Analysis of the Robustness of Neural Network-Based Target Activity Detection

Stefan Meier, Daniel Gerber, and Walter Kellermann
Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
Email:{stefan.a.meier, daniel.l.gerber, walter.kellermann}@fau.de

***Abstract*—Many applications in audio signal processing require a precise identification of time frames where a predefined target source is active. In previous work, Artificial Neural Networks (ANNs) with crosscorrelation features showed a considerable potential in this field. In this paper, the performance of ANN-based target activity detection is analyzed in more detail and compared with a well-performing "classical" signal processing method. On the one hand, the impact of the angular distance between target source and interferers is evaluated for both the neural network-based method and the classical one. On the other hand, the sensitivity of both methods to varying Signal-to-Noise Ratio (SNR) conditions is analyzed with respect to the importance of a proper choice of detection thresholds. In the evaluations, the ANN-based method proves its general superiority and also its robustness with respect to a non-ideal choice of detection thresholds.**

I. Introduction

Commonly, audio signal processing algorithms rely on knowledge of the activity of a target source in noisy background, which is denoted as Voice Activity Detection (VAD). Depending on the application, the goal can be to detect any activity of the target source, which can be exploited to, e.g., estimate noise power spectral density during speech absence or to activate an Automatic Speech Recognition (ASR) system in a later processing stage. Other applications require an identification of time instants and/or frequencies with target source dominance, which can be beneficial if source signal statistics or Relative Transfer Functions (RTFs) [1] of the target source need to be estimated during target dominance. Classical VAD methods address scenarios where the target source is a human speaker embedded into non-speech background noise. These conventional VAD methods typically exploit distinctive features of speech like stationarity, harmonic structure and spectral envelopes for discrimination against background noise [2], [3]. Beyond this, multichannel methods can also incorporate spatial information into the detection process [4].

If the target speech signal is corrupted by interfering speakers, these conventional single-channel VAD measures are no longer effective since multiple sources exhibit speech signal characteristics. As a result, spatial diversity of the sources becomes the most important feature for detecting target source activity. In order to be able to exploit spatial information, multi-microphone recordings are required. Conventional methods for acoustic source localization can be modified to allow a discrimination between multiple point sources [5] or between background noise (assumed to be incoherent) and point sources [6]. Similarly, the position of the null of an adaptive nullsteering beamformer can be tracked, indicating a dominant target source if the null is steered towards the target source position [7]. The cross-correlation (or cross power spectral density) between a pair of microphones also allows for detecting activity of a target speaker if the position of its main peak (or the phase difference) corresponds to the expected Time Difference of Arrival (TDoA) of the target source [8]–[11]. Similarly, the Magnitude Squared Coherence (MSC) can be used as feature to distinguish between a coherent target source and incoherent background noise [12]. By calculating the powers of the outputs of a beamformer and a nullformer steered towards the target source, target signal and interference-plus-noise power estimates can be calculated, allowing for estimating the Signal-to-Noise Ratio (SNR) as feature for Target Activity Detection (TAD) [13]–[15]. Finally, probabilistic methods have been discussed for TAD in recent years [16]–[19].

For combining multiple or multidimensional features to a single decision, ANNs have gained interest, especially in the context of single-channel VAD [3], [20]–[24]. Recently, we proposed combining features for multichannel TAD with knowledge on the target source position by means of an ANN [25]–[27]. The concept was proposed for robot audition and offers the advantages that scattering effects at the robot's head can be learned by the ANN, and that a flexible definition of desired detection thresholds is possible. In this paper, this method is evaluated in more detail. On the one hand, its sensitivity to the angular distance of interferers, and especially small target-to-interferer distances, is evaluated. On the other hand, its robustness with respect to the number of interferers is analyzed.

The remainder of this paper is organized as follows: In Section II, the problem is formulated, followed by a description of a conventional TAD method in Section III and the neural network-based one in Section IV. The above-mentioned evaluations are performed in Section V, followed by conclusions in Section VI.

II. Problem Description

At microphone $i \in \{0, \dots, M-1\}$, we record the signal $x_i(k) = s_i(k) + n_i(k)$, which contains target source com-

ISBN 978-0-9928626-7-1

ponents $s_i(k)$ and undesired (i.e., interferer and/or noise) components $n_i(k)$. By forming blocks of length L, we can calculate a time-dependent SNR estimate

$$\mathrm{SNR}_{\mathrm{dB}}(m) = 10\,\log_{10}\left(\frac{\sum_{i=0}^{M-1}\sum_{k=0}^{L-1}\bar{s}_i^2(k,m)}{\sum_{i=0}^{M-1}\sum_{k=0}^{L-1}\bar{n}_i^2(k,m)}\right) \quad (1)$$

for each block m, where $\bar{s}_i(k,m)$ and $\bar{n}_i(k,m)$ denote the mth blocks of $s_i(k)$ and $n_i(k)$, respectively. We define the ground truth for target activity at block m by checking if $\mathrm{SNR}_{\mathrm{dB}}(m)$ exceeds a certain threshold $\vartheta_{0,\mathrm{dB}}$, which yields the desired detection sequence

$$D_{\mathrm{des}}(m) = \begin{cases} 1, & \text{if } \mathrm{SNR}_{\mathrm{dB}}(m) > \vartheta_{0,\mathrm{dB}}, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The threshold $\vartheta_{0,\mathrm{dB}}$ is defined according to the application. If any target activity should be detected, a small value (according to the minimum target signal level) would be required. In this paper, we are interested in target source dominance, which means that the threshold is chosen as $\vartheta_{0,\mathrm{dB}} = 0\mathrm{dB}$. In [25], the generalized crosscorrelation (phase transform) (GCC-PHAT) [28]

$$\hat{r}_{x_i x_j}(\Delta k, m) = \mathrm{DFT}^{-1}\left(\frac{X_i^*(\mu,m)X_j(\mu,m)}{|X_i^*(\mu,m)X_j(\mu,m)|}\right) \quad (3)$$

was found to be the most valuable feature for ANN-based TAD, where $X_i(\mu,m)$ denotes the Short-Time Fourier Transform (STFT) of $x_i(k)$ for frequency bin μ and block m. Therefore, the evaluations in this paper are limited to crosscorrelation features. In order to be able to compare the ANN-based method with conventional methods, a very effective, recently proposed method by Taseska and Habets [8] is taken as reference. In the following section, this method is shortly explained.

III. Narrowband DoA-Based TAD

In [8], a method for TAD is proposed, which defines the hypotheses $\mathcal{H}_t$ for target dominance and $\mathcal{H}_u$ for dominance of undesired components. By using bin-wise Direction of Arrival (DoA) estimates $\phi_{\mathrm{max}}(m,\mu)$ under a free-field assumption, the decision on target source activity for frequency bin μ and frame m is made based on a ratio of posterior probabilities

$$D_{\mathrm{NB}}(m,\mu) = \begin{cases} 1, & \text{if } \frac{\tilde{p}\{\mathcal{H}_t(\mu)|\phi_{\mathrm{max}}(m,\mu)\}}{\tilde{p}\{\mathcal{H}_u(\mu)|\phi_{\mathrm{max}}(m,\mu)\}} > \vartheta_{\mathrm{NB}}, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

In order to estimate the posterior probabilities, the Probability Density Functions (PDFs) $p\left(\phi_{\mathrm{max}}(\mu)|\mathcal{H}_t(\mu);\phi_{\mathrm{tar}}\right)$ and $p\left(\phi_{\mathrm{max}}(\mu)|\mathcal{H}_u(\mu);\phi_{\mathrm{tar}}\right)$ are modeled by a von Mises distribution [29], and a PDF which is nearly uniform but has an anti-mode at ϕ_{tar}, respectively. Since we are interested in a broadband decision, we modify the narrowband method by adding up the posterior probability ratios, which yields the new detector

$$D_{\mathrm{BB}}(m) = \begin{cases} 1, & \text{if } \sum_{\mu=1}^{B_{\mathrm{max}}} \frac{\tilde{p}\{\mathcal{H}_t(\mu)|\phi_{\mathrm{max}}(m,\mu)\}}{\tilde{p}\{\mathcal{H}_u(\mu)|\phi_{\mathrm{max}}(m,\mu)\}} > \vartheta_{\mathrm{BB}}, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

IV. Neural Network-Based TAD

Conventional methods for TAD suffer from a number of shortcomings. On the one hand, the selection of detection thresholds may be cumbersome. This problem becomes crucial when shadowing effects occur and peaks in the crosscorrelation are no longer as pronounced as in the free-field case. In order to overcome these problems, using ANNs for mapping TAD features to a binary decision was proposed in [25]. In the original paper, several features including crosscorrelation sequences, SNR estimates and MSC were discussed. Since the crosscorrelation was found to be the most powerful feature, the other features will not be considered in this paper. The crosscorrelation vector $\mathbf{f}_{\mathrm{CC},i,j}(m)$ at time frame m for the microphone pair (i,j) is defined as

$$\mathbf{f}_{\mathrm{CC},i,j}(m) = \left[\hat{r}_{x_i x_j}(-K,m),\ldots,\hat{r}_{x_i x_j}(+K,m)\right]^{\mathsf{T}} \in \mathbb{R}^{(2K+1)\times 1}. \quad (6)$$

It is possible to stack multiple vectors for different microphone pairs $(i_p,j_p), p \in [0,P-1]$ to one vector $\mathbf{f}_{\mathrm{CC}}(m)$, solving ambiguity problems which can occur in the two-microphone case for sources located symmetrically to the axis defined by the microphone positions. Moreover, taking into account the elevation of the respective sources would be possible with more than one microphone pair.

If the crosscorrelation vector $\mathbf{f}_{\mathrm{CC}}(m)$ is used without any further information, no link to the target source can be established. Therefore, knowledge on the target source DoA relative to the array axis needs to be appended to the input vector of the ANN. To this end, we define a vector $\mathbf{f}_{\phi_{\mathrm{tar}}}(m)$ as

$$\mathbf{f}_{\phi_{\mathrm{tar}}}(m) = \left[\cos(\phi_{\mathrm{tar}}),\ \sin(\phi_{\mathrm{tar}})\right]^{\mathsf{T}} \quad \in \mathbb{R}^{2\times 1}, \quad (7)$$

containing the cosine and sine of ϕ_{tar}. The trigonometric functions were applied in order to account for the circular nature of ϕ_{tar} (e.g., the fact that $0°$ is equivalent to $360°$). By stacking the crosscorrelation vector $\mathbf{f}_{\mathrm{CC}}(m)$ and the target source position vector $\mathbf{f}_{\phi_{\mathrm{tar}}}(m)$, we define the feature vector $\mathbf{f}(m)$ as

$$\mathbf{f}(m) = \left[\mathbf{f}_{\mathrm{CC}}(m)^{\mathsf{T}}, \mathbf{f}_{\phi_{\mathrm{tar}}}(m)^{\mathsf{T}}\right]^{\mathsf{T}} \quad \in \mathbb{R}^{(P(2K+1)+2)\times 1}. \quad (8)$$

For conventional methods, the relation between the target source position ϕ_{tar} and the crosscorrelation sequence has to be established by hand (e.g., by assuming free-field propagation). The ANN-based method does not require any prior knowledge on the topology of the microphone array but allows to learn the relation between ϕ_{tar} and the crosscorrelation sequence during the training process.

V. Experiments

A. Setup

For evaluation, a separate training and test set were defined based on impulse responses measured with the robot NAO (Softbank) at a distance of 1m. A new 12-microphone head for the robot, designed in the European Union-funded project *Embodied Audition for RobotS (EARS)* (http://robot-ears.eu)

ISBN 978-0-9928626-7-1 © EURASIP 2017

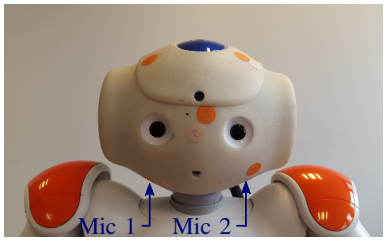

Fig. 1: Robot Nao with microphone positions.

TABLE I: *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)	$\phi_{\text{int},1} \in \{10°, 40°, \ldots, 160°\}$
	$\phi_{\text{int},2} \in \{10°, 40°, \ldots, 160°\}$
	$\phi_{\text{tar}} \neq \phi_{\text{int},1} \neq \phi_{\text{int},2}$

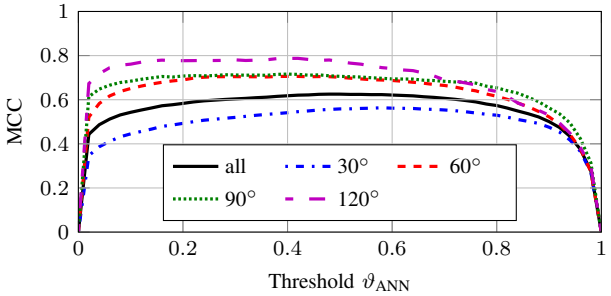

Fig. 2: Choice of the detection threshold for different Δ_{min}.

TABLE II: MCC (with different thresholds) and AUC dependent on the minimum target-to-interferer distance Δ_{min}.

(a) Proposed method.

Δ_{min}	MCC_{opt}	$\text{MCC}_{0.5}$	AUC
all	0.63	0.63	0.90
30°	0.56	0.56	0.87
60°	0.71	0.70	0.93
90°	0.72	0.71	0.94
120°	0.79	0.77	0.96

(b) Method based on [8].

Δ_{min}	MCC_{opt}	AUC
all	0.53	0.84
30°	0.48	0.81
60°	0.66	0.90
90°	0.67	0.92
120°	0.73	0.93

[30] was used for the measurements. Two microphones as marked in Fig. 1 were used (i.e., one microphone pair, $P = 1$), with a horizontal distance of 6.8 cm. The impulse responses from the source positions to the microphones were convolved with speech signals of length 5s. Scenarios with 1 or 2 interferers located at the positions summarized in Table I were simulated for T_{60} times of 150 ms and 400 ms, leading to 42 min of training data and 25 min of test data. In order to create different training and test sets, the test set does not contain the same positions as the training set and different clean source signals were used. For both training and test set, knowledge of the target source position ϕ_{tar} was assumed. The sampling rate f_s was set to 48 kHz, the maximum lag of the crosscorrelation was $K = 15$ and the block length was $L = 2048$, corresponding to 43ms, with an overlap of 50%.

The ANNs were implemented in Python based on the Lasagne library [31]. Feedforward classification neural networks with 2 hidden layers consisting of 30 nodes each were used, with a $\tanh$ function as nonlinearity. This network topology was found to be sensible (but not crucial) during the experiments. Moreover, not only the current feature vector, but also the one of the previous block was fed into the neural network. The training was performed with a ground truth computed according to (2) and a desired detection threshold $\vartheta_{0,\text{dB}} = 0\text{dB}$.

For the modified reference algorithm [8] as explained briefly in Section III, the concentration parameter of the von Mises distribution was chosen such that the best values were achieved ($\kappa = 25$). Since no background noise was considered, the corresponding hypothesis was also not taken into account for the reference algorithm. The bin-wise DoAs were estimated based on the phase (with free-field assumption) with an STFT length of 256. The maximum frequency bin B_{max} was chosen as 13 (≈2.5 kHz) in order to avoid ambiguities at high frequencies.

B. Evaluation measures

A common representation in detection theory is the Receiver Operating Characteristic (ROC), where the true positive rate is plotted over the false positive rate. The Area Under Curve (AUC) describes the area under the ROC curve and indicates whether a suitable threshold can be found for detection, where $\text{AUC} = 1$ denotes a perfect detection and $\text{AUC} = 0.5$ is the expectation value achieved by simple guessing. The AUC only gives an indication on the existence of a good detection threshold but not how precisely the ideal threshold needs to be found, and if slight deviations already lead to a performance decrease. Therefore, the Matthews Correlation Coefficient (MCC) defined as [32]

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (9)$$

will be the main measure in the experimental results, where TP, FP, TN and FN denote the numbers of true/false positives/negatives. $\text{MCC}=1$ describes a perfect detector, whereas simple guessing would lead to $\text{MCC}=0$.

C. Influence of the minimum target-to-interferer distance Δ_{min}

In a first experiment, it is evaluated how interferers located closely to the target source affect the performance of the TAD algorithms. To this end, the angular distance between the target source and the closest interferer is determined and denoted as Δ_{min}. Based on Δ_{min}, the test set is subdivided into subsets, which are evaluated separately. In Fig. 2, the resulting MCC is calculated dependent on the threshold ϑ_{ANN} applied to the probability for target activity returned by the ANN. The

ISBN 978-0-9928626-7-1 © EURASIP 2017

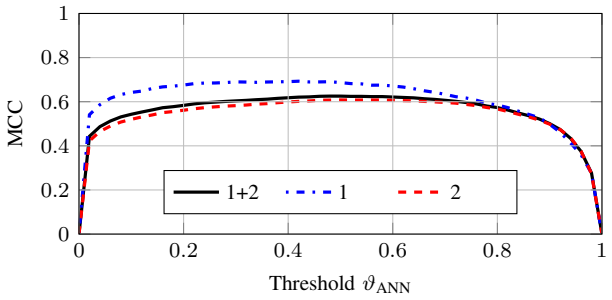

Fig. 3: Choice of the detection threshold for varying numbers of interferers.

curves illustrate the results obtained for the whole test set and four subsets with different Δ_{min}. Obviously, a large angular distance between target source and closest interferer leads to higher MCC values. It can be observed that all curves are relatively flat in the center, which indicates that the choice of the detection threshold ϑ_{ANN} is not critical and deviations from the optimum detection threshold can be accepted. In Table IIa, this observation is confirmed. Here, two different thresholds are applied to the output probabilities: On the one hand, the optimum value MCC_{opt} is shown, corresponding to the maxima of the respective curves in Fig. 2. On the other hand, the value $\text{MCC}_{0.5}$ corresponds to the MCC obtained for $\vartheta_{\text{ANN}} = 0.5$, which would be chosen without prior knowledge. Obviously, the uninformed decision threshold nearly leads to the same result as the optimum one. From $\Delta_{\text{min}} = 30°$ to $\Delta_{\text{min}} = 120°$, $\text{MCC}_{0.5}$ improves from 0.56 to 0.77.

In Table IIb, the results obtained with the reference method are summarized. As before, the detection threshold ϑ_{NB} was optimized by iterating over possible thresholds and choosing the one with the best MCC value. As expected, the reference method faces problems when it comes to distinguishing between the target source and a close-by interferer. In this case, phase differences are small and head shadowing effects would have to be incorporated. Therefore, a low MCC of 0.48 is achieved for $\Delta_{\text{min}} = 30°$ (compared to 0.56 for the proposed one), but the method becomes more competitive for $\Delta_{\text{min}} \geq 90°$.

D. Influence of the number of interferers N_{int}

Another important factor for the performance of TAD methods is the number of interferers. In a second experiment, this influence is evaluated in more detail. To this end, the test set is subdivided into scenarios with one interferer and those with two interferers. The resulting MCC values for the entire set and the two subsets are plotted in Fig. 3 (dependent on the threshold ϑ_{ANN}). Again, both curves are very flat around $\vartheta_{\text{ANN}} = 0.5$, indicating a low sensitivity to the actual choice of the threshold. Due to the greater number of possible combinations (according to Table I), the test set contains more scenarios with 2 interferers than with 1 interferer, which explains why the curve for the entire set is close to that obtained with 2 interferers. In the left column of

TABLE III: MCC (with different thresholds) and AUC dependent on the number of interferers N_{int}.

(a) Proposed method.

N_{int}	Non-matched			Matched		
	MCC_{opt}	$\text{MCC}_{0.5}$	AUC	MCC_{opt}	$\text{MCC}_{0.5}$	AUC
1+2	0.63	0.63	0.90			
1	0.69	0.69	0.92	0.69	0.69	0.91
2	0.61	0.61	0.90	0.63	0.62	0.90

(b) Reference method based on [8].

N_{int}	MCC_{opt}	AUC
1+2	0.53	0.84
1	0.59	0.87
2	0.53	0.84

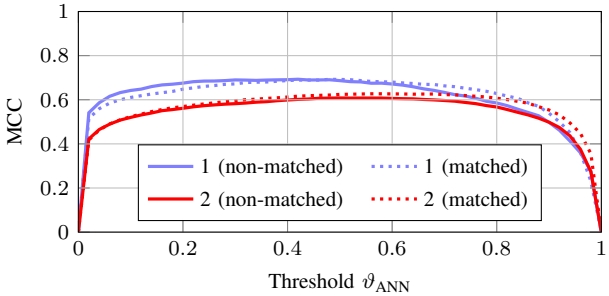

Fig. 4: Matched training for varying numbers of interferers.

Table IIIa, it is confirmed that, again, choosing a threshold of 0.5 nearly leads to the optimum results. Obviously, scenarios with two interferers make a detection of the target source more complicated (which is also due to the lower mean distance between target source and closest interferer). However, the choice of the detection threshold is not affected by the number of interferers – in both cases, 0.5 leads to a nearly optimum decision. As before, the reference method in Table IIIb leads to significantly worse results in terms of MCC than the neural network-based one (for both subsets).

A final experiment evaluates the question if the results in Fig. 3 can be improved by performing an individual training for scenarios with 1 interferer and for scenarios with 2 interferers. For this evaluation, it is assumed that the number of interferers is estimated correctly and the respective neural network model is loaded. In Fig. 4, the results obtained with this *matched training* are compared with those obtained by training only one ANN model for the whole training set. The corresponding values for MCC_{opt} and $\text{MCC}_{0.5}$ are summarized in the right half of Table IIIa. Both the curves and the numbers in the table show that, in fact, the matched training does not lead to a significant improvement for both subsets. In the 2-interferer case, a slight enhancement with an optimum detection threshold may be deduced. However, the curve is less symmetric than before, and when choosing a threshold of 0.5, there is almost no improvement. From this result, one can conclude that the original neural network

ISBN 978-0-9928626-7-1 © EURASIP 2017

model is well able to generalize for a varying number of interferers, since training over the complete set does not lead to a performance degradation compared with "matched" training. Again, the reference method (Table IIIb) performs worse than the ANN-based one.

VI. Conclusions

In this paper, a recently proposed method for ANN-based TAD was evaluated with respect to its sensitivity towards a proper choice of a detection threshold, with focus on interferers located close to the target source and varying numbers of interferers. It could be shown that the choice of the detection threshold is generally not crucial, and choosing a default threshold of 0.5 nearly leads to the same results as an optimized threshold would. Obviously, interferers located close to the target source affect the performance of this method. The reference method, however, suffers more severely from close interferers. It was also shown that the choice of a suitable detection threshold is not affected by the minimum target-to-interferer distance Δ_{min}. The number of interferers did not affect the choice of the detection threshold either, but more interferers degraded the overall performance. Matched training for a certain number of interferers did not lead to an improvement, proving that training for a large set of scenarios is sufficient. Future evaluations may include taking into account two angular dimensions and radial distances for the source positions. In this case, additional microphone pairs may be considered.

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 J. Advances Signal Process.*, vol. 2015, no. 1, pp. 1–15, 2015.

[4] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072–1077, 2010.

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

[6] 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 Hands-free Speech Commun. Microphone Arrays (HSCMA)*. IEEE, 2011, pp. 92–97.

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

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

[9] 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.

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

[10] 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. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2007, pp. 222–225.

[12] 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.

[13] 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.

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

[15] 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 Process. (ICASSP)*. IEEE, 2010, pp. 2834–2837.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, 2015, pp. 2282–2286.

[23] 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.

[24] N. Moritz, J. Drefs, H. Baumgartner, and J. Rennies, "Sprachaktivitätserkennung basierend auf Deep Neural Networks für Anwendungen in Film und Fernsehen," in *Proc. 42. Jahrestagung für Akustik (DAGA)*, 2016, pp. 960–963.

[25] S. Meier and W. Kellermann, "Artificial neural network-based feature combination for spatial voice activity detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, San Francisco, USA, September 2016, pp. 2987–2991.

[26] ——, "Relative impulse response estimation during doubletalk with an artificial neural network-based step size control," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Xi'an, China, September 2016.

[27] D. Gerber, S. Meier, and W. Kellermann, "Efficient target activity detection based on recurrent neural networks," in *Proc. IEEE Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA)*, San Francisco, USA, 2017.

[28] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[29] K. V. Mardia and P. E. Jupp, *Directional statistics*. New York, USA: Wiley, 1999.

[30] 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.

[31] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly *et al.*, "Lasagne: First release," *Zenodo: Geneva, Switzerland*, 2015.

[32] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

ISBN 978-0-9928626-7-1