
AFRL-AFOSR-JP-TR-2018-0005

# Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency response missions

**Hanseok Ko**
**Korea University Research and Business Foundation**

**01/04/2018**
**Final Report**



Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From - To)* |
|---|---|---|
| 07-11-2017 | Final | 08/08/2016-08/07/2017 |

**4. TITLE AND SUBTITLE**

Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
FA2386-16-1-4130

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Hanseok Ko
Sungkyu Mun,
Minkyu Lee
Sangwook Park
Sungjae Lee

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Korea University Research and Business Foundation
145 Anam-ro, Seongbuk-gu
Seoul 02841
Republic of Korea

**8. PERFORMING ORGANIZATION REPORT NUMBER**

AFOSR-hsko-2017-01

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
AOARD
7-23-17, ROPPONGI, MINATO-KU
TOKYO 106-0031 JAPAN

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AOARD

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Distribution unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Three key-techniques were investigated for achieving robust acoustics and speech perception of aerial robot for scene understanding during critical emergency response missions. For noise robust sound source localization, a noise robust desired sound direction estimation method is developed using LSTM based weighting function. The direction estimation experiments confirmed that the proposed method shows improved robustness under indoor surveillance noise environment characterized by presence of harmonic or non-stationary noise sources. For attaining signal enhancement under noisy environment, the GSC exploits spatial information and generates multi-channel enhanced signals on which the following DAE can act. As a result, the DAE can take advantage of the multi-channels by modeling the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels. The evaluation results demonstrate that utilizing the results of the proposed GSC structure as an input to the DAE is effective in improving noise reduction and speech recognition performance. To improve acoustic event recognition performance and overcome the deficit of acoustic event resource, a novel DNN based transfer learning approach is developed. By utilizing the information transferred from the universal source domain, the proposed approach improved AEC accuracy in indoor surveillance experiments.

**15. SUBJECT TERMS**
Aerial robot, ego noise, Deep Neural Networks, detection, localization, transfer learning

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| | | | | | Hanseok Ko |
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | **19b. TELEPHONE NUMBER** *(include area code)* |
| | | | | | +82-10-9001-3239 |

# Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency

This report describes the research work and results of the project entitled: **Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency response missions**, which took place during 2016.08.08~2017.08.07. Authors gratefully acknowledge the support of this research by the Air Force Office of Scientific Research.

## 1. Research Objectives

In this research, Hanseok Ko as Principal Investigator and his team of graduate students investigated acoustic signal based robust scene understanding techniques for aerial robot. Due to the added advantage of flying besides having the usual movement dexterity, an aerial robot creates a new set of opportunities to provide highly intelligent perceptions to engage and respond to time-critical emergency situations. It is desirable to equip aerial robots with the ability to detect evidence of emergency situations such as calling for help by human voice via using a microphone array to listen to surrounding zone, perform speech/acoustic event detection, and conduct sound source localization for directivity assessment and for possible tracking if the source is either stationary or moving.

Despite of its potential capability in emergency response missions over hazardous sites and observing wide areas, auditory processing in aerial robots is technically extremely challenging due to the inherent issues delineated as follows.

- High level of wind noise and ego-noise from rotor
- Constantly changing noise level and target to sensor distance while robot is moving
- Wide dynamic range of target signal power by changing target to sensor distance in outdoor environment
- High probability of overlapped acoustic events in wide searching area
- Difficulty of reliable target detection due to ever-present ego-noise
- Robust acoustic scene classification

To address the above issues and explore mitigating approaches, a 3-phase plan was set up at the onset of this project as shown in Figure 1. We explored the audio perception techniques under **idle state** as goal of 1st year.

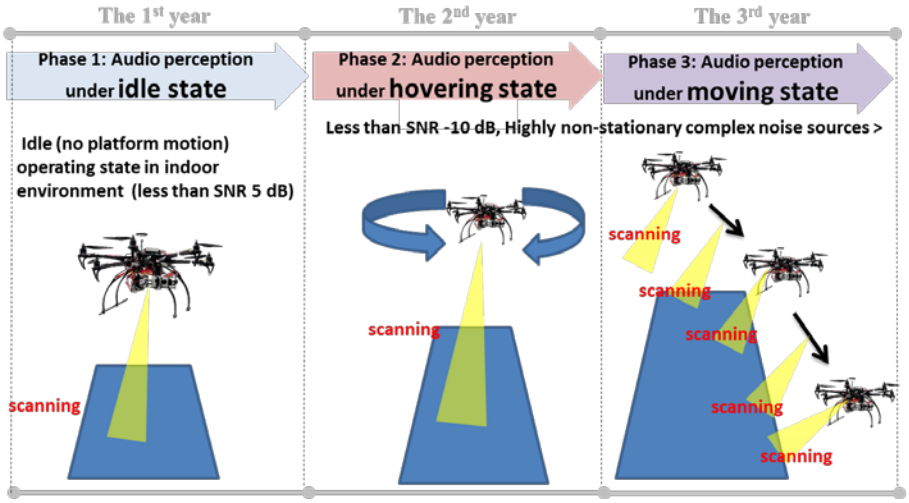


**Fig. 1** 3-phase (3-years) plan for audio perception under aerial robot environments

In order to achieve the research goal of Phase 1, **we investigated the following 3 specific research issues.**

(1) *How can we select desired signal dominant frequency bin in non-stationary noisy environment for* ***robust sound source localization****?*

(2) *For* ***signal enhancement****, what beamforming structure would be effective for utilizing the modeling ability of deep neural network?*

(3) *To achieve robustness of acoustic event (scene) classification, a powerful mitigating approach would be to provide a large database made available for training. What deep learning based approaches can be rendered effective to generate useful training database?*

## 2. Research tasks and results

### 2.1 (Task 1) Investigate novel sound source localization techniques in highly non-stationary noise dominant environments

We investigated the problem of finding desired signal using a microphone array in highly non-stationary noisy environment. Frequency bins with high noise levels can be inadvertently considered as desired sound sources. Reliable estimation performance can be achieved by using meaningful signal-dominant frequency bins while avoiding noise-dominant bins.

***Issue : How can we select desired signal dominant frequency bin in non-stationary noisy environment for robust sound source localization?***

To address this issue, hence, selecting (or weighting high) desired signal frequency bin in non-stationary noise, a weighting function is investigated using Deep Neural-Network (DNN). Using DNN approach for selecting desired signal frequency turned out to be effective. In DNN based time-frequency mask estimation step, the ideal binary masks for signal are set as

$$\widehat{M_X}(l,\omega) = \begin{cases} 1, & \frac{\|X(l,\omega)\|}{\|N(l,\omega)\|} > threshold \\ 0, & else \end{cases} \quad (1)$$

where $X(l,\omega)$ and $N(l,\omega)$ are the desired signal component and noise component in $l_{\text{th}}$ frame of input noisy signal in STFT domain respectively. A DNN composed of LSTM and feedforward layers is trained to estimate the ideal binary masks, $\widehat{M_X}(l,\omega)$ from a noisy spectrum. The noisy spectrum is generated for the training step by summing the signal and noise signal which is recorded in real environments. The specific structure of DNN is shown in the table below,

**Table 1.** LSTM network configuration for mask estimation

| Layers | Units | Type | Non-linearity | $p_{dropout}$ |
|---|---|---|---|---|
| L1 | 256 | LSTM | Tanh | 0.5 |
| L2 | 513 | FF | ReLU | 0.5 |
| L3 | 513 | FF | ReLU | 0.5 |
| L4 | 1026 | FF | Sigmoid | 0.0 |

The window size and frame shift are set at 1024 and 512 samples, respectively, at a 16 kHz sampling rate. $p_{dropout}$ means the ratio of nodes to apply dropout technique in training step. The DNN consists of one Long Short-Term Memory (LSTM) layer and three Feed-Forward (FF) layers.

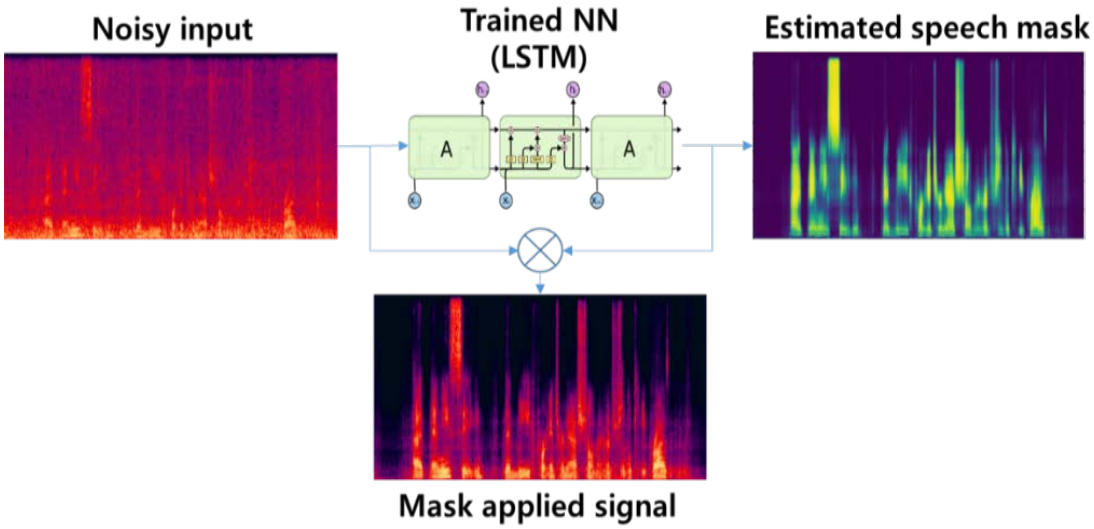


**Fig. 2** LSTM based speech mask estimation process (speech as desired signal)

**Result:**

**- We developed a novel masking method for sound source localization in the noisy environments by:**

**① Using the DNN based approaches for modeling the relationship between the noisy and clean signals as a nonlinear transformation.**

**② Exploiting the sequential information from the previous adjacent frames using feedback of LSTM network, which conventional approaches cannot utilize.**

After applying DNN based mask to each channel of input signal, Time Difference Of Arrival (TDOA) is estimated for each pair of microphones to find location of acoustic source. The TDOAs can be estimated from so-called angular spectrum, whose peaks indicate the TDOA of the source. Generalized Cross Correlation-Phase Transform (GCC-PHAT) is used as an angular spectrum in this system. GCC of the $m_{th}$ and $n_{th}$ microphone signals is calculated as eq. (2)

$$R_{mn}(l,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{mn}(l,\omega) X_m(l,\omega) X_n^*(l,\omega) e^{j\omega\tau} d\omega, \quad (2)$$

where $\Psi_{mn}(l,\omega)$ denotes a weight function. Although many different weighting functions can be applied, the PHAT has been found to perform quite well under realistic acoustical conditions.

$$\Psi_{mn}(l,\omega) = \frac{1}{|X_m(l,\omega) X_n^*(l,\omega)|} \quad (3)$$

The Figure 3 shows the outline of GCC-PHAT based source localization system.

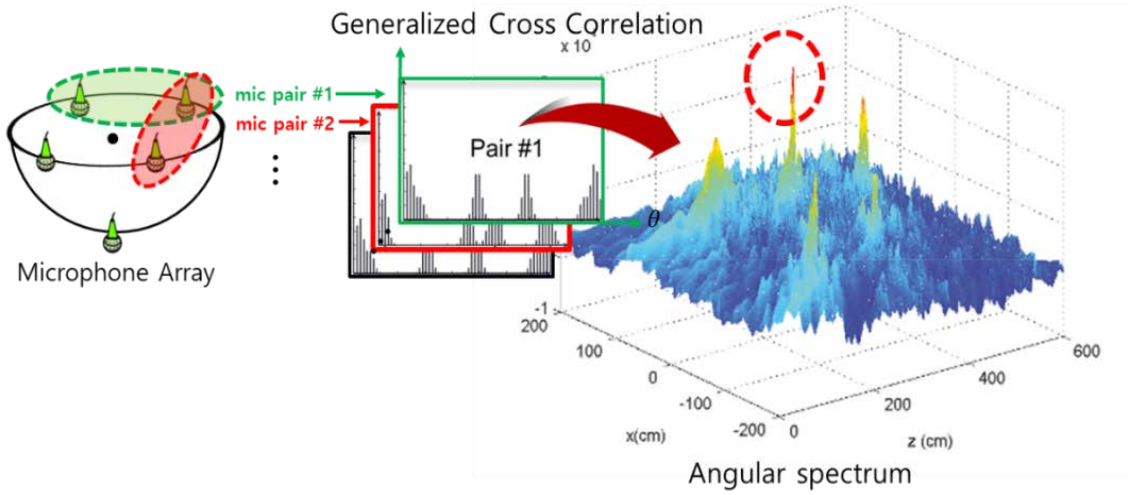


**Fig. 3** GCC-PHAT based source localization system

The proposed algorithm was evaluated in a 5.0 × 6.0 × 2.5 $m^3$ simulated room environment using image source method. The reverberation time setting (RT60) was 0.5sec. A pair of microphones with 20cm inter-spacing was located at the center of the room. The non-stationary harmonic noise, speech (desired signal) and background noise sources were generated at 1.5 m from the microphones in the direction of 75°, 90° and 105°, respectively. We compared the proposed algorithm performance to the conventional GCC-PHAT, Denda's method [1], local peak weight (LPW) [2], and SNR [3] based methods with correction rate [%] of direction estimation for each samples. As mentioned above, the ground truth direction of speech $\theta_s$ was 90°. Table 2 shows the desired sound source direction estimation performance of conventional and proposed methods. It shows that the proposed method attained improved performance under the ensuing harmonic and non-stationary noise environment. We achieved the most robust performance using LSTM based non-linear and temporal modeling, which other algorithm lacks as shown by the table below.

**Table 2.** Desired sound source direction estimation performance

| Correction rate of direction estimation (%) | Noisy input SNR (dB) | | | |
|---|---|---|---|---|
| | -5 | 0 | 5 | Avg. |
| GCC-PHAT (baseline) | 13.3 | 43.3 | 60.0 | 38.9 |
| Denda's weight [1] | 53.7 | 76.0 | 83.2 | 71.0 |
| SNR based weight [2] | 40.3 | 69.3 | 90.3 | 66.6 |
| LPW based weight [3] | 49.3 | 66.3 | 94.3 | 70.0 |
| **Proposed LSTM based weight** | **76.3** | **86.3** | **97.7** | **86.8** |
| At SNR over 10dB, the correction rate of both conventional and proposed methods shows over 95%. | | | | |

## 2.2 (Task 2) Explore effective multichannel signal enhancement techniques

To date, deep-learning approaches to far-field acoustic signal enhancement, particularly those that incorporate a Denoising Auto-Encoder (DAE), have had great success when applied to single-channel audio signals. However, the use of DAEs for multiple channels faces a number of challenges. The main reason for this is that phase information in the time-frequency domain plays a vital role in delivering the spatial information of multi-channel signals. Modelling this phase difference from time-domain or time-frequency domain requires large amounts of data to cover the various spatial configurations, while the effective way of utilizing spatial information in DAE structure is still under researching.

As an alternative to a direct DAE based approach, a conventional beamformer can be introduced prior to the implementation of the DAE. The spatial information utilized by the beamformer in the form of the ratio of acoustic transfer functions, i.e. Relative Transfer Functions (RTFs), is characterized by the path between the

speaker and each microphone. However, the modelling ability of the DAE is limited when applied to single-channel beamformer output.

***Issue : What bemforming structure would be effective for utilizing the modelling ability of the DAE?***

Therefore, in Task 2, a novel structure of multichannel signal enhancement system which adopts a DAE as part of the beamformer is proposed. The proposed structure of the Generalized Sidelobe Canceller (GSC) generates enhanced multi-channel signals, instead of merely one channel, to which the following DAE can be applied. Because the beamformer exploits spatial information and compensates for differences in the transfer functions of each channel, the proposed technique is expected to resolve the difficulty of modelling relative transfer functions consisting of complex numbers which are hard to model with a DAE. As a result, the modelling capability of the DAE can concentrate on removing artefacts caused by the beamformer. Unlike conventional beamformers, which combine these artefacts into one channel, they remain separated for each channel in the proposed method. As a result, the DAE can remove the artefacts by referring to other channels.

We use GSC which can estimate noise statistics adaptively and can be implemented using only the direction of target speech. We assume that the beamformer is working on each frame instead of a whole utterance so that the proposed algorithm can be applied to not only recognition system, but also the real-time communication system. This approach estimates noise statistics by each frame, thus eliminating the need to model noise in advance. Since no prior information of noise statistics is provided, limiting noise types to predefined ones can also be avoided.

**Result:**

**- We developed a novel structure of beamformer for**

**① Exploiting spatial information and generates multi-channel enhanced signals on which the following DAE can act.**

**② Taking advantage of the multi-channels by modelling the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels.**

The proposed system is illustrated in Figure 4. $z_m(l,k)$ is the received signal and noise at the *m*-th sensor in the short-time Fourier transform domain.

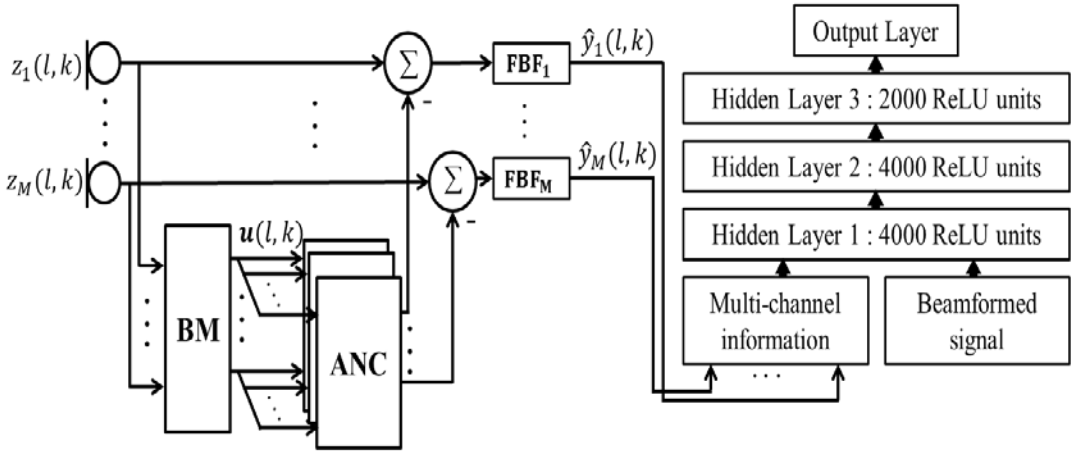


**Fig. 4** Structure of the proposed GSC and DAE

The BM is designed to project the input signals into the orthogonal complement of the target signal RTF. The filter weights of BM can be calculated using the target signal RTF. In our case, the satisfactory estimation of target signal RTF is not feasible due to noise being non-stationary. As a result, the RTF is

simplified as a pure time delay and calculated from the estimated target signal direction. In the proposed system, each individual ANC filter is adapted separately to minimize each input signal by removing the noise component estimate.

$$\hat{\mathbf{q}}_m(l,k) = argmin_{\mathbf{q}_m} E\{\|z_m(l,k) - \mathbf{q}_m^{\mathrm{H}}(l,k)\mathbf{u}(l,k)\|^2\}, \quad (1)$$

where $\mathbf{q}_m$ is the ANC coefficient corresponding to the m-th channel. The FBF takes $M$ enhanced channel signals from $M$ separated ANC filters and compensates the RTF to generate the multi-channel output features:

$$\hat{y}_m(l,k) = w^*_{FBF,m}(l,k)\left(z_m(l,k) - \mathbf{q}_m^{\mathrm{H}}(l,k)\mathbf{u}(l,k)\right), \quad (2)$$

where $w^*_{FBF,m}(l,k)$is the mth channel component of the fixed beamformer and this compensate the time delay between each channel using estimated target signal direction. Note that the proposed structure has the same filter coefficient as a conventional GSC if the output signals are summed into one channel. The distortion caused by imperfect noise cancellation is placed on the multi-channel spectrum domain of a two-dimensional space with a channel axis and frequency axis for each frame. The DAE is expected to model the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels.

To analyze effectiveness of the proposed structure, target signal enhancement is conducted in a manner similar to [4], as depicted in Figure 4. Note that the use of mask estimation before the beamformer and more sophisticated DAE structures are not considered because these improvements can be used in both the conventional and the proposed system. This experiment aims to judge the effectiveness of the proposed algorithm in its most typical configuration. The proposed system uses the results of the proposed GSC as the multi-channel information and their summation as the beamformed signal. In the baseline system, the output of the conventional GSC is used as the beamformed signal, and the noisy input signal itself (GSC-NOISY) [4] and the interaural phase difference (GSC-IPD) [5] are used as the multichannel information. To assess the advantages of using multi-channel information, these systems are also compared with a single channel DAE (GSC-ONLY) which uses only the conventional GSC output without multi-channel information.

To evaluate performance, six-channel data from CHiME [6] is used. This database provides noisy signal recorded using 6 microphones. The signal to distortion ratio (SDR) which is defined as energy ratio criteria [6] and the short-time objective intelligibility (STOI) described in [7] are used to measure speech enhancement performance. The word error rate in automatic speech recognition (ASR) is scored with an acoustic model trained on a clean database. The LibriSpeech database [8] is used in a time delay neural network based ASR system [6]. Note that ASR evaluation is performed in mismatched conditions in terms of noise and RIRs on the assumption that target signal enhancement is performed without prior knowledge of the environment. Evaluation results show that the proposed method consistently outperforms the conventional methods. Note that the STOI score is expected to have a monotonic relation with subjective speech intelligibility, where a higher value denotes more intelligible speech.

**Table. 3.** Evaluation results for target signal enhancement and word error rate

| Score / Mult. Info. | SDR | STOI | WER(%) |
|---|---|---|---|
| Noisy input | -0.694 | 0.674 | 84.43 |
| GSC- ONLY | 7.915 | 0.835 | 30.57 |
| GSC-NOISY | 7.320 | 0.837 | 27.13 |
| GSC-IPD | 7.445 | 0.835 | 26.74 |
| Proposed | 8.687 | 0.856 | 20.83 |

## 2.3 (Task 3) Develop an acoustic event recognition for aerial robot platform

One of the fundamental issues in deep learning is availability of large labeled data set. It has been consistently shown over the last decade that larger labeled data set with deeper network layers can lead to improved results. However, it is not easy to collect large amounts of labeled data, especially in Acoustic Event Recognition (AER) for specific target event. Hence, it is necessary to transfer knowledge for domain specific event recognition task from the network independently trained by a relatively large acoustic DB.

***Issue :*** *To achieve robustness of acoustic event (scene) classification, a powerful mitigating approach would be to provide a large database made available for training. What deep learning based approaches can be rendered effective for generating useful training database?*

The "transfer learning" scheme aims at transferring knowledge between the source domain used for pre-training and the target domain of interest [10]. In computer vision, transfer learning overcomes deficit of target domain training samples by adapting classifiers that are pre-trained for other large-scaled DB [11]. In recent VOC fields, CNN based supervised transfer learning methods pre-train lower layers in source domain first and then transfer these lower layer parameters for training target domain categories [11]. Transfer learning can address the issue of AER DB being significantly smaller compared to that of other audio signal applications. Therefore, we proposed to pre-train a classifier with large-scaled source domain DB and transfer the parameters for training with target DB. Figure 5 shows the proposed structure for transfer learning.

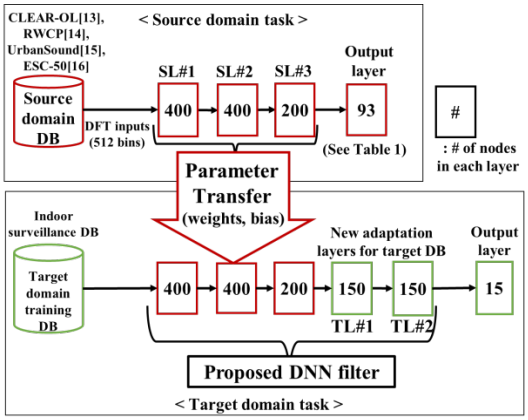


**Fig. 5** Structure of the proposed transfer learning based AER

Additionally, we explored using generative model for DB augmentation. To generate additional samples using the training DB, we proposed to use Generative Adversarial Net (GAN). The GAN learns two sub-networks: a generator and a discriminator. The discriminator reveals whether a sample is generated or real, while the generator produces samples to pass through the discriminator as real data. Although additional data generated by GAN may lead to improved classifier training, it is not clear whether every data point generated by GAN would have equal impact in classifier performance. As it has been shown by Support Vector Machine (SVM), those support vectors that reside near decision boundary are generally crucial in providing key information for classification [12]. It is believed that the performance could be improved by selecting the generated data by measuring decision value (distance) from decision hyper-plane of SVM for each class. Figure 6 shows GAN based iteration routine for DB augmentation.

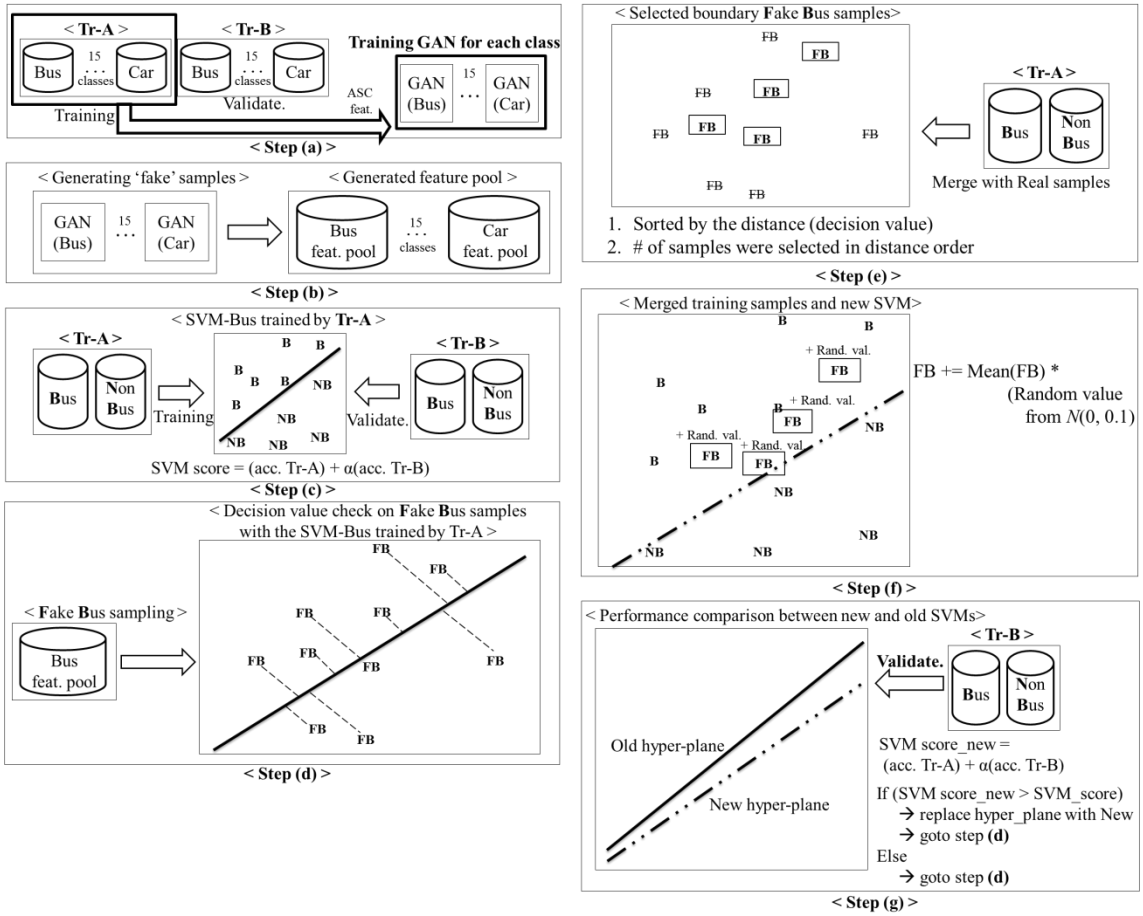


**Fig. 6** The iterative routine of the DB generation and selection

**Result:**

**- To overcome the deficit of training DB and thereby improve the performance in acoustic scene classification and event recognition, we proposed to**

**① Use transfer learning for utilizing information from DB which has relatively large amount of volume and various class types.**

**② Incorporate a GAN based DB augmentation approach using SVM criterion.**

In transfer learning based approach, as shown in Figure 5, the network for source domain is composed of three hidden fully connected layers which use sigmoid activation function and a single output layer with a SoftMax function. For filter training in the target domain, similar to the transfer learning in VOC [10], the output layer of the pre-trained network is removed and two hidden fully connected layers and a new single output layer are added to enable adaptation. Because the transferred layers have been pre-trained to classify various classes within the source domain, the layer outputs may capture the discriminative features of different sounds [10]. In target domain training, the outputs of the transferred layer are adapted to target domain labels by using them as inputs for training the additional two hidden layers. In summary, the parameters for layers SL#1-3 are first trained in the source domain then transferred to the target domain and fixed. Only the additional adaptation layers (TL#1-2) are trained using the target domain training data.

After the target domain training step, output layer and activation functions of the last hidden layer (TL#2) are removed. This process is motivated by the bottleneck feature studies [17], which follow a similar approach in using DNN mid-layers and demonstrate effective performance. Finally, five hidden layers from SL#1 to TL#2 are used as a DNN filter and the output values of layer TL#2 without activation function are used as the input features for the AEC system.

In GAN based approach, as shown in Figure 6-step (a), a GAN for each class was trained using a part of the training set, which excludes the validation part for following steps. Using the trained GANs, we generated 'fake' samples and organized the sample feature pools for each class as shown in step (b). Before using the generated samples, an SVM hyper-plane for each class (target class vs. the others) was first determined from the real data set to establish a baseline performance. We chose the bus class as an example. Note that half of the training set was used for training and the other half was used for validating SVM performance. As shown in step (c), we checked classification performance of SVM with sum of the training and validation set accuracy. Considering the SVM update in the next step, we added a weight ($\alpha$, which is bigger than 1) to the unseen data, i.e. validation accuracy. In step (d), we subsampled 'fake bus' features from the generated bus feature pool and checked decision values on the SVM hyper-plane trained from Tr-A set. As shown in step (e), we sorted the fake samples by the distance order, and chose a preset number of the nearest samples. Additionally, we also included small number of samples near the hyper-plane that were classified as non-bus by handicapping their decision value. We then merged the near boundary fake samples with the real samples of Tr-A set. Step (f) shows the new SVM hyper-plane trained by the merged set. Before training the new SVM, we added random vectors, which are scaled to the magnitude of the samples, to reduce the sample bias of the generation using GAN. As was done in step (c), the classification performance of new SVM was checked with the sum of the training (Tr-A) and validation set (Tr-B) accuracy. If the accuracy score of the new SVM outperforms the previous SVM score, the reference SVM hyper-plane was replaced with the new one and the iteration continues again with the fake sample subsampling in the step (d). If not, the iteration proceeds to the step (d) without replacing the reference hyper-plane. Once the SVM performance is optimized, the associated support vectors of fake bus features were used for the augmented training set. The entire process is repeated with the Tr-B as the training set for GAN and SVM, and Tr-A as the validation set. The whole processes are repeated for each acoustic scene class.

Table 4 shows the source domain DB for transfer learning. The target indoor surveillance DB consists of 15 events (a crying child, breaking glass, water drops, chirping birds, a doorbell, home appliance beeping, screaming, a dog barking, music, speech, a cat meowing, a gunshot, a siren, an explosion, and footsteps). For checking noise robustness of AER, noise was added to the event DB at 5, 10, 15 dB SNR. In addition, compared with other DNN-based feature extraction methods, such as the Deep Belief Network (DBN) feature, which is used for music genre classification [18], and DNN bottleneck feature [17], the proposed method demonstrated improved accuracy by effectively utilizing the information transferred from the source domain. Table 5 shows performance comparison with aforementioned conventional approaches.

**Table 4.** Source domain database description

| DB set | Contents |
|---|---|
| Clear-OL [13] | Alert, cough, door slam, drawer, key, keyboard, knocking, laughing, mouse, page turn, pen drop, phone, printer, speech, switch, clear throat |
| RWCP [14] | Air-cap, bell, break stick, buzzer, castanet, ceramic collision, clap, clock ringing, coin, cymbals, drum, dryer, grinding coffee, kara, maracas, metal collision, article dropping, plastic collision, pump, punch stapler, rubbing, shaver, spray, string, tambourine, toy, whistle, wood collision |
| Urban-Sound [15] | Air-conditioner, dog bark, drilling, engine idling, car horn, jackhammer, children playing, siren, street music, shot |
| ESC-50 [16] | Airplane, breathing, brushing teeth, can opening, cat, chainsaw, chirping birds, church bells, clapping, clock alarm, clock tick, coughing, cow, crackling fire, crickets, crow, door - wood creaks, door knock, drinking – sipping, engine, fireworks, footsteps, frog, hand saw, helicopter, hen, insects (flying), pig, pouring water, rooster, sea waves, sheep, sneezing, snoring, thunderstorm, toilet flush, vacuum cleaner, washing machine, wind |
| **Total 93 classes / The similar classes from the different DB set had been merged / 16 kHz resampled, 16 bit resolution** | |

**Table 5. Average acoustic event classification rate [%] for ETSI background noise using various features with SVM classifier**

| | Living room noise | | | Office noise | | | Clean DB | **Average** |
|---|---|---|---|---|---|---|---|---|
| SNR [dB] | 5 | 10 | 15 | 5 | 10 | 15 | | |
| MFCC | 79.7 | 85.5 | 94.5 | 81.1 | 87.6 | 95.1 | 96.1 | 88.5 |
| DBN feature [18] | 86.4 | 89.9 | 93.9 | 89.9 | 93.3 | 95.7 | 96.4 | 92.2 |
| DNN-bottleneck feature [17] | 86.3 | 90.9 | 95.5 | 90.7 | 92.5 | 95.9 | 96.5 | 92.6 |
| **Proposed transfer learning approach** | **92.5** | **96.3** | **96.3** | **93.7** | **96.5** | **96.5** | **98.9** | **95.8** |

Table 6 shows performance comparison between original DB set and GAN based augmented DB set. We used IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 task 1 DB [21] for Acoustic Scene Classification (ASC). It contains 15 different acoustic scene classes such as, Bus, Café, Car, City center, Forest path, Grocery store, Home, Lakeside beach, Library, Metro station, Office, Residential area, Train, Tram, and Park. We used Discrete Fourier Transform (DFT) based feature and Mel-Filtered Bank (MFB) as feature input, and Fully Connected Neural Network (FCNN) and SVM for classifier. Based on the experimental results of AER and ASC, we achieved improved performance in noisy surveillance environment utilizing information of universal background DB and generative method. Additionally, as shown in Figure 7, we achieved the best performance in DCASE 2017 grand challenge Task using the GAN based approach.

**Table 6**. Comparing the performance of the conventional and the proposed method (average accuracy on 4-fold validation of DCASE 2017 development set)

| Avg. acc. [%] | with original development set | | | | with augmented set | | | |
|---|---|---|---|---|---|---|---|---|
| | DFT-FCNN | MFB-FCNN | DFT-SVM | MFB-SVM | DFT-FCNN | MFB-FCNN | DFT-SVM | MFB-SVM |
| | 75.4 | 75.1 | 78.2 | 79.3 | 83.2 | 83.7 | 81.6 | 85.6 |


**Fig. 7** IEEE DCASE challenge 2017 task 1 results, (http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results)

## 3. List of relevant papers published by this project

- **[Task 1] Investigate novel sound source localization techniques in highly non-stationary noise dominant environments**
  - Seongkyu Mun, Suwon Shon, Wooli Kim, David K. Han, and Hanseok Ko, "Acoustic Signal based Noise Robust Speaker Direction Estimation using Recurrent Neural Network", *IEICE transactions on Information & System*, 2017 [submitted]

- **[Task 2] Explore various multichannel signal enhancement techniques**
  - Minkyu Shin, Seongkyu Mun, David K. Han and Hanseok Ko, "New Generalized Sidelobe Canceller with Denoising Auto-Encoder for Improved Speech Enhancement", *IEICE transactions on Fundamentals of Electronics Communications and Computer Sciences*, vol.100-A, no.12, Dec, 2017

- **[Task 3] Develop the acoustic event recognition system for aerial robot platform**
  - Seongkyu Mun, Sangwook Park, David K. Han and Hanseok Ko, "Generative Adversarial Network based Acoustic Scene Training Set Augmentation and Selection using SVM Hyper-Plane", *IEEE DCASE challenge workshop*, Nov. 2017 **[Winner of the Grand Challenge Task 1]**
  - Seongkyu Mun, Minkyu Shin, Suwon Shon, Wooil Kim, David K. Han and Hanseok Ko, "DNN Transfer Learning based Non-linear Feature Extraction for Acoustic Event Classification", *IEICE transactions on Information and Systems*, Vol. 100-D, No. 9, pp. 2249-2252, Sep. 2017
  - Seongkyu Mun, Suwon Shon, Wooli Kim, David K. Han, and Hanseok Ko, "A Novel Discriminative Feature Extraction for Acoustic Scene Classification using RNN based Source Separation", *IEICE transactions on Information & System*, 2017 [in press]

## 4. Conclusions and future work

We explored three key-techniques for robust acoustics and speech perception of aerial robot for scene understanding during critical emergency response missions. For noise robust sound source localization, we proposed a noise robust desired sound direction estimation method using LSTM based weighting function. The direction estimation experiments confirmed that the proposed method shows improved robustness under indoor surveillance noise environment characterized by presence of harmonic or non-stationary noise sources. Our future work will investigate effective methods for applying RNN method to phase-spectrogram as Non-negative Matrix Factorization (NMF) was applied to phase-spectrogram previously in [19].

In terms of our signal enhancement performance under noisy environment, the GSC exploits spatial information and generates multi-channel enhanced signals on which the following DAE can act. As a result, the DAE can take advantage of the multi-channels by modeling the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels. The evaluation results demonstrate that utilizing the results of the proposed GSC structure as an input to the DAE is effective in improving noise reduction and speech recognition performance.

To improve acoustic event recognition performance and overcome the deficit of acoustic event resource, we proposed a novel DNN based transfer learning approach. By utilizing the information transferred from the universal source domain, the proposed approach was characterized by improved AEC accuracy in indoor surveillance experiments. Once DNN filter training has been completed in the source domain, this DNN filter can be utilized in other domains, repeatedly. Therefore, future work will investigate an effective transfer learning scheme for various acoustic applications and determine how performance changes depending on the configuration of the data.

The acoustic perception during the hovering and moving of the aerial robot, which will be conducted in the future steps, will be more challenging task, due to the severe noisy environment. Furthermore, we assumed that an only single sound event occurs within a restricted event class in this phase, but in a real environment hundreds of multiple sounds occur simultaneously. To address the issue above, the next research phase is to

investigate acoustic event recognition using the Google Audio-Set [20]. The Google Audio-Set is acoustic DB based on real life video uploaded in YouTube. It is the latest and largest video-based sound event recognition DB released in March 2017 with 5.8K hour long consisting of 527 sound classes in total. Based on the acoustic database, we will address the DB deficit issue of this year and investigate the simultaneous occurrence of the hundreds of sounds mentioned above. In the IEEE DCASE 2017 challenge task 4 [21], there was a competition using the Audio-Set DB. It was the competition to recognize 17 types of warning and vehicle sounds for a self-driving smart car environment similar to aerial robot environments. The participating teams using the convolutional recurrent neural network showed the best performance in the competition. As a future plan, we plan to explore the domain transformation (adaptation) for the aerial robot environment using the GAN-based various approaches and the audio feature augmentation using other generative methods such as, Variational Auto Encoder (VAE).

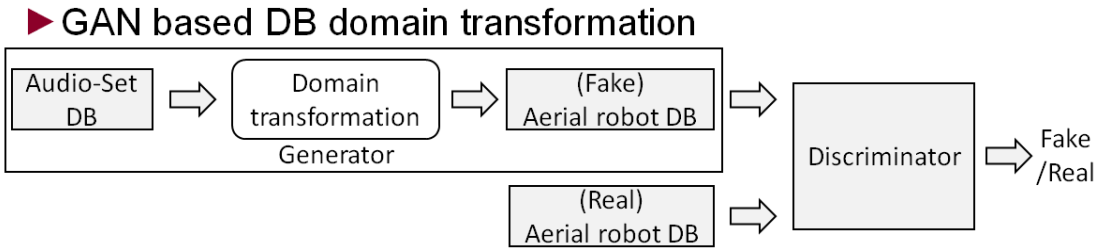


**Fig. 8** The example structure of GAN based domain transformer

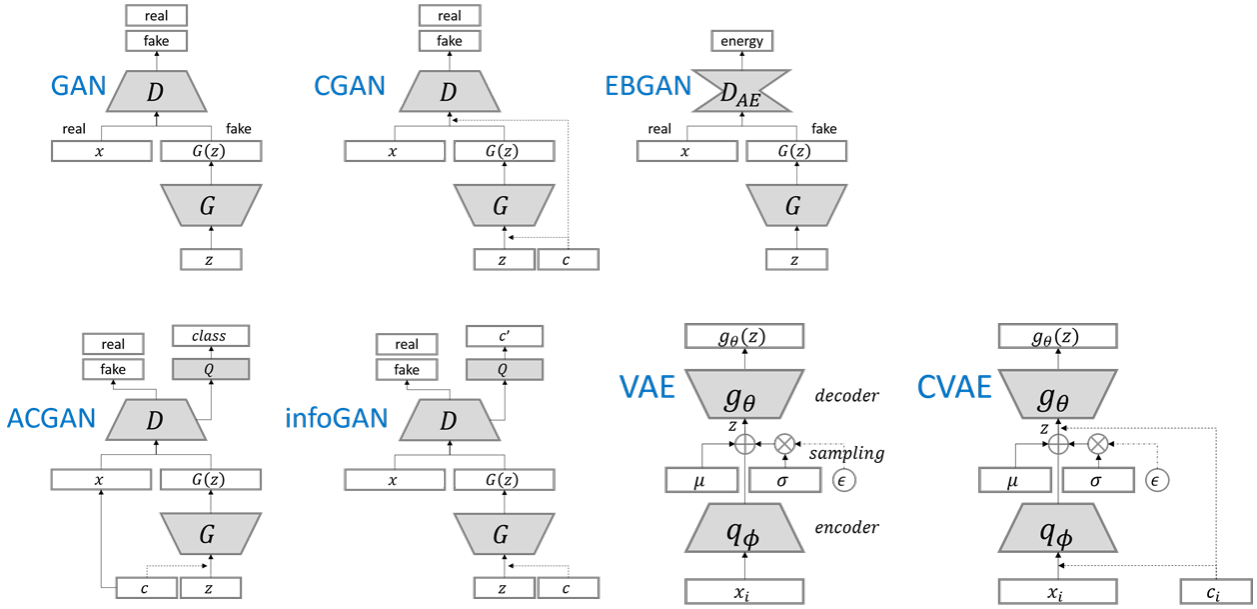


**Fig. 9** Possible structure examples of generative approaches [22]

## 5. References

[1] Y. Denda, T. Nishiura, and Y. Yamashita, “Robust talker direction estimation based on weighted csp analysis and maximum likelihood estimation” IEICE Trans. on Information and Systems, vol. E89-D, no. 3, pp. 1050–1057, Jan. 2006

[2] I. Markovic et. al., “Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering”, Robotics and Autonomous Systems, vol. 58, no. 11, pp.1185-1196, Jan. 2010

[3] O. Ichikawa, T. Fukuda, and M. Nishimura, “DOA estimation with local-peak-weighted CSP” EURASIP Journal on Advances in Signal Processing, vol. 2010, no.1, pp. 1–10, Jan. 2010.

[4] P. Swietojanski, A. Ghoshal, and S. Renals, “Hybrid acoustic models for distant and multichannel large vocabulary speech recognition,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 285–290, 2013

[5] M. I. Mandel, R. J. Weiss, and D. Ellis, “Model-Based Expectation-Maximization Source Separation and Localization,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[6] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, 2006.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech,” IEEE Trans., vol. 19, no. 7, pp. 2125–2136, 2011.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2015–Augus, pp. 5206–5210, 2015

[9] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2015–Janua, pp. 3214–3218, 2015.

[10] M. Oquab et al, “Learning and transferring mid-level image representations using convolutional neural networks,” IEEE Conf. on Computer Vision and Pattern Recog. (CVPR), Columbus, USA, pp. 1717-1724, June 2014.

[11] J. Gehring et al, “Extracting deep bottleneck features using stacked auto-encoders,” IEEE Int. Conf. on Acoustics, Speech and Signal Proc., Vancouver, Canada, pp. 3377-3381, May 2013.

[12] C. Cortes and V. Vapnik, “Support-vector networks”, Ma-chine learning, vol. 20, no.3, pp. 273-297, 1995.
[13] A. Temko et al, “CLEAR evaluation of acoustic event detection and classification systems,” Proc. of Int. Eval. Work. on Classification of Events, Act. and Relation., pp. 311–322, 2007.

[14] S. Nakamura et al,“Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition’, in Proc. of EUROSPEECH, pp. 2255–2258, 1999.

[15] J. Salamon et al, “A dataset and taxonomy for urban sound research,” ACM 2014 Int. Conf. on Multimedia, New York, USA, pp. 1041-1044, Oct. 2014.

[16] K. Piczak, “ESC: Dataset for environmental sound classification,” ACM 2015 Int. Conf. on Multimedia, Brisbane, Australia, pp. 1015-1018, Oct. 2015.

[17] S. Mun, S. Shon, W. Kim, H. Ko, “Deep neural network bottleneck features for acoustic event recognition,” Proc. of the Int. Speech Comm. Association, INTERSPEECH 2016, San Francisco, USA, pp. 2954-2957, Sep. 2016

[18] P. Hamel, D. Eck, “Learning features from music audio with deep belief networks,” Int. Society for Music Infor. Retri. Conf., ISMIR 2010, Utrecht, Netherlands, pp. 339-344, Aug. 2010.

[19] S. Shon, S. Mun, D. Han, H. Ko, “Non-negative matrix factorisation-based subband decomposition for acoustic source localization”, Electronics Letters, vol. 51, no. 22, pp 1723-1724, Oct. 2015

[20] J. F. Gemmeke, et. al., “Audio Set: An ontology and human-labeled dataset for audio events”, ICASSP 2017, Mar. 2017.

[21] http://www.cs.tut.fi/sgn/arg/dcase2017/

[22] https://github.com/hwalsuklee/tensorflow-generative-model-collections

# Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency

Hanseok Ko (Korea University)

This report describes the research work and results of the project entitled: **Robust acoustics and speech perception of aerial robot under ego noise for scene understanding during critical emergency response missions**, which took place during 2016.08.08~2017.08.07. Authors gratefully acknowledge the support of this research by the Air Force Office of Scientific Research.

## 1. Research Objectives

In this research, Hanseok Ko as Principal Investigator and his team of graduate students investigated acoustic signal based robust scene understanding techniques for aerial robot. Due to the added advantage of flying besides having the usual movement dexterity, an aerial robot creates a new set of opportunities to provide highly intelligent perceptions to engage and respond to time-critical emergency situations. It is desirable to equip aerial robots with the ability to detect evidence of emergency situations such as calling for help by human voice via using a microphone array to listen to surrounding zone, perform speech/acoustic event detection, and conduct sound source localization for directivity assessment and for possible tracking if the source is either stationary or moving.

Despite of its potential capability in emergency response missions over hazardous sites and observing wide areas, auditory processing in aerial robots is technically extremely challenging due to the inherent issues delineated as follows.

- High level of wind noise and ego-noise from rotor
- Constantly changing noise level and target to sensor distance while robot is moving
- Wide dynamic range of target signal power by changing target to sensor distance in outdoor environment
- High probability of overlapped acoustic events in wide searching area
- Difficulty of reliable target detection due to ever-present ego-noise
- Robust acoustic scene classification

To address the above issues and explore mitigating approaches, a 3-phase plan was set up at the onset of this project as shown in Figure 1. We explored the audio perception techniques under **idle state** as goal of 1$^{st}$ year.

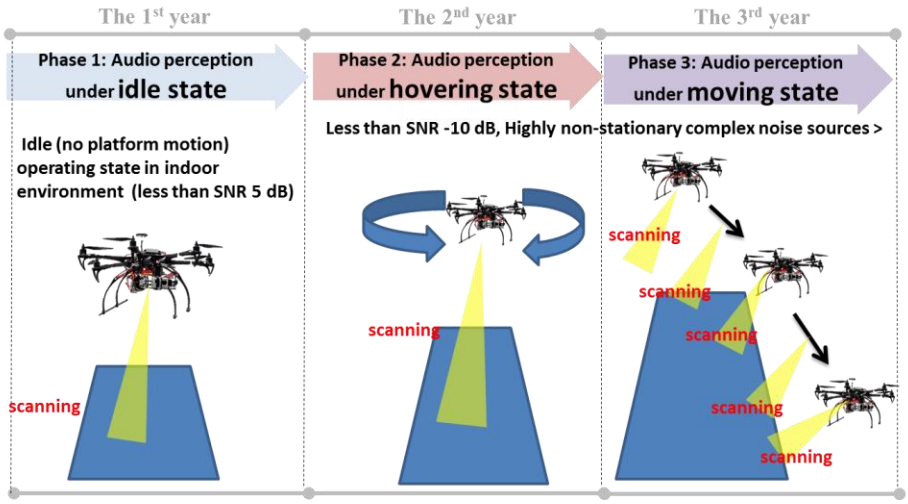


**Fig. 1** 3-phase (3-years) plan for audio perception under aerial robot environments

In order to achieve the research goal of Phase 1, **we investigated the following 3 specific research issues.**

(1) *How can we select desired signal dominant frequency bin in non-stationary noisy environment for* ***robust sound source localization****?*

(2) *For* ***signal enhancement****, what beamforming structure would be effective for utilizing the modeling ability of deep neural network?*

(3) *To achieve robustness of acoustic event (scene) classification, a powerful mitigating approach would be to provide a large database made available for training. What deep learning based approaches can be rendered effective to generate useful training database?*

## 2. Research tasks and results

### 2.1 (Task 1) Investigate novel sound source localization techniques in highly non-stationary noise dominant environments

We investigated the problem of finding desired signal using a microphone array in highly non-stationary noisy environment. Frequency bins with high noise levels can be inadvertently considered as desired sound sources. Reliable estimation performance can be achieved by using meaningful signal-dominant frequency bins while avoiding noise-dominant bins.

***Issue : How can we select desired signal dominant frequency bin in non-stationary noisy environment for robust sound source localization?***

To address this issue, hence, selecting (or weighting high) desired signal frequency bin in non-stationary noise, a weighting function is investigated using Deep Neural-Network (DNN).

Using DNN approach for selecting desired signal frequency turned out to be effective. In DNN based time-frequency mask estimation step, the ideal binary masks for signal are set as

$$\widehat{M_X}(l,\omega) = \begin{cases} 1, & \frac{\|X(l,\omega)\|}{\|N(l,\omega)\|} > threshold \\ 0, & else \end{cases} \quad (1)$$

where $X(l,\omega)$ and $N(l,\omega)$ are the desired signal component and noise component in $l_{\text{th}}$ frame of input noisy signal in STFT domain respectively. A DNN composed of LSTM and feedforward layers is trained to estimate the ideal binary masks, $\widehat{M_X}(l,\omega)$ from a noisy spectrum. The noisy spectrum is generated for the training step by summing the signal and noise signal which is recorded in real environments. The specific structure of DNN is shown in the table below,

**Table 1.** LSTM network configuration for mask estimation

| Layers | Units | Type | Non-linearity | $p_{\text{dropout}}$ |
|---|---|---|---|---|
| L1 | 256 | LSTM | Tanh | 0.5 |
| L2 | 513 | FF | ReLU | 0.5 |
| L3 | 513 | FF | ReLU | 0.5 |
| L4 | 1026 | FF | Sigmoid | 0.0 |

The window size and frame shift are set at 1024 and 512 samples, respectively, at a 16 kHz sampling rate. $p_{\text{dropout}}$ means the ratio of nodes to apply dropout technique in training step. The DNN consists of one Long Short-Term Memory (LSTM) layer and three Feed-Forward (FF) layers.

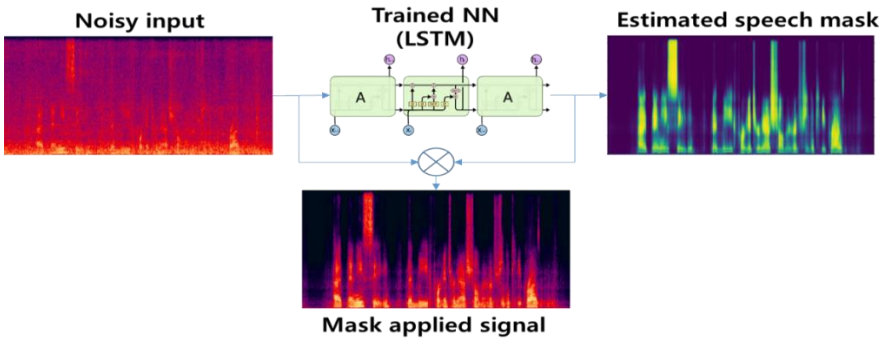


**Fig. 2** LSTM based speech mask estimation process (speech as desired signal)

**Result:**

**- We developed a novel masking method for sound source localization in the noisy environments by:**

**① Using the DNN based approaches for modeling the relationship between the noisy and clean signals as a nonlinear transformation.**

**② Exploiting the sequential information from the previous adjacent frames using feed-back of LSTM network, which conventional approaches cannot utilize.**

After applying DNN based mask to each channel of input signal, Time Difference Of Arrival (TDOA) is estimated for each pair of microphones to find location of acoustic source. The TDOAs can be estimated from so-called angular spectrum, whose peaks indicate the TDOA of the source. Generalized Cross Correlation-Phase Transform (GCC-PHAT) is used as an angular spectrum in this system. GCC of the $m_{th}$ and $n_{th}$ microphone signals is calculated as eq. (2)

$$\mathrm{R_{mn}}(l,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{mn}(l,\omega) X_m(l,\omega) X_n^*(l,\omega) e^{j\omega\tau} d\omega, \tag{2}$$

where $\Psi_{mn}(l,\omega)$ denotes a weight function. Although many different weighting functions can be applied, the PHAT has been found to perform quite well under realistic acoustical conditions.

$$\Psi_{mn}(l,\omega) = \frac{1}{|X_m(l,\omega)X_n^*(l,\omega)|} \tag{3}$$

The Figure 3 shows the outline of GCC-PHAT based source localization system.

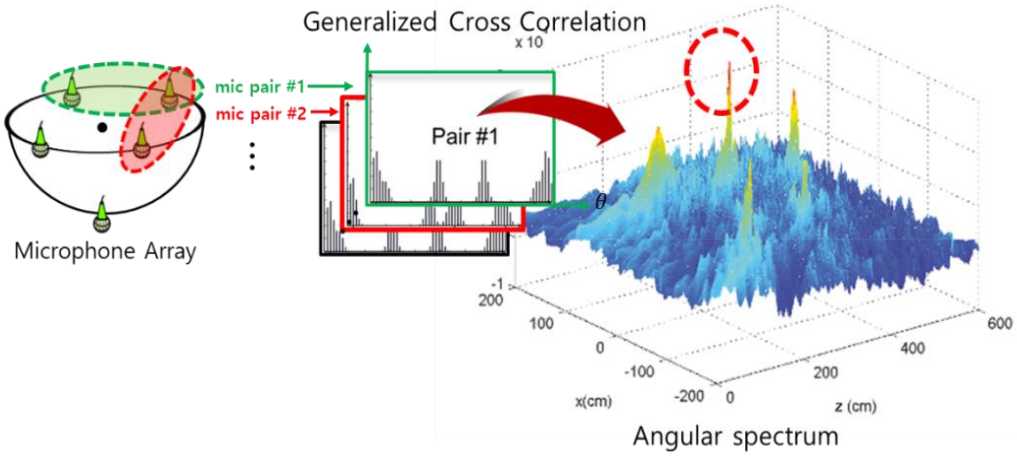


**Fig. 3** GCC-PHAT based source localization system

The proposed algorithm was evaluated in a 5.0 × 6.0 × 2.5 $m^3$ simulated room environment using image source method. The reverberation time setting (RT60) was 0.5sec. A pair of microphones with 20cm inter-spacing was located at the center of the room. The non-stationary harmonic noise, speech (desired signal) and background noise sources were generated at 1.5 m from the microphones in the direction of 75°, 90° and 105°, respectively. We compared the proposed algorithm performance to the conventional GCC-PHAT, Denda's method [1], local peak weight (LPW) [2], and SNR [3] based methods with correction rate [%] of direction estimation for each samples. As mentioned above, the ground truth direction of speech $\theta_s$ was 90°. Table 2 shows the desired sound source direction estimation performance of conventional and proposed methods. It shows that the proposed method attained improved performance under the ensuing harmonic and non-stationary noise environment. We achieved the most robust performance using LSTM based non-linear and temporal modeling, which other algorithm lacks as shown by the table below.

**Table 2.** Desired sound source direction estimation performance

| Correction rate of direction estimation (%) | Noisy input SNR (dB) | | | |
|---|---|---|---|---|
| | -5 | 0 | 5 | Avg. |
| GCC-PHAT (baseline) | 13.3 | 43.3 | 60.0 | 38.9 |
| Denda's weight [1] | 53.7 | 76.0 | 83.2 | 71.0 |
| SNR based weight [2] | 40.3 | 69.3 | 90.3 | 66.6 |
| LPW based weight [3] | 49.3 | 66.3 | 94.3 | 70.0 |
| **Proposed LSTM based weight** | **76.3** | **86.3** | **97.7** | **86.8** |
| At SNR over 10dB, the correction rate of both conventional and proposed methods shows over 95%. | | | | |

## 2.2 (Task 2) Explore effective multichannel signal enhancement techniques

To date, deep-learning approaches to far-field acoustic signal enhancement, particularly those that incorporate a Denoising Auto-Encoder (DAE), have had great success when applied to single-channel audio signals. However, the use of DAEs for multiple channels faces a number of challenges. The main reason for this is that phase information in the time-frequency domain plays a vital role in delivering the spatial information of multi-channel signals. Modelling this phase difference from time-domain or time-frequency domain requires large amounts of data to cover the various spatial configurations, while the effective way of utilizing spatial information in DAE structure is still under researching.

As an alternative to a direct DAE based approach, a conventional beamformer can be introduced prior to the implementation of the DAE. The spatial information utilized by the beamformer in the form of the ratio of acoustic transfer functions, i.e. Relative Transfer Functions (RTFs), is characterized by the path between the speaker and each microphone. However, the modelling ability of the DAE is limited when applied to single-channel beamformer output.

***Issue : What bemforming structure would be effective for utilizing the modelling ability of the DAE?***

Therefore, in Task 2, a novel structure of multichannel signal enhancement system which adopts a DAE as part of the beamformer is proposed. The proposed structure of the Generalized Sidelobe Canceller (GSC) generates enhanced multi-channel signals, instead of merely one channel, to which the following DAE can be applied. Because the beamformer exploits spatial information and compensates for differences in the transfer functions of each channel, the proposed technique is expected to resolve the difficulty of modelling relative transfer functions consisting of complex numbers which are hard to model with a DAE. As a result, the modelling capability of the DAE can concentrate on removing artefacts caused by the beamformer. Unlike conventional beamformers, which combine these artefacts into one channel, they remain separated for each channel in the proposed method. As a result, the DAE can remove the artefacts by referring to other channels.

We use GSC which can estimate noise statistics adaptively and can be implemented using only the direction of target speech. We assume that the beamformer is working on each frame instead of a whole utterance so that the proposed algorithm can be applied to not only recognition system, but also the real-time communication system. This approach estimates noise statistics by each frame, thus eliminating the need to model noise in advance. Since no prior information of noise statistics is provided, limiting noise types to predefined ones can also be avoided.

**Result:**

**- We developed a novel structure of beamformer for**

**① Exploiting spatial information and generates multi-channel enhanced signals on which the following DAE can act.**

**② Taking advantage of the multi-channels by modelling the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels.**

The proposed system is illustrated in Figure 4. $z_m(l,k)$ is the received signal and noise at the *m*-th sensor in the short-time Fourier transform domain.

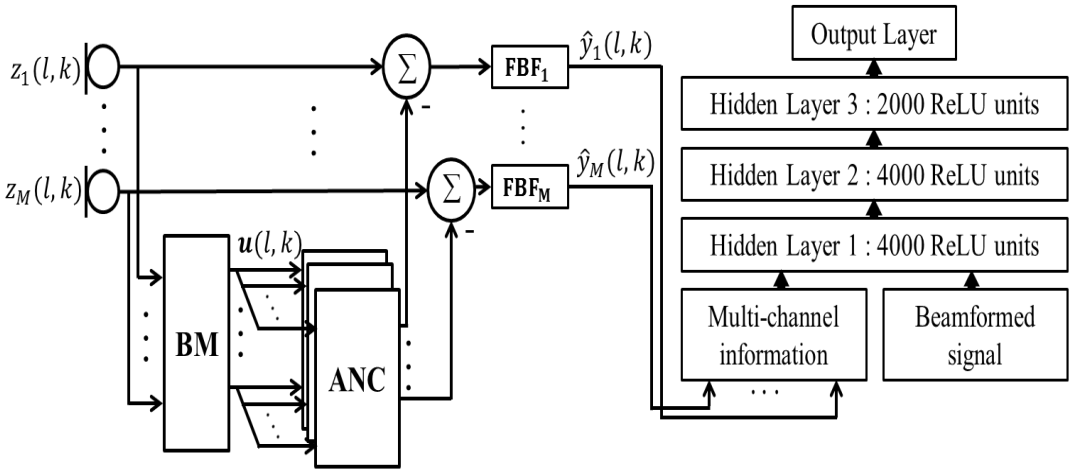


**Fig. 4** Structure of the proposed GSC and DAE

The BM is designed to project the input signals into the orthogonal complement of the target signal RTF. The filter weights of BM can be calculated using the target signal RTF. In our case, the satisfactory estimation of target signal RTF is not feasible due to noise being non-stationary. As a result, the RTF is simplified as a pure time delay and calculated from the estimated target signal direction. In the proposed system, each individual ANC filter is adapted separately to minimize each input signal by removing the noise component estimate.

$$\widehat{\mathbf{q}}_m(l,k) = argmin_{\mathbf{q}_m} E\{\|z_m(l,k) - \mathbf{q}_m^{\mathrm{H}}(l,k)\mathbf{u}(l,k)\|^2\}, \quad (1)$$

where $\mathbf{q}_m$ is the ANC coefficient corresponding to the m-th channel. The FBF takes *M* enhanced channel signals from *M* separated ANC filters and compensates the RTF to generate the multi-channel output features:

$$\hat{y}_m(l,k) = w^*_{FBF,m}(l,k)\Big(z_m(l,k) - \mathbf{q}^{\mathrm{H}}_m(l,k)\mathbf{u}(l,k)\Big), \quad (2)$$

where $w^*_{FBF,m}(l,k)$is the mth channel component of the fixed beamformer and this compensate the time delay between each channel using estimated target signal direction. Note that the proposed structure has the same filter coefficient as a conventional GSC if the output signals are summed into one channel. The distortion caused by imperfect noise cancellation is placed on the multi-channel spectrum domain of a two-dimensional space with a channel axis and frequency axis for each frame. The DAE is expected to model the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels.

To analyze effectiveness of the proposed structure, target signal enhancement is conducted in a manner similar to [4], as depicted in Figure 4. Note that the use of mask estimation before the beamformer and more sophisticated DAE structures are not considered because these improvements can be used in both the conventional and the proposed system. This experiment aims to judge the effectiveness of the proposed algorithm in its most typical configuration. The proposed system uses the results of the proposed GSC as the multi-channel information and their summation as the beamformed signal. In the baseline system, the output of the conventional GSC is used as the beamformed signal, and the noisy input signal itself (GSC-NOISY) [4] and the interaural phase difference (GSC-IPD) [5] are used as the multichannel information. To assess the advantages of using multi-channel information, these systems are also compared with a single channel DAE (GSC-ONLY) which uses only the conventional GSC output without multi-channel information.

To evaluate performance, six-channel data from CHiME [6] is used. This database provides noisy signal recorded using 6 microphones. The signal to distortion ratio (SDR) which is defined as energy ratio criteria [6] and the short-time objective intelligibility (STOI) described in [7] are used to measure speech enhancement performance. The word error rate in automatic speech recognition (ASR) is scored with an acoustic model trained on a clean database. The LibriSpeech database [8] is used in a time delay neural network based ASR system [6]. Note that ASR evaluation is performed in mismatched conditions in terms of noise and RIRs on the assumption that target signal enhancement is performed without prior knowledge of the environment. Evaluation results show that the proposed method consistently outperforms the conventional methods. Note that the STOI score is expected to have a monotonic relation with subjective speech intelligibility, where a higher value denotes more intelligible speech.

**Table. 3.** Evaluation results for target signal enhancement and word error rate

| Score / Mult. Info. | SDR | STOI | WER(%) |
|---|---|---|---|
| Noisy input | -0.694 | 0.674 | 84.43 |
| GSC- ONLY | 7.915 | 0.835 | 30.57 |
| GSC-NOISY | 7.320 | 0.837 | 27.13 |
| GSC-IPD | 7.445 | 0.835 | 26.74 |
| Proposed | 8.687 | 0.856 | 20.83 |

## 2.3 (Task 3) Develop an acoustic event recognition for aerial robot platform

One of the fundamental issues in deep learning is availability of large labeled data set. It has been consistently shown over the last decade that larger labeled data set with deeper network layers can lead to improved results. However, it is not easy to collect large amounts of labeled data, especially in Acoustic Event Recognition (AER) for specific target event. Hence, it is necessary to transfer knowledge for domain specific event recognition task from the network independently trained by a relatively large acoustic DB.

***Issue :*** *To achieve robustness of acoustic event (scene) classification, a powerful mitigating approach would be to provide a large database made available for training. What deep learning based approaches can be rendered effective for generating useful training database?*

The "transfer learning" scheme aims at transferring knowledge between the source domain used for pre-training and the target domain of interest [10]. In computer vision, transfer learning overcomes deficit of target domain training samples by adapting classifiers that are pre-trained for other large-scaled DB [11]. In recent VOC fields, CNN based supervised transfer learning methods pre-train lower layers in source domain first and then transfer these lower layer parameters for training target domain categories [11]. Transfer learning can address the issue of AER DB being significantly smaller compared to that of other audio signal applications. Therefore, we proposed to pre-train a classifier with large-scaled source domain DB and transfer the parameters for training with target DB. Figure 5 shows the proposed structure for transfer learning.

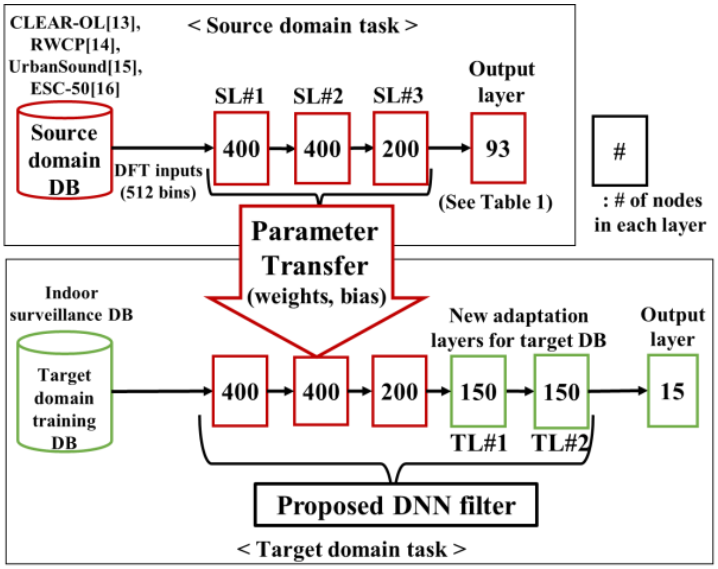


**Fig. 5** Structure of the proposed transfer learning based AER

Additionally, we explored using generative model for DB augmentation. To generate additional samples using the training DB, we proposed to use Generative Adversarial Net (GAN). The GAN learns two sub-networks: a generator and a discriminator. The discriminator reveals whether a sample is generated or real, while the generator produces samples to pass through the discriminator as real data. Although additional data generated by GAN may lead to improved

classifier training, it is not clear whether every data point generated by GAN would have equal impact in classifier performance. As it has been shown by Support Vector Machine (SVM), those support vectors that reside near decision boundary are generally crucial in providing key information for classification [12]. It is believed that the performance could be improved by selecting the generated data by measuring decision value (distance) from decision hyper-plane of SVM for each class. Figure 6 shows GAN based iteration routine for DB augmentation.

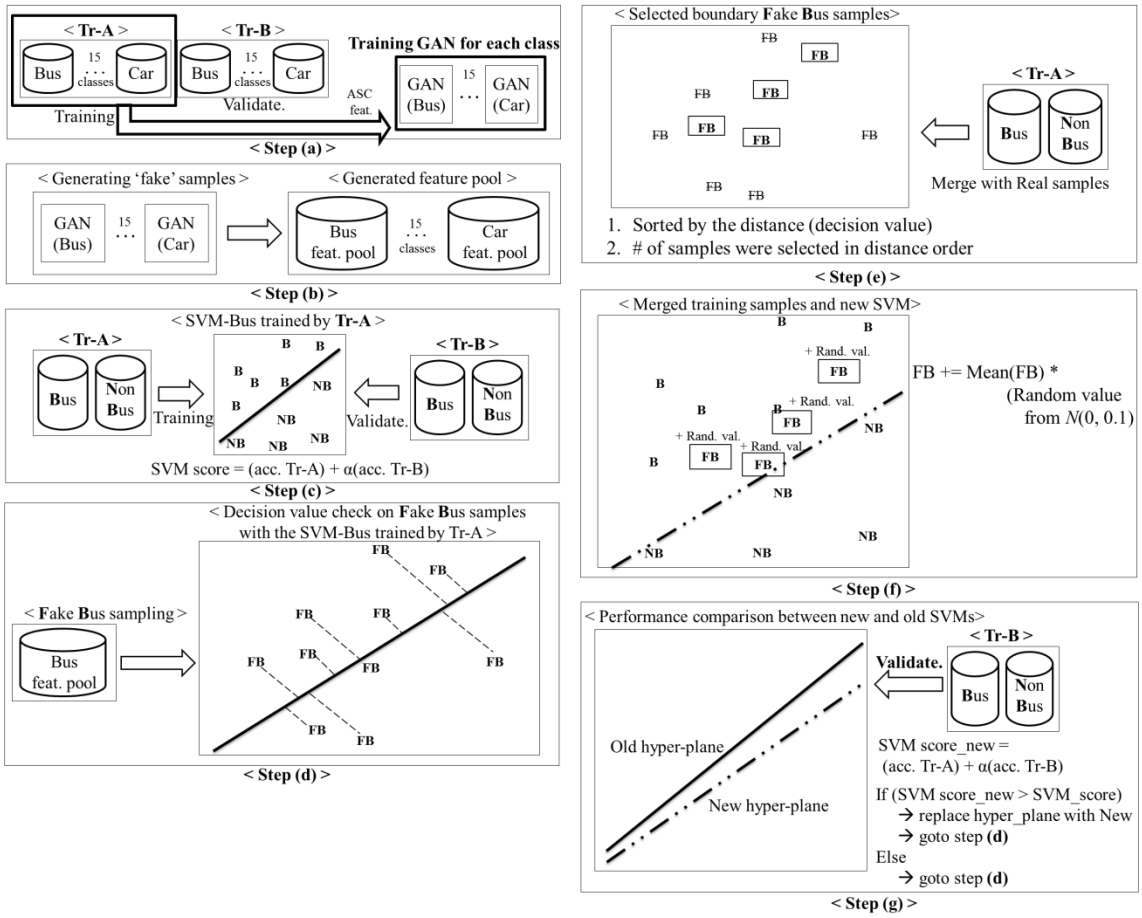


**Fig. 6** The iterative routine of the DB generation and selection

**Result:**

**- To overcome the deficit of training DB and thereby improve the performance in acoustic scene classification and event recognition, we proposed to**

① **Use transfer learning for utilizing information from DB which has relatively large amount of volume and various class types.**

② **Incorporate a GAN based DB augmentation approach using SVM criterion.**

In transfer learning based approach, as shown in Figure 5, the network for source domain is composed of three hidden fully connected layers which use a Sigmoid activation function and a single output layer with a SoftMax function. For filter training in the target domain, similar to the transfer learning in VOC [10], the output layer of the pre-trained network is removed and two hidden fully connected layers and a new single output layer are added to enable adaptation. Because the transferred layers have been pre-trained to classify various classes within the source domain, the layer outputs may capture the discriminative features of different sounds [10]. In target domain training, the outputs of the transferred layer are adapted to target domain labels by using them as inputs for training the additional two hidden layers. In summary, the parameters for layers SL#1-3 are first trained in the source domain then transferred to the target domain and fixed. Only the additional adaptation layers (TL#1-2) are trained using the target domain training data.

After the target domain training step, output layer and activation functions of the last hidden layer (TL#2) are removed. This process is motivated by the bottleneck feature studies [17], which follow a similar approach in using DNN mid-layers and demonstrate effective performance. Finally, five hidden layers from SL#1 to TL#2 are used as a DNN filter and the output values of layer TL#2 without activation function are used as the input features for the AEC system.

In GAN based approach, as shown in Figure 6-step (a), a GAN for each class was trained using a part of the training set, which excludes the validation part for following steps. Using the trained GANs, we generated ‘fake’ samples and organized the sample feature pools for each class as shown in step (b). Before using the generated samples, an SVM hyper-plane for each class (target class vs. the others) was first determined from the real data set to establish a baseline performance. We chose the bus class as an example. Note that half of the training set was used for training and the other half was used for validating SVM performance. As shown in step (c), we checked classification performance of SVM with sum of the training and validation set accuracy. Considering the SVM update in the next step, we added a weight ($\alpha$, which is bigger than 1) to the unseen data, i.e. validation accuracy. In step (d), we subsampled ‘fake bus’ features from the generated bus feature pool and checked decision values on the SVM hyper-plane trained from Tr-A set. As shown in step (e), we sorted the fake samples by the distance order, and chose a preset number of the nearest samples. Additionally, we also included small number of samples near the hyper-plane that were classified as non-bus by handicapping their decision value. We then merged the near boundary fake samples with the real samples of Tr-A set. Step (f) shows the new SVM hyper-plane trained by the merged set. Before training the new SVM, we added random vectors, which are scaled to the magnitude of the samples, to reduce the sample bias of the generation using GAN. As was done in step (c), the classification performance of new SVM was checked with the sum of the training (Tr-A) and validation set (Tr-B) accuracy. If the accuracy score of the new SVM outperforms the previous SVM score, the reference SVM hyper-plane was replaced with the new one and the iteration continues again with the fake sample subsampling in the step (d). If not, the iteration proceeds to the step (d) without replacing the reference hyper-plane. Once the SVM performance is optimized, the associated support vectors of fake bus features were used for the augmented training set. The entire process is repeated with the Tr-B as the training set for GAN and SVM, and Tr-A as the validation set. The whole processes are repeated for each acoustic scene class.

Table 4 shows the source domain DB for transfer learning. The target indoor surveillance DB consists of 15 events (a crying child, breaking glass, water drops, chirping birds, a doorbell, home appliance beeping, screaming, a dog barking, music, speech, a cat meowing, a gunshot, a

siren, an explosion, and footsteps). For checking noise robustness of AER, noise was added to the event DB at 5, 10, 15 dB SNR. In addition, compared with other DNN-based feature extraction methods, such as the Deep Belief Network (DBN) feature, which is used for music genre classification [18], and DNN bottleneck feature [17], the proposed method demonstrated improved accuracy by effectively utilizing the information transferred from the source domain. Table 5 shows performance comparison with aforementioned conventional approaches.

**Table 4.** Source domain database description

| DB set | Contents |
|---|---|
| Clear-OL [13] | Alert, cough, door slam, drawer, key, keyboard, knocking, laughing, mouse, page turn, pen drop, phone, printer, speech, switch, clear throat |
| RWCP [14] | Air-cap, bell, break stick, buzzer, castanet, ceramic collision, clap, clock ringing, coin, cymbals, drum, dryer, grinding coffee, kara, maracas, metal collision, article dropping, plastic collision, pump, punch stapler, rubbing, shaver, spray, string, tambourine, toy, whistle, wood collision |
| Urban-Sound [15] | Air-conditioner, dog bark, drilling, engine idling, car horn, jackhammer, children playing, siren, street music, shot |
| ESC-50 [16] | Airplane, breathing, brushing teeth, can opening, cat, chainsaw, chirping birds, church bells, clapping, clock alarm, clock tick, coughing, cow, crackling fire, crickets, crow, door - wood creaks, door knock, drinking – sipping, engine, fireworks, footsteps, frog, hand saw, helicopter, hen, insects (flying), pig, pouring water, rooster, sea waves, sheep, sneezing, snoring, thunderstorm, toilet flush, vacuum cleaner, washing machine, wind |
| **Total 93 classes / The similar classes from the different DB set had been merged / 16 kHz resampled, 16 bit resolution** | |

**Table 5.** Average acoustic event classification rate [%] for ETSI background noise using various features with SVM classifier

| | Living room noise | | | Office noise | | | Clean DB | **Average** |
|---|---|---|---|---|---|---|---|---|
| SNR [dB] | 5 | 10 | 15 | 5 | 10 | 15 | | |
| MFCC | 79.7 | 85.5 | 94.5 | 81.1 | 87.6 | 95.1 | 96.1 | 88.5 |
| DBN feature [18] | 86.4 | 89.9 | 93.9 | 89.9 | 93.3 | 95.7 | 96.4 | 92.2 |
| DNN-bottleneck feature [17] | 86.3 | 90.9 | 95.5 | 90.7 | 92.5 | 95.9 | 96.5 | 92.6 |
| **Proposed transfer learning approach** | **92.5** | **96.3** | **96.3** | **93.7** | **96.5** | **96.5** | **98.9** | **95.8** |

Table 6 shows performance comparison between original DB set and GAN based augmented DB set. We used IEEE Detection and Classification of Acoustic Scenes and Events

(DCASE) 2017 task 1 DB [21] for Acoustic Scene Classification (ASC). It contains 15 different acoustic scene classes such as, Bus, Café, Car, City center, Forest path, Grocery store, Home, Lakeside beach, Library, Metro station, Office, Residential area, Train, Tram, and Park. We used Discrete Fourier Transform (DFT) based feature and Mel-Filtered Bank (MFB) as feature input, and Fully Connected Neural Network (FCNN) and SVM for classifier. Based on the experimental results of AER and ASC, we achieved improved performance in noisy surveillance environment utilizing information of universal background DB and generative method. Additionally, as shown in Figure 7, we achieved the best performance in DCASE 2017 grand challenge Task using the GAN based approach.

**Table 6**. Comparing the performance of the conventional and the proposed method (average accuracy on 4-fold validation of DCASE 2017 development set)

| Avg. acc. [%] | with original development set | | | | with augmented set | | | |
|---|---|---|---|---|---|---|---|---|
| | DFT-FCNN | MFB-FCNN | DFT-SVM | MFB-SVM | DFT-FCNN | MFB-FCNN | DFT-SVM | MFB-SVM |
| | 75.4 | 75.1 | 78.2 | 79.3 | 83.2 | 83.7 | 81.6 | 85.6 |

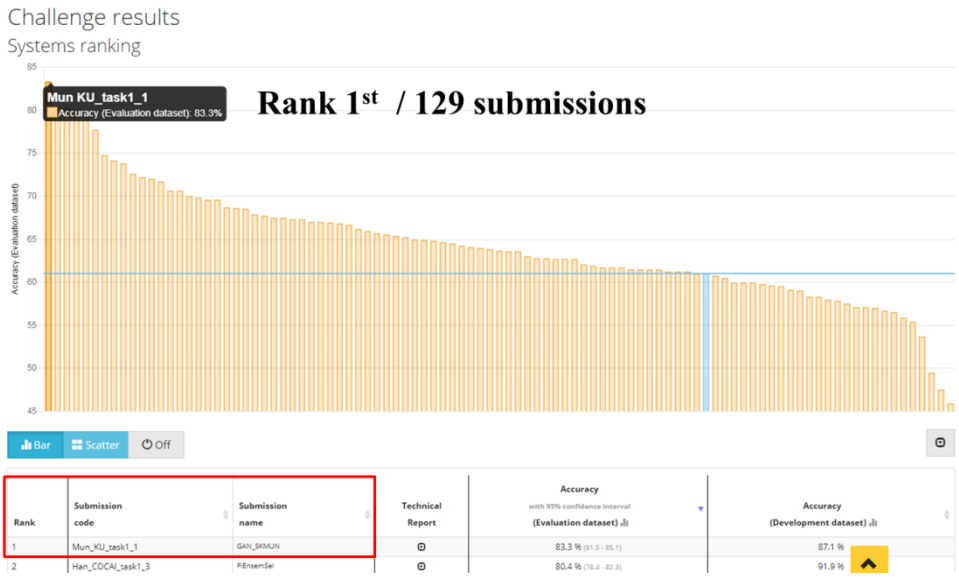


**Fig. 7** IEEE DCASE challenge 2017 task 1 results, (http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results)

## 3. List of relevant papers published by this project

- **[Task 1] Investigate novel sound source localization techniques in highly non-stationary noise dominant environments**
  - Seongkyu Mun, Suwon Shon, Wooli Kim, David K. Han, and Hanseok Ko, "Acoustic Signal based Noise Robust Speaker Direction Estimation using Recurrent Neural Network", *IEICE transactions on Information & System*, 2017 [submitted]

- **[Task 2] Explore various multichannel signal enhancement techniques**
  - Minkyu Shin, Seongkyu Mun, David K. Han and Hanseok Ko, "New Generalized Sidelobe Canceller with Denoising Auto-Encoder for Improved Speech Enhancement", *IEICE transactions on Fundamentals of Electronics Communications and Computer Sciences*, vol.100-A, no.12, Dec, 2017

- **[Task 3] Develop the acoustic event recognition system for aerial robot platform**
  - Seongkyu Mun, Sangwook Park, David K. Han and Hanseok Ko, "Generative Adversarial Network based Acoustic Scene Training Set Augmentation and Selection using SVM Hyper-Plane", *IEEE DCASE challenge workshop*, Nov. 2017 **[Winner of the Grand Challenge Task 1]**
  - Seongkyu Mun, Minkyu Shin, Suwon Shon, Wooil Kim, David K. Han and Hanseok Ko, "DNN Transfer Learning based Non-linear Feature Extraction for Acoustic Event Classification", *IEICE transactions on Information and Systems*, Vol. 100-D, No. 9, pp. 2249-2252, Sep. 2017
  - Seongkyu Mun, Suwon Shon, Wooli Kim, David K. Han, and Hanseok Ko, "A Novel Discriminative Feature Extraction for Acoustic Scene Classification using RNN based Source Separation", *IEICE transactions on Information & System*, 2017 [in press]

## 4. Conclusions and future work

We explored three key-techniques for robust acoustics and speech perception of aerial robot for scene understanding during critical emergency response missions. For noise robust sound source localization, we proposed a noise robust desired sound direction estimation method using LSTM based weighting function. The direction estimation experiments confirmed that the proposed method shows improved robustness under indoor surveillance noise environment characterized by presence of harmonic or non-stationary noise sources. Our future work will investigate effective methods for applying RNN method to phase-spectrogram as Non-negative Matrix Factorization (NMF) was applied to phase-spectrogram previously in [19].

In terms of our signal enhancement performance under noisy environment, the GSC exploits spatial information and generates multi-channel enhanced signals on which the following DAE can act. As a result, the DAE can take advantage of the multi-channels by modeling the underlying relationship of the distortion with adjacent frequency bins in other frequencies and other channels. The evaluation results demonstrate that utilizing the results of the proposed GSC structure as an input to the DAE is effective in improving noise reduction and speech recognition performance.

To improve acoustic event recognition performance and overcome the deficit of acoustic event resource, we proposed a novel DNN based transfer learning approach. By utilizing the information transferred from the universal source domain, the proposed approach was characterized by improved AEC accuracy in indoor surveillance experiments. Once DNN filter training has been completed in the source domain, this DNN filter can be utilized in other domains, repeatedly. Therefore, future work will investigate an effective transfer learning scheme for various acoustic applications and determine how performance changes depending on the configuration of the data.

The acoustic perception during the hovering and moving of the aerial robot, which will be conducted in the future steps, will be more challenging task, due to the severe noisy

environment. Furthermore, we assumed that an only single sound event occurs within a restricted event class in this phase, but in a real environment hundreds of multiple sounds occur simultaneously. To address the issue above, the next research phase is to investigate acoustic event recognition using the Google Audio-Set [20]. The Google Audio-Set is acoustic DB based on real life video uploaded in YouTube. It is the latest and largest video-based sound event recognition DB released in March 2017 with 5.8K hour long consisting of 527 sound classes in total. Based on the acoustic database, we will address the DB deficit issue of this year and investigate the simultaneous occurrence of the hundreds of sounds mentioned above. In the IEEE DCASE 2017 challenge task 4 [21], there was a competition using the Audio-Set DB. It was the competition to recognize 17 types of warning and vehicle sounds for a self-driving smart car environment similar to aerial robot environments. The participating teams using the convolutional recurrent neural network showed the best performance in the competition. As a future plan, we plan to explore the domain transformation (adaptation) for the aerial robot environment using the GAN-based various approaches and the audio feature augmentation using other generative methods such as, Variational Auto Encoder (VAE).

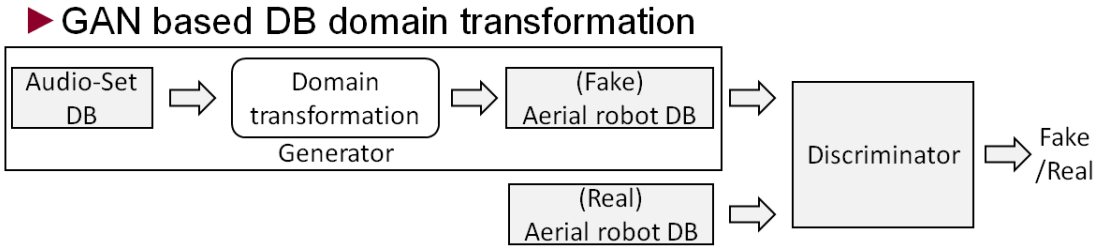


**Fig. 8** The example structure of GAN based domain transformer

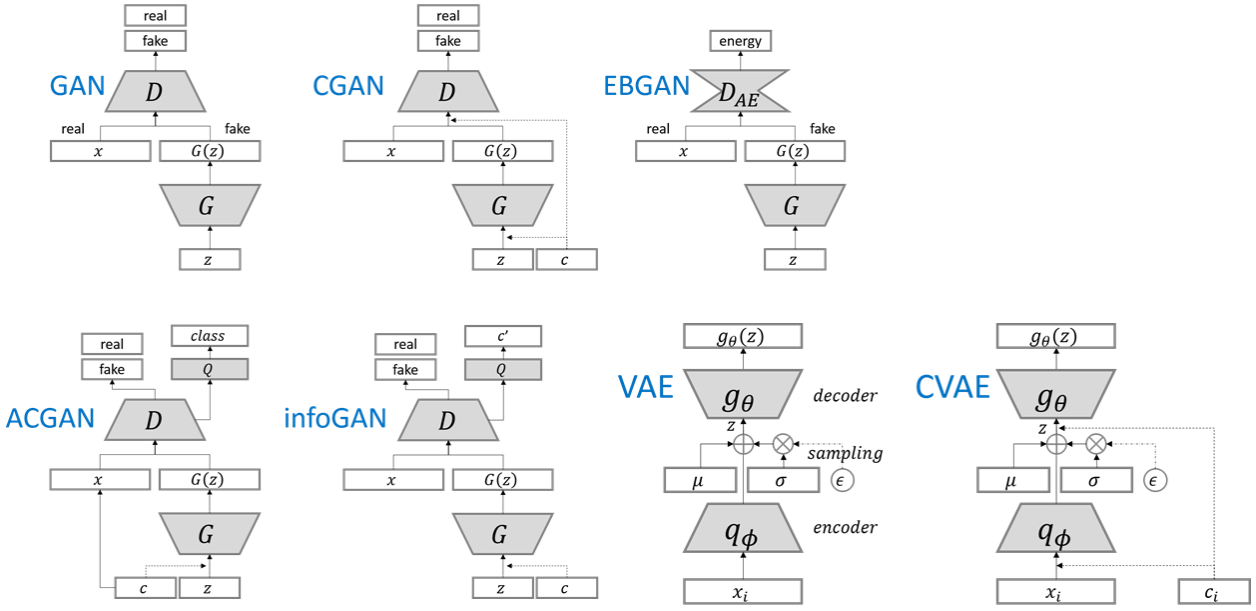


**Fig. 9** Possible structure examples of generative approaches [22]

## 5. References

[1] Y. Denda, T. Nishiura, and Y. Yamashita, “Robust talker direction estimation based on weighted csp analysis and maximum likelihood estimation” IEICE Trans. on Information and Systems, vol. E89-D, no. 3, pp. 1050–1057, Jan. 2006

[2] I. Markovic et. al., “Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering”, Robotics and Autonomous Systems, vol. 58, no. 11, pp.1185-1196, Jan. 2010

[3] O. Ichikawa, T. Fukuda, and M. Nishimura, “DOA estimation with local-peak-weighted CSP” EURASIP Journal on Advances in Signal Processing, vol. 2010, no.1, pp. 1–10, Jan. 2010.

[4] P. Swietojanski, A. Ghoshal, and S. Renals, “Hybrid acoustic models for distant and multichannel large vocabulary speech recognition,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 285–290, 2013

[5] M. I. Mandel, R. J. Weiss, and D. Ellis, “Model-Based Expectation-Maximization Source Separation and Localization,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[6] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, 2006.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech,” IEEE Trans., vol. 19, no. 7, pp. 2125–2136, 2011.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2015–Augus, pp. 5206–5210, 2015

[9] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2015–Janua, pp. 3214–3218, 2015.

[10] M. Oquab et al, “Learning and transferring mid-level image representations using convolutional neural networks,” IEEE Conf. on Computer Vision and Pattern Recog. (CVPR), Columbus, USA, pp. 1717-1724, June 2014.

[11] J. Gehring et al, “Extracting deep bottleneck features using stacked auto-encoders,” IEEE Int. Conf. on Acoustics, Speech and Signal Proc., Vancouver, Canada, pp. 3377-3381, May 2013.

[12] C. Cortes and V. Vapnik, “Support-vector networks”, Ma-chine learning, vol. 20, no.3, pp. 273-297, 1995.

[13] A. Temko et al, “CLEAR evaluation of acoustic event detection and classification systems,” Proc. of Int. Eval. Work. on Classification of Events, Act. and Relation., pp. 311–322, 2007.

[14] S. Nakamura et al,"Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition', in Proc. of EUROSPEECH, pp. 2255–2258, 1999.

[15] J. Salamon et al, "A dataset and taxonomy for urban sound research," ACM 2014 Int. Conf. on Multimedia, New York, USA, pp. 1041-1044, Oct. 2014.

[16] K. Piczak, "ESC: Dataset for environmental sound classification," ACM 2015 Int. Conf. on Multimedia, Brisbane, Australia, pp. 1015-1018, Oct. 2015.

[17] S. Mun, S. Shon, W. Kim, H. Ko, "Deep neural network bottleneck features for acoustic event recognition," Proc. of the Int. Speech Comm. Association, INTERSPEECH 2016, San Francisco, USA, pp. 2954-2957, Sep. 2016

[18] P. Hamel, D. Eck, "Learning features from music audio with deep belief networks," Int. Society for Music Infor. Retri. Conf., ISMIR 2010, Utrecht, Netherlands, pp. 339-344, Aug. 2010.

[19] S. Shon, S. Mun, D. Han, H. Ko, "Non-negative matrix factorisation-based subband decomposition for acoustic source localization", Electronics Letters, vol. 51, no. 22, pp 1723-1724, Oct. 2015

[20] J. F. Gemmeke, et. al., "Audio Set: An ontology and human-labeled dataset for audio events", ICASSP 2017, Mar. 2017.

[21] http://www.cs.tut.fi/sgn/arg/dcase2017/

[22] https://github.com/hwalsuklee/tensorflow-generative-model-collections