**(3), 536-537 (1963).

13. J. L. Bauck and D. H. Cooper, *"Generalized transaural stereo and applications"*, J. Audio Eng. Soc. **44**(9), 683-705 (1996).

14. M. R. Bai, and C. C. Lee, *"Objective and subjective analysis of effects of loudspeaker span on crosstalk cancellation in spatial sound reproduction"*, J. Acoust.

Soc. Am., **120**(4), 1976-1989, (2006).
15. T. N. Sainath, and C. Parada, *"Convolutional Neural Networks for Small-Footprint Keyword Spotting"*, Interspeech (2015).
16. G. Chen, C. Parada, and H, Georg, *"Small-footprint keyword spotting using deep neural networks"*, ICASSP (2014).
17. S. Liu and W. Deng, "*Very deep convolutional neural network based image classification using small training sample size"*, IAPR (2015).
18. O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "*Convolutional Neural Networks for Speech Recognition"*, IEEE/ACM (2014).
19. I, Sergey and C. Szegedy. *"Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift"*, ICML (2015).
20. C. H. Knapp and G. C. Carter, *"The generalized correlation method for estimation of time delay"*, IEEE *Trans. Acoust*., Speech, Signal Processing, vol. ASSP-24, pp. 320-327 (1976).
21. C. W Groetsch, *The theory of Tikhonov regularization for Fredholm equation of the first kind*, Pitman Advanced Pub. Program, Boston (1984).
22. P. Warden, "Speech commands dataset", Internet: https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html, (2017).
23. L. Ilya and F. Hutter, *"SGDR: Stochastic Gradient Descent with Warm Restarts"*, ICLR (2017).
24. ITU-R Recommendation BS.1534-1, *"Method for the subjective assessment of intermediate sound quality (MUSHRA)"*, International Telecommunications Union, Geneva, Switzerland (2001).