INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Artificial Neural Network-Based Feature Combination for Spatial Voice Activity Detection

Stefan Meier and Walter Kellermann

Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

{stefan.a.meier, walter.kellermann}@fau.de

Abstract

For many applications in speech communications and speech-based human-machine interaction, a reliable Voice Activity Detection (VAD) is crucial. Conventional methods for VAD typically differentiate between a target speaker and background noise by exploiting characteristic properties of speech signals. If a target speaker should be distinguished from other speech sources, these conventional concepts are no longer applicable, and other methods, typically exploiting the spatial diversity of the individual sources, are required. Often, it is beneficial to combine several features in order to improve the overall decision. Optimum combinations of features, however, depend strongly on the scenario, especially on the position of the target source, the characteristics of noise and interference and the Signal-to-Interference Ratio (SIR). Moreover, choosing detection thresholds which are robust to changing scenarios is often a difficult problem. In this paper, these issues are addressed by introducing Artificial Neural Networks (ANNs) for spatial voice activity detection, which allow to combine several features with background information. The experimental results show that already small ANNs can significantly and robustly improve the detection rates, offering a valuable tool for VAD.

Index Terms: voice activity detection, artificial neural network, generalized cross-correlation, receiver operating characteristic

1. Introduction

VAD, which aims at detecting activity of a target source in noisy background, is a common problem in audio signal processing and has been widely investigated over the last decades. The goal of VAD methods is to determine either if a target source (typically assumed to be a speech source) is active or if the target source is dominant. The former case may, e.g., be important for Automatic Speech Recognition (ASR), where the speech recognizer should only be active during target source activity. On the other hand, detecting target source dominance is of interest, e.g., if the adaptation step size of adaptive filters should be controlled, for instance, when the relative transfer functions of the target source should be estimated [1]. Most conventional methods address the scenario where a speech source is distinguished from stationary background noise, and distinctive properties of speech signals like stationarity, harmonic structure and spectral envelopes are exploited to distinguish between speech and noise [2, 3]. When it comes to distinguishing a target speaker from other interfering speakers, however, the above-mentioned features can no longer be exploited. In this case, spatial information can be incorporated by taking microphone arrays instead of using a single microphone. Conventional acoustic source localization techniques for multi-microphone arrays can be modified to provide information on target source activity. For instance, the Steered Response Power (SRP) method can be exploited to either distinguish between multiple point sources [4] or between point sources and incoherent background noise [5]. Similarly, the cross-correlation function between two microphones can be calculated, allowing for a detection of target activity when a peak is observed for the time lag corresponding to a target source position [6, 7, 8]. Also the Magnitude Squared Coherence (MSC) gives an indication on the activity of a dominant coherent point source in the presence of incoherent background noise [9]. Alternatively, beamforming techniques can be exploited for Spatial Voice Activity Detection (SVAD): On the one hand, one can estimate the Signal-to-Noise Ratio (SNR) by steering a beamformer to the known target source direction in order to obtain a target signal power estimate and, equivalently, a nullsteering beamformer for a noise (and interference) estimate [10, 11, 12]. On the other hand, one can monitor the look direction of an adaptive nullsteering beamformer and detect target source activity if the null is pointing towards the known Direction of Arrival (DoA) of the target source [13]. Finally, several probabilistic methods have been proposed in the literature in the past years [14, 15, 16, 17, 18].

Some of the above-mentioned methods combine several features in order to derive a decision on source activity. A reliable combination of features as well as the choice of thresholds, however, depend on various factors like, e.g., the target source position, the number of interferers and the SNR. In this paper, the problem of finding an appropriate decision in different scenarios is addressed by applying an ANN for feature combination. While using ANNs has already been considered in classical single-channel VAD applications [3, 19, 20, 21, 22], this paper focuses on combining several features with information on the target source position. The method is evaluated in the context of robot audition, where a considerable angle-dependency exists due to scattering at the robot's head.

The remainder of this paper is organized as follows: First, three different features for SVAD are introduced in Section 2. Hereupon, the concept of combining SVAD features with additional information based on ANNs is proposed in Section 3. Finally, an evaluation is performed in Section 4, leading to conclusions and an outlook in Section 5.

2. Features for SVAD

In this paper, three different kinds of features for SVAD are considered. First of all, an SNR estimate can be obtained by steer-

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

http://dx.doi.org/10.21437/Interspeech.2016-1184

ing a beamformer and a nullformer into the target source direction in order to estimate target source power and interference-plus-noise power, respectively, as further described in Section 2.1. A second feature for SVAD is the cross-correlation of the microphone signals, which will be addressed in Section 2.2. Finally, the MSC will be evaluated as feature.

2.1. SNR estimates

A first feature is the SNR, which can be estimated by steering a beamformer into the target source direction in order to obtain for each block l a target signal power estimate $\hat{\sigma}_\text{s}^2(l)$, and a nullsteering beamformer for an interference-plus-noise estimate $\hat{\sigma}_\text{n}^2(l)$. The resulting SNR estimate could be exploited similarly to [10, 11, 12] as standalone feature by detecting time frames where a predefined threshold γ_SNR is exceeded:

$$D_\text{SNR}(l) = \begin{cases} 1, & \text{if } \widehat{\text{SNR}}(l) = \frac{\hat{\sigma}_\text{s}^2(l)}{\hat{\sigma}_\text{n}^2(l)} > \gamma_\text{SNR}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In order to estimate the target signal power, a beamformer is steered towards the target source. We used an Robust Least-Squares Frequency-Invariant (RLSFI) Filter-and-Sum Beamformer (FSB) as originally proposed in [23], where a desired beamformer response $B_\text{des}(\omega, \phi, \theta)$ is approximated at a discrete set of P frequencies ω_p, $p \in \{0, \ldots, P-1\}$ and for M look directions (ϕ_m, θ_m), $m \in \{0, \ldots, M-1\}$ in the Least-Squares (LS) sense by minimizing [23]

$$\underset{\mathbf{w}_\text{f}(\omega_p)}{\text{argmin}} \, \|\mathbf{G}(\omega_p)\mathbf{w}_\text{f}(\omega_p) - \mathbf{b}_\text{des}\|_2^2, \quad (2)$$

where $\mathbf{b}_\text{des} = [B_\text{des}(\phi_0, \theta_0), \ldots, B_\text{des}(\phi_{M-1}, \theta_{M-1})]^T$ is a vector containing the frequency-invariant desired responses for all look directions, matrix $[\mathbf{G}(\omega_p)]_{mn} = \exp\left(-j\mathbf{k}_m^T\mathbf{p}_n\right)$ with wave vector $\mathbf{k}_m^T$ and N microphone positions $\mathbf{p}_n$, and $\mathbf{w}_\text{f}(\omega_p) = [W_0(\omega_p), \ldots, W_{N-1}(\omega_p)]^T$ contains the filter weights. Additionally, the robust design in [23] imposes a distortionless response constraint for the DoA of the target source and a constraint restricting the White Noise Gain (WNG), respectively,

$$\frac{|\mathbf{w}_\text{f}^T(\omega_p)\mathbf{d}(\omega_p)|^2}{\mathbf{w}_\text{f}^H(\omega_p)\mathbf{w}_\text{f}(\omega_p)} \geq \gamma > 0, \quad \mathbf{w}_\text{f}^T(\omega_p)\mathbf{d}(\omega_p) = 1, \quad (3)$$

where $\mathbf{d}(\omega_p) = [\exp(-j\mathbf{k}_\text{d}^T\mathbf{p}_0), \ldots, \exp(-j\mathbf{k}_\text{d}^T\mathbf{p}_{N-1})]^T$ is the steering vector corresponding to the DoA of the target source $(\phi_\text{d}, \theta_\text{d})$ with the wave vector $\mathbf{k}_\text{d}$ and operator $(\cdot)^H$ denoting transposition of conjugate complex vectors or matrices. In [24], this concept was modified for scenarios where free-field propagation cannot be assumed by replacing the steering vector $\mathbf{d}(\omega_p)$ with impulse responses, taking into account shadowing effects of the robot's head. In order to obtain a noise and interference power estimate, a nullsteering beamformer is used. For this task, a conventional LS beamformer has been chosen [25].

2.2. Cross-correlation

As a second feature, we consider the Generalized Cross-Correlation (GCC) with phase transform (PHAT) weighting between two microphones i and j, which can be estimated as

$$\hat{r}_{x_i x_j}(\Delta k, l) = \text{DFT}^{-1}\left(\frac{X_i^*(\mu, l)X_j(\mu, l)}{|X_i^*(\mu, l)X_j(\mu, l)|}\right), \quad (4)$$

where $X_i(\mu, l)$ and $X_j(\mu, l)$ are the Short-Time Fourier Transforms (STFTs) of the microphone signals $x_i(k)$ and $x_j(k)$, and l and μ are the block index and frequency index, respectively, and Δk is the time lag. The cross-correlation itself could be applied for SVAD as proposed, e.g., in [6, 7, 8] by determining the Time Difference of Arrival (TDoA) expected for the target source (denoted as ΔT), and detecting target source activity if the time lag of the maximum of $r_{x_i x_j}(\Delta k, l)$ (denoted as Δk_max) equals the target source TDoA, i.e.,

$$D_\text{GCC}(l) = \begin{cases} 1, & \text{if } \Delta k_\text{max}(l) f_s = \Delta T, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

A drawback of this conventional method is the fact that the actual values of $r_{x_i x_j}(\Delta k, l)$ are discarded, although they might carry valuable additional information. Moreover, head scattering in the robot audition context smears the GCC function and makes an identification of peaks difficult.

2.3. Coherence

As a third and final feature, we consider the MSC estimate $\Gamma_{ij}(l)$ between two microphones i and j, which can be calculated as follows

$$\Gamma_{ij}(\mu, l) = \frac{\left|P_{x_i x_j}(\mu, l)\right|^2}{P_{x_i x_i}(\mu, l)P_{x_j x_j}(\mu, l)}, \quad (6)$$

where $P_{x_i x_j}(\mu, l)$, $P_{x_i x_i}(\mu, l)$ and $P_{x_j x_j}(\mu, l)$ are the respective Cross-Spectral Density (CSD) and Power Spectral Densities (PSDs). While the GCC exploits the estimated phase information, the MSC provides information on the power of the crosspower spectrum. As shown in [9], the MSC can to some extent act as SVAD feature since a high MSC indicates activity of a dominant (coherent) point source, and a decision $D_\text{MSC}(l)$ can be made by averaging the MSC over all frequency bands M and setting a threshold γ_MSC:

$$D_\text{MSC}(l) = \begin{cases} 1, & \text{if } \sum_{\mu=0}^{M-1} \Gamma_{ij}(\mu, l) > \gamma_\text{MSC}, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Naturally, the MSC is not suited to distinguish between the target source and single interferers but needs to be chosen in combination with other features.

3. ANN-based SVAD

In the previous sections, three different features for SVAD were presented, which have a number of drawbacks in common. For the SNR and MSC features, thresholds γ_SNR and γ_MSC need to be chosen, which are typically highly dependent on the position ϕ_tar of the target source, while the GCC method as defined in (5) even does not directly provide a threshold which can be adjusted. On the other hand, a combination of several features would traditionally be performed by a simple AND operation, which does not take into account the relative importance of the individual features. In order to cope with these problems, an ANN can be applied to the SVAD features. To this end, we define several feature vectors containing the information of the individual methods for a given time frame l:

$$\mathbf{f}_\text{SNR}(l) = \left[\hat{\sigma}_\text{s}^2(l), \hat{\sigma}_\text{n}^2(l)\right]^\mathsf{T}, \quad (8)$$

$$\mathbf{f}_{\text{GCC},i,j}(l) = \left[r_{x_i x_j}(-L, l), \ldots, r_{x_i x_j}(L, l)\right]^\mathsf{T}, \quad (9)$$

$$\mathbf{f}_{\text{MSC},i,j}(l) = \left[\sum_{\mu=0}^{M-1} \Gamma_{ij}(\mu, l)\right], \quad (10)$$

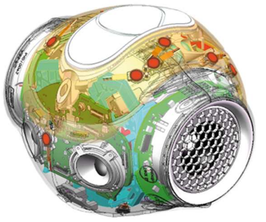

Figure 1: Head of the NAO robot with visible microphones highlighted in red (©Aldebaran Robotics).

where $\mathbf{f}_{\text{SNR}}(l) \in \mathbb{R}^{2\times1}$, $\mathbf{f}_{\text{GCC}}(l) \in \mathbb{R}^{2L+1\times1}$ and $\mathbf{f}_{\text{MSC}}(l) \in \mathbb{R}^{1\times1}$. Feature vector $\mathbf{f}_{\text{SNR}}(l)$ contains the individual components $\hat{\sigma}^2_{\text{s}}(l)$ and $\hat{\sigma}^2_{\text{n}}(l)$ instead of their ratio since the absolute powers additionally allow for a judgement on the overall power of all active signals, which would be lost if the ratio was taken instead. Moreover, a feature vector containing cosine and sine of the target source position is defined:

$$\mathbf{f}_{\phi_{\text{tar}}}(l) = [\cos(\phi_{\text{tar}}),\ \sin(\phi_{\text{tar}})]^{\mathsf{T}} \quad \in \mathbb{R}^{2\times1}. \tag{11}$$

The target DoA feature vector in (11) has been included as reference information for various reasons: First, the position of the main peaks of the GCC function carries the main information of the feature vector $\mathbf{f}_{\text{GCC},i,j}(l)$, which is only informative once a link to the target source position ϕ_{tar} can be established. Moreover, the performance of each of the individual features will differ depending on the target source position, which also calls for including information on the target source position into the feature vector. Second, the cosine and sine of ϕ_{tar} were chosen instead of ϕ_{tar} itself since cosine and sine explicitly represent the periodicity of the azimuth.

In our concept, the feature vectors of (8), (9), (10) and (11) are combined to a stacked feature vector $\mathbf{f}(l)$. For the features $\mathbf{f}_{\text{GCC},i,j}(l)$ and $\mathbf{f}_{\text{MSC},i,j}(l)$, several microphone pairs (i, j) can be exploited simultaneously. In order to define a ground truth for training, the block-wise SNR is calculated based on the knowledge of target source and interference components in the training set. With this feature vector, a feedforward classification ANN is trained with threshold 0dB. In order to allow for a soft treatment of values around 0dB, the range between −5dB and 5dB is not assigned to a single class but mapped linearly to the range between 0 and 1. A benefit of the ANN-based method, however, is the fact that this ground truth can be easily modified in the training phase. ANNs with up to three hidden layers and are used and as nonlinearity, a sigmoid function is chosen. In what follows, the notation *10-10* describes an ANN with two layers consisting of 10 nodes each.

4. Evaluation

4.1. Evaluation scenarios

The ANN-based SVAD method was evaluated in the robot audition context with the robot NAO (Aldebaran Robotics), equipped with a new 12-microphone array as illustrated in Figure 1, designed in the European Union-funded project *Embodied Audition for RobotS (EARS)* (http://robot-ears.eu) [26]. Simulated impulse responses obtained by means of a Boundary Element Method (BEM) simulation [27] were used and convolved with speech signals. The simulations were performed at a sampling rate $f_{\text{s}} = 10\text{kHz}$ for a source-microphone distance of 1m. Instantaneous and mutually independent decisions were made for blocks of 512 samples with an overlap of 50%. A training set and a test set consisting of scenarios of duration 5s with one target source and one or two interferers were created. In Table 1, the source positions are summarized, among which all possible combinations were formed. In order to avoid that training set and test set contain the same positions, an offset of 10° was added to the positions of the test set. All sources were located in the front half plane, where 90° corresponds to the position in front of the robot. Short pauses of length 1s were added to all individual source signals at random positions in order to create representative test and training sets containing both target source-only and interference-only time frames.

Table 1: *Evaluation scenarios.*

Training set	
Number of interferers	$\{1, 2\}$
Target source position	$\phi_{\text{tar}} \in \{0°, 30°, \ldots, 180°\}$
Interferer position(s)	$\phi_{\text{int},1} \in \{0°, 30°, \ldots, 180°\}$
	$\phi_{\text{int},2} \in \{0°, 30°, \ldots, 180°\}$
	$\phi_{\text{tar}} \neq \phi_{\text{int},1} \neq \phi_{\text{int},2}$
Test set	
Number of interferers	$\{1, 2\}$
Target source position	$\phi_{\text{tar}} \in \{10°, 40°, \ldots, 160°\}$
Interferer position(s)	$\phi_{\text{int},1} \in \{10°, 40°, \ldots, 160°\}$
	$\phi_{\text{int},2} \in \{10°, 40°, \ldots, 160°\}$
	$\phi_{\text{tar}} \neq \phi_{\text{int},1} \neq \phi_{\text{int},2}$

4.2. Evaluation results

For evaluation, the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) was calculated. The ROC curve is an illustration of the true positive rate plotted over the false positive rate with a varying detection threshold, as shown exemplarily in Figure 2 (more details in Section 4.2.1). The AUC measure denotes the area under the ROC curve. Ideally, with 100% true positive rate at 0% false positive rate, the AUC would be 1. The evaluation was performed in two steps. At first, the GCC feature was evaluated in order to find the optimum number of microphone pairs. In a second step, combinations with the other two features are analyzed.

4.2.1. Correlation features

In Figure 3, the performance of the correlation features dependent on the number of microphone pairs and the feedforward ANN topologies is plotted. For both training and test, a T_{60} time of 300ms was used. For the feature vector $\mathbf{f}_{\text{GCC}}$, a maximum lag of $L = 2$ was found to be sufficient, leading to a vector length of 5 (per microphone pair) plus the two elements of the position vector $\mathbf{f}_{\phi_{\text{tar}}}$. Among the possible combinations of microphone pairs, the best ones (in terms of AUC) were chosen, respectively. One can see that increasing the number of microphone pairs from one to three can lead to significant improvements while combining more microphone pairs is no longer beneficial. Even the small ANN consisting of 10 nodes performs well, with a performance decrease for 7 microphone pairs (i.e., feature vector length 37), where the small topology cannot sufficiently model the relations between features and decisions. Deeper networks, however, can still slightly improve the performance. Exemplarily, the ROC curve for 3 microphones with network topology *20-20-20* is plotted in Figure 2. Applying the decision in (5)

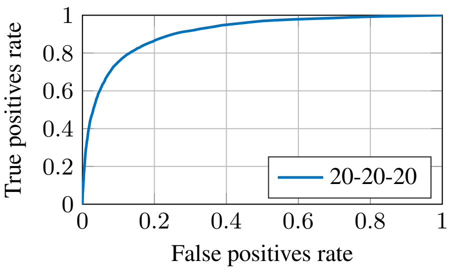

Figure 2: ROC curve for GCC features with 3 microphone pairs.

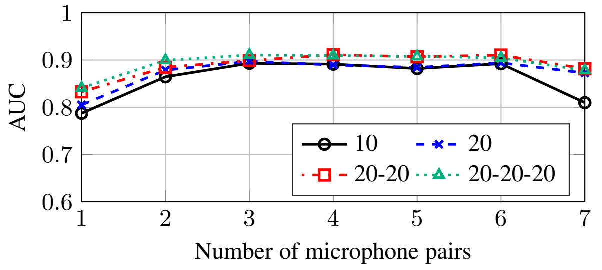

Figure 3: ANN with GCC features and a varying number of microphone pairs ($T_{60}^{(\text{train})} = 300\text{ms}$, $T_{60}^{(\text{test})} = 300\text{ms}$).

based on a free-field assumption without an ANN, for comparison, only yielded an AUC of 0.54, which was due to a difficult identifiability of the peaks in the GCC function caused by head scattering and the low sampling rate. In Figure 4, the ANN trained for a T_{60} of 300ms is applied to test data with a T_{60} of 800ms . Generally, a decrease by approx. 0.1 can be observed, but still an AUC between 0.7 and 0.8 is achieved.

4.2.2. Combined features

In the previous section, the best performance with GCC features was achieved with 3 microphone pairs, leading to a feature vector length of 17 (including $\mathbf{f}_{\phi_\text{tar}}$). In this section, it is evaluated if shorter feature vectors can achieve a similar performance. To this end, the GCC feature vector is reduced to only one microphone pair (i.e., length 5) and all possible combinations of the SNR, GCC and MSC features presented in Section 2 (always in combination with $\mathbf{f}_{\phi_\text{tar}}$) are evaluated. In Figure 5, the resulting AUC values are are summarized. Even with only one microphone pair, the GCC features clearly outperform the other features (however, at the expense of a longer feature vector). The MSC feature obviously is not well suited to distinguish between target source and interferers and achieves an AUC value of 0.57 (the result of (7) without the ANN would be 0.49). The SNR feature suffers from the close source positions and broad beams and notches, which makes it hard to distinguish between target source and interferers in many scenarios, and achieves an AUC of only 0.61 (also with (1) and no ANN, 0.62 is achieved). Combining SNR and MSC features can lead to a slight improvement to 0.63. Some combinations of several features achieve worse results than the individual features, which could be addressed by including more training data. In Figure 6, the target source is placed at $\phi_\text{tar} = 90^\circ$ (the regular case in the robot audition context) instead of variable target source positions, while the interferer positions remain the same. Now, the MSC feature becomes more valuable and the combination of MSC and SNR features leads to an AUC of 0.76, while the GCC feature achieves 0.85. With all combinations, however, the performance of the (single-pair) GCC feature cannot be significantly improved by adding other features and the three-pair GCC feature vector in Figure 3 still achieves the best AUC.

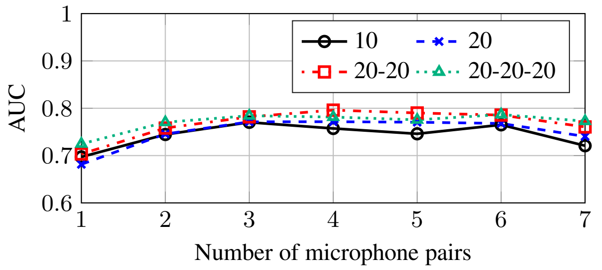

Figure 4: ANN with GCC features and a varying number of microphone pairs ($T_{60}^{(\text{train})} = 300\text{ms}$, $T_{60}^{(\text{test})} = 800\text{ms}$).

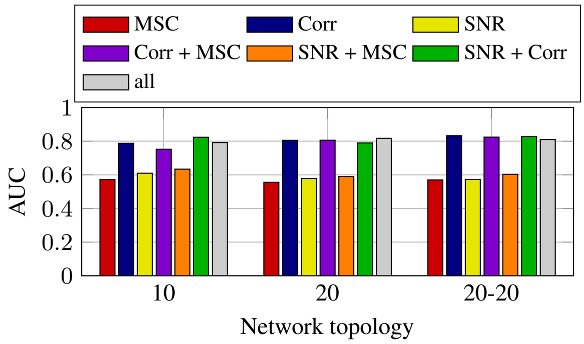

Figure 5: Combination of SNR, GCC and MSC features ($\phi_\text{tar} \in \{0^\circ, 30^\circ, \ldots, 180^\circ\}$, $T_{60}^{(\text{train})} = T_{60}^{(\text{test})} = 300\text{ms}$).

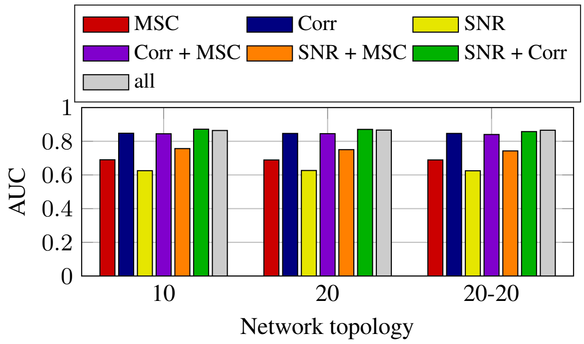

Figure 6: Combination of SNR, GCC and MSC features ($\phi_\text{tar} = 90^\circ$, $T_{60}^{(\text{train})} = T_{60}^{(\text{test})} = 300\text{ms}$).

5. Conclusions

In this paper, an ANN-based method for combination of SVAD features was proposed. The method proved to be very powerful with GCC features and was still convincing when training and test were performed for different reverberation times. Also a combination of other features with smaller feature vector length was evaluated but did not yield a competitive performance. Future work involves an extension with single-channel VAD features. Moreover, recurrent neural networks will offer the possibility to incorporate memory into the ANN.

6. References

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[2] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Improved performance measures for voice activity detection," in *Proc. ITG Conf. Speech Communication*. VDE, 2014, pp. 1–4.

[3] ——, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[4] H. Lee and D. Yook, "Space-time voice activity detection," *IEEE Trans. Consumer Electronics*, vol. 55, no. 3, pp. 1471–1476, 2009.

[5] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2011, pp. 92–97.

[6] A. Koul and J. E. Greenberg, "Using intermicrophone correlation to detect speech in spatially separated noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–14, 2006.

[7] Y. Denda, T. Nishiura, and Y. Yamashita, "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," *IEICE Trans. Information and Systems*, vol. 89, no. 3, pp. 1050–1057, 2006.

[8] Y. Denda, T. Tanaka, M. Nakayama, T. Nishiura, and Y. Yamashita, "Noise-robust hands-free voice activity detection with adaptive zero crossing detection using talker direction estimation," in *Proc. Annual Conf. Int. Speech Communication Association (Interspeech)*, 2007, pp. 222–225.

[9] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech communication*, vol. 16, no. 3, pp. 245–254, 1995.

[10] M. W. Hoffman, L. Zhao, and D. Khataniar, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 175–179, 2001.

[11] W. Herbordt, H. Buchner, and W. Kellermann, "An acoustic human-machine front-end for multimedia applications," *EURASIP Journal on Applied Signal Processing*, pp. 21–31, 2003.

[12] T. Yu and J. H. Hansen, "An efficient microphone array based voice activity detector for driver's speech in noise and music rich in-vehicle environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*. IEEE, 2010, pp. 2834–2837.

[13] S. Srinivasan and K. Janse, "Spatial audio activity detection for hearing aids," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*. IEEE, 2008, pp. 4021–4024.

[14] I. Potamitis and E. Fishler, "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2406–2415, 2004.

[15] G. Kim and N. I. Cho, "Voice activity detection using phase vector in microphone array," *Electronics Letters*, vol. 43, no. 14, pp. 783–784, 2007.

[16] H.-D. Kim, J. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. IEEE, 2008, pp. 1705–1711.

[17] D. P. Jarrett, M. Taseska, E. A. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband DOA estimates," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 5, pp. 967–978, 2014.

[18] M. Taseska and E. A. Habets, "Minimum Bayes risk signal detection for speech enhancement based on a narrowband DOA model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*. IEEE, 2015, pp. 539–543.

[19] Y. Qi and B. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, Apr 1993.

[20] F. Albu and A. Mateescu, "Application of multilayer feedforward network to the voiced-unvoiced-silence detection problem," in *Proc. Int. Symp. Communications*, Bucharest, Romania, Nov 1996, pp. 532–537.

[21] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, and C.-H. Lee, "A universal VAD based on jointly trained deep neural networks," in *Proc. Annual Conf. Int. Speech Communication Association (Interspeech)*, Dresden, Germany, 2015, pp. 2282–2286.

[22] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 252–264, 2016.

[23] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*. IEEE, 2009, pp. 77–80.

[24] H. Barfuss, C. Huemmer, G. Lamani, A. Schwarz, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant beamforming," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz (NY), October 2015, pp. 1–5.

[25] H. L. Van Trees, *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory, Part IV. John Wiley & Sons, 2002.

[26] V. Tourbabin and B. Rafaely, "Design of pseudo-spherical microphone array with extended frequency range for robot audition," in *Proc. 42. Jahrestagung für Akustik (DAGA)*, Aachen, 2016.

[27] ——, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1803–1814, Dec 2014.