JOINT EGO-NOISE SUPPRESSION AND KEYWORD SPOTTING ON SWEEPING ROBOTS

Yueyue Na[1]*, Ziteng Wang*[1]*, Liang Wang*[2]*, Qiang Fu*[1]

[1] Alibaba Group, China
{yueyue.nyy, ziteng.wzt, fq153277}@alibaba-inc.com
[2] School of Electronics and Communication Engineering
Sun Yat-sen University (SYSU), Guangzhou, Guangdong, 510275, China
wangliang7@mail.sysu.edu.cn

ABSTRACT

Keyword spotting is necessary for triggering human-machine speech interaction. It is a challenging task especially in low signal-to-noise ratio and moving scenarios, such as on a sweeping robot with strong ego-noise. This paper proposes a novel approach for joint ego-noise suppression and keyword detection. The keyword detection model accepts outputs from multi-look adaptive beamformers. The noise covariance matrix in the beamformer is in turn updated using the keyword absence probability given by the model, forming an end-to-end loop-back. The keyword model also adopts a multi-channel feature fusion using self-attention, and a hidden Markov model for online decoding. The performance of the proposed approach is verified on real-word datasets recorded on a sweeping robot.

Index Terms— sweeping robot, ego-noise, beamforming, speech enhancement, keyword spotting

1. INTRODUCTION

For the application of audio interfaces on autonomous systems, such as self-driving cars, robots, drones, etc., the influence of ego-noise (self-created noise) is inevitable [1]. Small footprint keyword spotting (KWS) [2] on sweeping robot is considered in this paper, which is used as the conversation triggering module of the audio interface. Sweeping robot ego-noise is a kind of composite noise resulted from many sources, such as multiple electric motors, sweeping and/or mopping brushes, robot wheels, vacuum cleaner, etc., both directional and non-sparse (compared with speech) diffuse noise components are contained. Since the microphone array is mounted on the robot, which is much closer to the noise sources than to the speaker, audio signals with low (-10 dB) signal-to-noise ratio (SNR) must be handled. In addition, moving conditions must be considered, which will result in the time variant transfer function from the speaker to the robot.

Most on-device distant-talking KWS utilizes the sequential architecture of speech enhancement (SE) submodule followed by KWS submodule. For SE, beamforming techniques can be used. Either fixed [3] or adaptive [4, 5] beamformers can be generated according to the microphone array geometry and/or masked array data. To suppress ego-noise, the authors in [6] use independent component analysis (ICA) plus permutation alignment postprocessing to separate speech from drone ego-noise. A dictionary based approach is also proposed in [7] to suppress the ego-noise generated by a humanoid robot.

For KWS, deep neural network (DNN) [2], convolutional neural network (CNN) [8], attention mechanism [9, 10], etc., can be used to detect a group of predefined keywords from speech data. A decoder, such as hidden Markov model (HMM) [11–13], is used to smooth neural network acoustic model (AM)'s outputs and calculate the confidence of a certain keyword. The keyword is detected if its confidence exceeds the corresponding threshold.

A joint approach for sweeping robot ego-noise suppression and keyword spotting is introduced in this paper. To overcome the difficulties of low SNR and time variant transfer function, two improvements are proposed. First, unlike the common sequential architecture with independent submodules, a feedback [14] is established from KWS to SE to perform better noise suppression with supervised information. Second, an online algorithm is used to update the HMM decoder's transition matrix, which brings more adaptability in different working scenes. The rest of this paper is organized as follows, section 2 depicts the proposed approach, experiments and comparisons are given in section 3, at last, section 4 gives the conclusion of this paper.

2. THE PROPOSED APPROACH

A circular microphone array with 65 mm of diameter and $M = 6$ microphones is used. Short-time Fourier transform is used to transform 16 kHz time domain signals into time-frequency domain, with 40 ms FFT size and 20 ms data block shift. The signal model in (1) is considered, where $\boldsymbol{x}$ is M dimensional time-frequency domain microphone array data,

k is frequency bin index, τ is data block index, s is target speech, and $\boldsymbol{v}$ is M dimensional sweeping robot ego-noise. The time variant transfer function $\boldsymbol{a}$ models the moving condition of the sweeping robot.

$$\boldsymbol{x}(k,\tau) = \boldsymbol{a}(k,\tau)s(k,\tau) + \boldsymbol{v}(k,\tau) \tag{1}$$

Due to the low SNR, time variant transfer function, and non-sparseness of the sweeping robot ego-noise, target signal statistics (direction-of-arrival, mask, transfer function, etc.) is difficult to estimate. Instead, multi-look beamforming [3, 15] is used to enhance target from multiple fixed orientations, and the estimation of $\boldsymbol{a}$ in (1) is avoided. Our approach is similar as [3], however, adaptive beamforming is used, which is guided by the information from the KWS submodule.

2.1. System overview

The system overview of our approach is shown in Fig. 1. In the SE submodule, $\boldsymbol{x}$ is enhanced by N beamformers. In the KWS submodule, multi-channel features $\boldsymbol{f}_1, ..., \boldsymbol{f}_N$ are extracted from the enhanced signals $\boldsymbol{y} = [y_1, ..., y_N]^T$, where ^{T} denotes transpose, then fused to a single feature vector $\bar{\boldsymbol{f}}$ according to the attention weights. The AM generates the observation probabilities of five classes, which are sent to the HMM decoder. At last, the best transition path is determined by the Viterbi algorithm, then keyword appearance and corresponding confidence can be determined. Meanwhile, the keyword absence probability $P(Filler|\bar{\boldsymbol{f}})$ is looped back to the SE submodule to guide the noise covariance adaption.

2.2. The SE submodule

Multiple minimum variance distortionless response (MVDR) beamformers $\boldsymbol{w}_n$, $n = 1, ..., N$ are generated according to (2), and the corresponding enhanced signal is given in (3), where ^{H} denotes conjugated transpose, $\boldsymbol{d}(\varphi_n, \theta)$ is the steering vector calculated from the array geometry, φ and θ are azimuth and elevation (degree). N azimuths are uniformly sampled from the unit circle. Since the user usually looks down on the sweeping robot, an empirical θ is set. The parameter δ is used to balance noise reduction and speech distortion [16], $\boldsymbol{I}$ is the identity matrix of proper size, ϵ is a small positive number to prevent singular. Since all frequency bins has similar operation, the frequency bin index k is omitted.

$$\boldsymbol{w}_n(\tau) = \frac{\left[\boldsymbol{\Phi}(\tau) + \epsilon\boldsymbol{I}\right]^{-1}\boldsymbol{d}(\varphi_n,\theta)}{\boldsymbol{d}^H(\varphi_n,\theta)\left[\boldsymbol{\Phi}(\tau) + \epsilon\boldsymbol{I}\right]^{-1}\boldsymbol{d}(\varphi_n,\theta) + \delta} \tag{2}$$

$$y_n(\tau) = \boldsymbol{w}_n^H(\tau)\boldsymbol{x}(\tau) \tag{3}$$

Since $\boldsymbol{v}$ in (1) is unknown, the noise covariance matrix $\boldsymbol{\Phi}$ is estimated according to (4) and (5), where η is the dynamic forgetting factor [13], which is controlled by the fixed forgetting factor $\eta^{(1)}$ and the keyword absence probability $P(Filler|\bar{\boldsymbol{f}})$. It can be seen that if the keyword is present ($P(Filler|\bar{\boldsymbol{f}}) \approx 0$), noise adaptation will be slowed down to prevent keyword speech being cancelled. While if the keyword is absent, noise adaptation will continue. Since $P(Filler|\bar{\boldsymbol{f}})$ is obtained from the subsequent submodule, which uses enhanced signals as its input, higher accuracy of the supervised information is expected, and virtuous circle can be established, which is verified in section 3.

$$\boldsymbol{\Phi}(\tau) = \eta\boldsymbol{\Phi}(\tau-1) + (1-\eta)\boldsymbol{x}(\tau)\boldsymbol{x}^H(\tau) \tag{4}$$

$$\eta = \eta^{(1)} + (1-\eta^{(1)})(1 - P(Filler|\bar{\boldsymbol{f}})) \tag{5}$$

2.3. The KWS submodule

2.3.1. Acoustic model

Forty dimensional FBank features [17] $\boldsymbol{f}_1, ..., \boldsymbol{f}_N$ are extracted from N enhanced channels from (3). It is shown in [3] that soft self-attention is helpful to improve multi-channel KWS performance as well as reduce computational complexity. N feature vectors are combined to a single vector $\bar{\boldsymbol{f}}$ according to (6). Attention weights $\bar{g}_n$ are calculated according to (7) and (8), where $\mathrm{Affine}(\cdot)$ denotes an affine transform layer, $\boldsymbol{u}^T$ is one-row linear layer weight.

$$\bar{\boldsymbol{f}}(\tau) = \sum_{n=1}^{N} \bar{g}_n(\tau)\boldsymbol{f}_n(\tau) \tag{6}$$

$$g_n(\tau) = \boldsymbol{u}^T \tanh(\mathrm{Affine}(\boldsymbol{f}_n(\tau))) \tag{7}$$

$$\bar{g}_n(\tau) = \frac{\exp(g_n(\tau))}{\sum_{i=1}^{N}\exp(g_i(\tau))} \tag{8}$$

Any single-channel KWS network can be used as the AM part in Fig. 1. In this paper, feedforward sequential memory networks (FSMN) [18] is used. The AM is constructured as (9) to (11), where $\mathrm{ReLU}(\cdot)$ and $\mathrm{Softmax}(\cdot)$ denote the corresponding activation functions, $\mathrm{FSMN}(\cdot)$ denotes the operation of FSMN layer, L is the number of stacked FSMN layers, $\boldsymbol{h}$ denotes the temporary output from hidden layers.

$$\boldsymbol{h}_0(\tau) = \mathrm{ReLU}(\mathrm{Affine}(\bar{\boldsymbol{f}}(\tau))) \tag{9}$$

$$\boldsymbol{h}_l(\tau) = \mathrm{FSMN}(\boldsymbol{h}_{l-1}(\tau)), l = 1, ..., L \tag{10}$$

$$\boldsymbol{p}(\tau) = \mathrm{Softmax}(\mathrm{Affine}(\boldsymbol{h}_L(\tau))) \tag{11}$$

Chinese character based modeling is adopted. A keyword with four different characters, denoted as A, B, C, and D, is used. The AM outputs the generation probabilities of five classes, as shown in Fig. 1 and (11), where $\boldsymbol{p} = [P(\bar{\boldsymbol{f}}|Filler), P(\bar{\boldsymbol{f}}|A), P(\bar{\boldsymbol{f}}|B), P(\bar{\boldsymbol{f}}|C), P(\bar{\boldsymbol{f}}|D)]^T$. "Filler" models silence, noise, and non-keyword speech.

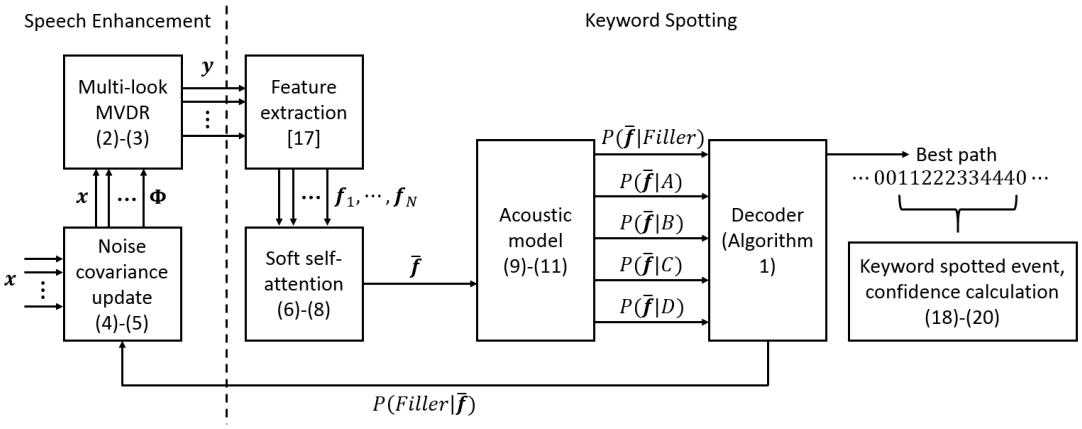

Fig. 1. System overview.

2.3.2. Decoder

Raw probabilities from the AM are noisy, it should be smoothed by a posterior handling scheme [2]. In this paper, HMM is used to smooth the data under the maximum likelihood criterion. The AM is used as the HMM's generation model, which is fixed after training. The HMM's transition matrix is updated online according to Algorithm 1 [13], where $\boldsymbol{T}$ is the 5×5 transition matrix, $\boldsymbol{\alpha} = [P(Filler|\bar{\boldsymbol{f}}), P(A|\bar{\boldsymbol{f}}), P(B|\bar{\boldsymbol{f}}), P(C|\bar{\boldsymbol{f}}), P(D|\bar{\boldsymbol{f}})]^T$ is the forward probability[1], α_1 is sent back to (5).

To calculate the keyword confidence, max presence probabilities $q_i, i = A, B, C, D$ of each keyword state are calculated from $\boldsymbol{\alpha}$ and the best path $[z(\tau_0 - Z + 1), ..., z(\tau_0)]^T$ derived by the Viterbi algorithm, where Z is the backtrack window size, τ_0 is current data block index.

$$q_i = \max_{\tau=\tau_0-Z+1,\ldots,\tau_0} P(z(\tau) = i|\bar{\boldsymbol{f}}(\tau)) \tag{18}$$

Then, the confidence q is calculated as:

$$q = \max_{i=A,B,C,D} \bar{q}_i \tag{19}$$

$$\bar{q}_i = \begin{cases} \prod_{j=A,B,C,D} q_j/q_i & q_i \neq 0 \\ 0 & otherwise \end{cases} \tag{20}$$

It is shown in experiments that (19) is robust against unclear pronunciation in distant-talking and noisy conditions.

3. EXPERIMENTS

3.1. Model training

A pipeline is established to perform data simulation, speech enhancement, and feature extraction for model training. To make sure the SE submodule has enough data to converge, minibatches with 1 minute length are generated. Multiple positive and negative utterances are randomly distributed in the minibatch. The probability used in (5) is provided by one minus the clean speech's voice activity detection (VAD) result. Multi-channel room impulse responses (RIR) are generated by the image method [19, 20] to simulate single channel data source to multi-channel microphone array data. Sweeping robot ego-noise is simulated as diffuse noise according to [21] with data sources from the noise training set.

Detailed acoustic and training set descriptions are listed in Table 1 and Table 2. Cross entropy is used as the loss function, which is optimized by the ADAM optimizer, the learning rate is 0.001. Three SNR ranges with increasing difficulty are used during the 0% to 10%, 10% to 70%, and 70% to 100% of the training procedure. The training is terminated after 50k hours of simulated data is consumed. To compare the performance of the proposed approach, the same KWS model with 120k parameter size is used in all experiments.

3.2. Real-world experiments

Experiments are performed on real-world datasets. The positive test set contains 6 hours of long audio recorded from different scenes with different robot working modes and noise levels, totally 4k keyword utterances are contained. The negative test set contains 24 hours of long audio, recorded when the sweeping robot is working, meanwhile a TV is playing news program. Parameters are tuned and then fixed in all experiments, which are given in Table 3. Experimental results are shown as the form of receiver operating characteristic (ROC) curves in Fig. 2, lower area under the curve (AUC) means better performance. The proposed approach is denoted as MVDR + KWS(A), which is implemented with multi-look MVDR beamforming with feedback, plus the adaptive HMM transition matrix update.

The baseline approach (DB + KWS(A)) is implemented

[1] Since the backward probability $\boldsymbol{\beta} = \mathbf{1}$ in online HMM, the (normalized) state presence probability $P(i|\bar{\boldsymbol{f}}) = \alpha_i\beta_i/P(\bar{\boldsymbol{f}}) = \alpha_i$, [13].

Algorithm 1: Online HMM transition matrix update.

Initialize: $\boldsymbol{\alpha}(0), \boldsymbol{T}(0)$ empirically
Input: $\boldsymbol{\alpha}(\tau-1), \boldsymbol{T}(\tau-1), \boldsymbol{p}(\tau)$
Output: $\boldsymbol{\alpha}(\tau), \boldsymbol{T}(\tau)$
1. Update each γ_{ij} in $\boldsymbol{\Gamma}$ and each α_j in $\boldsymbol{\alpha}$:

$$\gamma_{ij}(\tau) = \alpha_i(\tau-1)t_{ij}(\tau-1)p_j(\tau) \tag{12}$$

$$\alpha_j(\tau) = \sum_{i=1}^{5} \gamma_{ij}(\tau) \tag{13}$$

Then normalize:

$$\boldsymbol{\Gamma}(\tau) = \frac{\boldsymbol{\Gamma}(\tau)}{\sum_{i=1}^{5}\sum_{j=1}^{5}\gamma_{ij}(\tau)} \tag{14}$$

$$\boldsymbol{\alpha}(\tau) = \frac{\boldsymbol{\alpha}(\tau)}{\sum_{i=1}^{5}\alpha_i(\tau)} \tag{15}$$

2. Update each t_{ij} in $\boldsymbol{T}$, where $\eta^{(2)}$ is the fixed forgetting factor for HMM.

$$\eta_i = \eta^{(2)} + (1-\alpha_i(\tau))(1-\eta^{(2)}) \tag{16}$$

$$t_{ij}(\tau) = \eta_i t_{ij}(\tau-1) + (1-\eta_i)\gamma_{ij}(\tau)/\alpha_i(\tau) \tag{17}$$

Then, normalize $\boldsymbol{T}$ to row stochastic form.

Table 1. Acoustic configuration for data simulation.

Parameter	Range
Room size (x, y, z) (m)	([3, 10], [3, 8], [2.5, 6])
Source/receiver distance (m)	[0.5, 8]
RT60 (s)	[0.2, 1.5]
Volume (average RMS) (dB)	[-45, -15]
SNR (dB)	[5, 20], [-5, 15], [-10, 5]

with differential multi-look beamformers[2] like in [3], and the proposed KWS submodule. The performance without SE (KWS(A) ONLY) is also shown for comparison.

To show the performance improvement of the online HMM transition scheme (KWS(A)), HMM decoder with fixed transition matrix is compared (MVDR + KWS(F)). The fixed transition probabilities can easily be derived from the ratio of keyword and non-keyword data in the positive test set. Performance improvement can be observed from MVDR + KWS(F) to MVDR + KWS(A). It can be explained that the fixed transition matrix is optimized with respect to the average condition in all scenes of the test set, however, adaptive transition matrix may give extra tuning in individual scenes, which will improve the overall performance.

At last, the supervised information from the feedforward

[2]The code for differential beamforming simulation is available: https://github.com/nay0648/ego2022

Table 2. Training set description.

Name	Description	Size (hours)
Positive	Keyword utterances.	352
Negative	Non-keyword utterances.	969
Noise1	TV, non-speech noise.	386
Noise2	Sweeping robot ego-noise.	78

Table 3. Parameters used in all experiments.

$N = 3$	$\theta = 45$	$\epsilon = 1e-6$	$\delta = 3$
$\eta^{(1)} = 0.999$	$L = 5$	$\eta^{(2)} = 0.9$	$Z = 120(2.4s)$

scheme (KWS(A) + MVDR + KWS(A)) is compared, which is implemented like the structure in [14]. The first KWS(A) is only used to provide $P(Filler|\bar{\boldsymbol{f}})$. Most mask and neural beamforming approaches also utilize the similar architecture, with neural networks used before SE. Since noisy signals must be handled, the supervised information is expected less accurate than the information from enhanced signals. This conjecture is verified by the performance improvement from KWS(A) + MVDR + KWS(A) to MVDR + KWS(A).

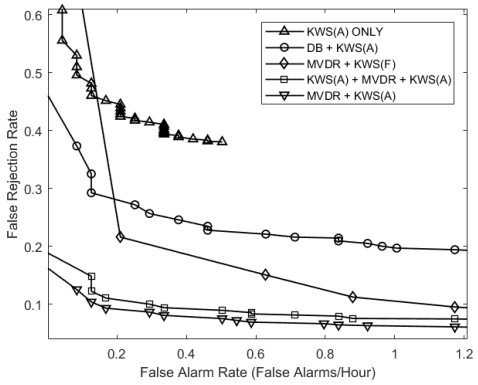

Fig. 2. ROC curves on real-world datasets.

4. CONCLUSION

The ego-noise with complicated compositions, low SNR, and time variant target transfer function prevents the directly usage of audio interfaces on sweeping robots. To overcome these difficulties, an approach for joint sweeping robot ego-noise suppression and keyword spotting is proposed. First, multi-look MVDR beamformers are used to perform speech enhancement. Then, multi-channel features are fused by an attention network, then sent to the AM part of KWS. The AM's generation probabilities are postprocessed by an online updated HMM decoder. At last, the keyword absence probability is sent back to guide the noise covariance adaptation for better ego-noise suppression. The performance of the proposed approach is verified on real-world datasets.

5. REFERENCES

[1] Alexander Schmidt, Heinrich W Löllmann, and Walter Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 1127–1149, 2020.

[2] Guoguo Chen, Carolina Parada, and Georg Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.

[3] Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, and Dong Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.

[4] Yuki Kubo, Tomohiro Nakatani, Marc Delcroix, Keisuke Kinoshita, and Shoko Araki, "Mask-based mvdr beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6855–6859.

[5] Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, Chao Weng, Jianming Liu, and Dong Yu, "Neural spatio-temporal beamformer for target speech separation," *arXiv preprint arXiv:2005.03889*, 2020.

[6] Lin Wang and Andrea Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[7] Alexander Schmidt, Heinrich W Löllmann, and Walter Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6583–6587.

[8] Tara N Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie, "Attention-based end-to-end models for small-footprint keyword spotting," *arXiv preprint arXiv:1803.10916*, 2018.

[10] Axel Berg, Mark O'Connor, and Miguel Tairum Cruz, "Keyword transformer: A self-attention model for keyword spotting," *arXiv preprint arXiv:2104.00769*, 2021.

[11] Jan Robin Rohlicek, William Russell, Salim Roukos, and Herbert Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *International Conference on Acoustics, Speech, and Signal Processing,*. IEEE, 1989, pp. 627–630.

[12] Mark Stamp, "A revealing introduction to hidden markov models," *Department of Computer Science San Jose State University*, pp. 26–56, 2004.

[13] Dongwen Ying and Yonghong Yan, "Noise estimation using a constrained sequential hidden markov model in the log-spectral domain," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 6, pp. 1145–1157, 2013.

[14] Yiteng Huang, Thad Hughes, Turaj Z Shabestary, and Taylor Applebaum, "Supervised noise reduction for multichannel keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5474–5478.

[15] Meng Yu, Xuan Ji, Bo Wu, Dan Su, and Dong Yu, "End-to-end multi-look keyword spotting," *arXiv preprint arXiv:2005.10386*, 2020.

[16] Mehrez Souden, Jacob Benesty, and Sofiene Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[17] Haytham M. Fayek, "Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between," 2016.

[18] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[19] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[21] Emanuël AP Habets and Sharon Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.