# The Minimum Overlap-Gap Algorithm for Speech Enhancement

**POUL HOANG** [1,2], **ZHENG-HUA TAN** [1], **(Senior Member, IEEE),**
**JAN MARK DE HAAN** [2], **AND JESPER JENSEN** [1,2]
[1] Department of Electronic Systems, Aalborg University, 9000 Aalborg, Denmark
[2] Oticon A/S, 2765 Smoerum, Denmark

Corresponding author: Poul Hoang (phoa@demant.com)

**ABSTRACT** In this paper, we propose a novel speech enhancement paradigm which can effectively solve the problem of retrieving a desired speech signal in a multi-talker environment. The proposed speech enhancement paradigm involves a three-step procedure consisting of separation, ranking, and enhancement. First, a speech separation system – which could be a conventional spatial filter bank or more advanced separation systems – separates mixtures of speech signals captured by microphones into speech signals from candidate speakers. Next, novel ranking algorithms – proposed in this paper – are applied to determine the talker-of-interest amongst the separated speech signals. Finally, the speech signal of the talker-of-interest is estimated as a linear combination of the separated signals, whose weights are determined by the ranking algorithms. We propose ranking algorithms, which exploit turn-taking patterns between conversational partners in order to determine the talker-of-interest amongst competing speakers. Unlike some existing solutions, our ranking algorithms do not require access to additional sensors, e.g., EEG electrodes, cameras, etc., but only rely on microphone signals. Specifically, the proposed algorithms rank the separated speech signals based on the probability of speech overlaps and gaps with the user's own voice. The speech signal with highest ranking is the talker with *minimum* probability of speech overlap and gap with the user's own voice. The proposed ranking algorithms are shown highly effective at determining the talker-of-interest, since conversational partners, i.e., the user and the talker-of-interest, behaviorally avoid speech overlaps and gaps. We evaluate the proposed speech enhancement paradigm in two practical hearing aid related applications, where the objective is to enhance a speech signal of a conversational partner in a multi-talker environment. The results of the evaluation demonstrate that the proposed speech enhancement systems in both applications significantly outperform conventional speech enhancement systems.

**INDEX TERMS** Speech enhancement, turn-taking, multichannel noise reduction, DOA estimation, multi-talker problem, estimation of the talker-of-interest.

## I. INTRODUCTION

The cocktail party problem is often regarded as one of the most difficult situations any speech enhancement system may encounter. The complexity in the acoustic environment is vast and its composition may include multiple competing speakers, music, reverberation, and noise. Solving the cocktail party problem, i.e., the speech signal(s)-of-interest, i.e. the *target signal(s)*, is commonly the goal for speech enhancement systems in applications such as hearing assistive devices (HADs) and speaker-phone systems.

The enhancement system in these applications is crucial for many humans as they rely on the aid to communicate more efficiently in noisy environments, particularly when competing speech and noise become dominant. However, achieving effective suppression of loud competing speech and noise remains a remarkably difficult problem to solve even with the most recent state-of-the-art speech enhancement systems.

The problem of interest in this paper is to enhance a conversational partner, i.e., the talker-of-interest, in the presence of multiple competing speakers and noise. The competing speakers are obviously undesired and can potentially be louder than the conversational partner. In order to be able to enhance the conversational partner in such

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman [ID].

multi-speaker situations, any enhancement system faces the question: *"Who is the user listening and talking to?"*. The traditional speech enhancement paradigm for single-microphone systems involves estimation of temporal statistics of the conversational partner and noise for implementation of linear filters. For the multiple microphone case, beamformers are often implemented and typically require estimation of the direction-of-arrival (DOA) and/or spatio-temporal statistics of the conversational partner and noise [1]–[5]. However, the presence of multiple speakers poses a great estimation challenge, since the conversational partner and competing speakers are often indistinguishable from an acoustic perspective. In worst case scenarios, speech enhancement algorithms might in fact suppress the conversational partner and enhance competing speech. For example, current DOA estimators such as SRP-PHAT [6], maximum likelihood [7], [8], and deep learning-based DOA estimators [9], are not able to robustly handle a conversational partner in a multi-speaker environment, without additional a priori information on the conversational partner's location or voice activity. Consequently, these DOA estimators will indecisively switch between the candidate speakers as being the conversational partner leading to an enhanced signal of unacceptable intelligibility and quality.

In this paper, we propose a speech enhancement paradigm that can efficiently identify the conversational partner in a multi-speaker environment and retrieve the desired speech signal. The paradigm is described through the three step-procedure as shown in Fig. 1.

In the first step, the noisy microphone signals are fed into a speech separation system to separate mixtures of speech signals into individual source signals/components, which we refer to as candidate speakers. Example of speech separation systems include beamforming systems which separate speech using beams steered in different directions, or deep neural network (DNN) based separation algorithm e.g. uPit [10], [11] and TasNet [12], [13]. Some applications allow microphones to be placed physically on the candidate talker, in which case the separation is trivial.

In the second step, the separated candidate speakers are ranked according to their likelihood of being the conversational partner. Existing ranking strategies may involve additional sensor signals and prior knowledge to support the decision of estimating the conversational partner channel after speech separation. As an example, beamforming systems in HADs often rank, or simply assume, the frontal speaker as the most likely conversational partner [2]. However, unfortunately, the user may not always face the conversational partner in all situations, which leads to a loss of performance. Alternatively, estimated candidate speakers may be ranked using EEG-signals, retrieved from EEG-electrodes placed on the scalp of the user, to detect the user's attention on conversational partner, EOG-signals to estimate eye-gaze from in-ear electrodes, and cameras to track eye-movements and estimate eye-gaze [14]–[18]. While these signals have the potential to support the decision

of determining the talker-of-interest, they require additional sensors which increase equipment cost, increase wearing inconvenience, and likely also increase computational cost and power consumption. These trade-offs make acquisition of EEG, EOG, and visual signals impractical for small devices such as HADs where power consumption and wearing inconvenience matters for the end user.

Finally, the last step involves enhancement of the conversational partner signal. The enhanced signal is formed as a linear combination of the separated speech signals where the weights are determined from the speaker ranking algorithm.

Additionally, we propose a method to the ranking problem in Fig. 1, which does not require additional sensors apart from microphones. A microphone-only system is highly desirable from a practical perspective, both due to the cost of additional sensors and from a algorithm complexity perspective. Our method is based on exploiting the conversational behavior between the user and the conversational partner. We use the so-called *turn-taking* behavior between two conversational partners [19]–[23] to rank the candidate speakers according to the talker which is most likely the user's conversational partner. Specifically, the method analyses the speech overlaps and gaps between the user and a candidate speaker to quantify turn-taking, and then selects the speaker with minimum probability of speech overlap and gap with the user as the talker-of-interest.

This paper is organized as follows. Sec. II introduces the basics of conversation and turn-taking behavior and its potential use in ranking the candidate speakers and determining the talker-of-interest. In Sec. III, we derive our minimum overlap-gap (MOG) method and propose statistical models of speech overlap and gap behavior between a user and a conversational partner. Based on the statistical model, we propose an extension, namely, the Bayesian MOG (BMOG) algorithm. In Sec. IV, we describe the estimation of the parameters for the proposed statistical models of turn-taking from datasets of real conversations. We use the statistical models to derive the theoretical performance of the (B)MOG algorithm. Finally in Sec. V, we evaluate the performance of the proposed speech enhancement paradigm and (B)MOG algorithms in two speech enhancement applications.

## II. SPEECH INTERACTION IN CONVERSATIONS
Determining the talker-of-interest and ranking the candidate speakers are needed for the proposed speech enhancement paradigm and can be an extremely difficult problem to solve. We propose to rank the candidate talker using the turn-taking model presented in [19]. Human interaction is a group of behavioral mechanisms that are taught since childhood to use when engaged in conversations to structurize exchange of information [20]. *Addressing* and *turn-taking* mechanisms found in conversations are examples of interaction management between conversational partners [20].

Addressing is used by the addressee, i.e., the talking person, to indicate whom the speech is directed to. For example, humans may use gaze, gestures, and speech to indicate
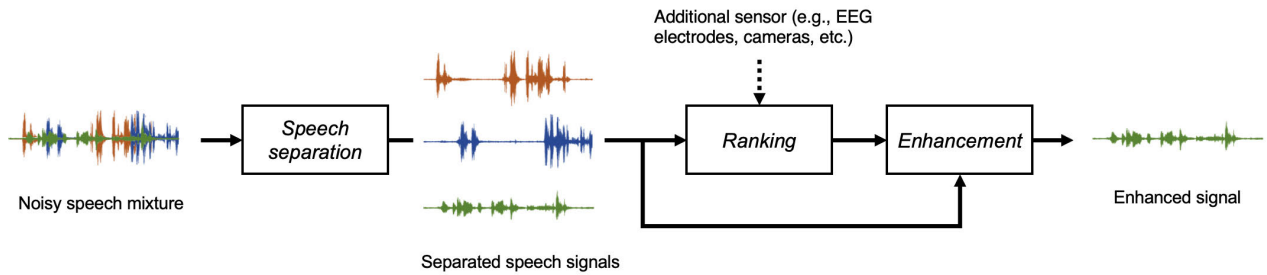
**FIGURE 1.** Speech enhancement paradigm for enhancement of a conversational partner in multi-talker situations.

the conversational partner. Strong indicators are typically head pose and eye-gaze which potentially could be utilized by speech enhancement systems to determine the talker-of-interest [20]. However, measuring the head pose and eye-gaze would usually require additional sensors such as accelerometers, electrodes, or cameras for applications such as HADs.

The turn-taking mechanism is another type of interaction management and is universal across cultures and languages. Turn-taking is used to structurize conversations. Turn-taking is used to coordinate who should speak next and when, to ensure that only one speaker is talking at a time, while others remain silent. Conversational partners may occasionally overlap and gap in conversations, but these are often of short duration such as when the listener responds the talker by saying "yes" or "uh hm" [19], [24]. In order to maintain rapid turn-taking, listeners also try to predict the end of a speech utterance of their conversational partner to minimize speech overlap and gap.

We use the turn-taking model in [19] to model a) the conversational behavior between the user and the conversational partner, and b) the voice activity pattern between the user and a competing speaker. We may describe a) and b) in terms of four voice activity states.

$\mathcal{S}_1$ : Conversational partner/competing speaker speaks while user is silent.

$\mathcal{S}_2$ : User speaks while conversational partner/competing speaker is silent.

$\mathcal{S}_3$ : Conversational partner/competing speaker and user are both silent.

$\mathcal{S}_4$ : Conversational partner/competing speaker and user are both speaking.

State $\mathcal{S}_1$ and $\mathcal{S}_2$ are *turns* of the conversational partner/competing speaker and user, respectively, while state $\mathcal{S}_3$ is referred to as *gaps* or pauses and state $\mathcal{S}_4$ is referred to as *overlaps* [19], [24], [25]. In [24] it was found that 77% of all recorded conversations between a user and a conversational partner were in state $\mathcal{S}_1$ or $\mathcal{S}_2$, 19.2 % belonged to state $\mathcal{S}_3$, and 3.8% were in state $\mathcal{S}_4$. For a user and a competing speaker, the proportion of time spent in each state, may be argued to be significantly different compared to a user and a conversational partner. Specifically, a larger proportion of speech overlaps and gaps would be expected between a user and a competing speaker, since the turn-taking mechanisms

would not exist. In addition, when the conversational partners are exposed to noisy environments, the proportion of time spend in each state changes, with overlaps becoming more common as the noise level increases. In [26] it was found that in very noisy environments the proportion of time spent in state $\mathcal{S}_1$ or $\mathcal{S}_2$ decreased from 70% at a noise level of 54 dB SPL to 50% at 78 dB SPL, $\mathcal{S}_3$ increased from 8% at 54 dB SPL to 24% at 78 dB SPL, and for $\mathcal{S}_4$ from approximately 22% at 54 dB SPL to 26% at 78 dB SPL, where normal conversation breaks down. A possible reason for these observations is that conversational partners insist on maintaining rapid turn-taking during conversations, resulting in poorer timing and prediction of their partners end of a turn, hence increasing the proportion of overlaps and gaps.

These results indicate that humans rely significantly on turn-taking to maintain normal conversations even in very noisy environments as conversations otherwise would break down. Although speech overlaps and gaps become more frequent in noisy environments, these conversational patterns remain robust in noisy condition and the turn-taking patterns between a user and a conversational partner would presumably still be significantly different than the voice activity patterns between a user and a competing speaker. Hence, in the following we propose a method that exploits these turn-taking patterns to determine the talker-of-interest in a multi-talker environment.

## III. THE MINIMUM OVERLAP-GAP ALGORITHM

In this section, we derive the proposed algorithm for ranking the candidate speakers using expected turn-taking patterns. Our primary focus in this section is the task of ranking the speakers by their likelihood of being the conversational partner, i.e. the *Ranking* block in Fig. 1.

First, the speech separation system separates mixtures of speech signals into individual discrete time-sequences $s_i(n)$, $i = 0, 1, \ldots, I$, where $s_0(n)$ is the user's own voice, and the remaining $s_i(n)$, $i = 1, \ldots, I$ are the $I$ candidate speech signals. For each speech signal $s_i(n)$, a binary output $\alpha_i(n)$ of voice activity detector (VAD) is defined as

$$\alpha_i(n) = \begin{cases} 1, & \text{if } s_i(n) \text{ contains speech at time } n \\ 0, & \text{if } s_i(n) \text{ contains no speech at time } n. \end{cases} \quad (1)$$

where $\alpha_0(n)$ is the user's own voice VAD (OVAD). We assume that $\alpha_i(n)$ represents the actual speech activity of the various speech sources – as we demonstrate in Sec. V-E, the proposed ranking and enhancement system work well, even when $\alpha_i(n)$ are estimated from sources separated with a practical beamforming system. Fig. 2 shows an example of VAD outputs a real conversation between the user and the conversational partner in addition to two competing speakers. The outputs of the VADs are used to determine the voice activity state between the user and a candidate speaker $i$ i.e.

$\mathcal{S}_1$ : if $\alpha_0(n) = 0$ and $\alpha_i(n) = 1$.
$\mathcal{S}_2$ : if $\alpha_0(n) = 1$ and $\alpha_i(n) = 0$.
$\mathcal{S}_3$ : if $\alpha_0(n) = 0$ and $\alpha_i(n) = 0$.
$\mathcal{S}_4$ : if $\alpha_0(n) = 1$ and $\alpha_i(n) = 1$.

As discussed in Sec. II, conversational partners use turn-taking when engaged in a conversation. A consequence of the turn-taking mechanism is that conversational partners avoid speech overlaps and gaps, i.e., they minimize the proportion of time spent in state $\mathcal{S}_3$ and $\mathcal{S}_4$. In the following, we use this observation to propose an algorithm exploiting this to determine the talker-of-interest.
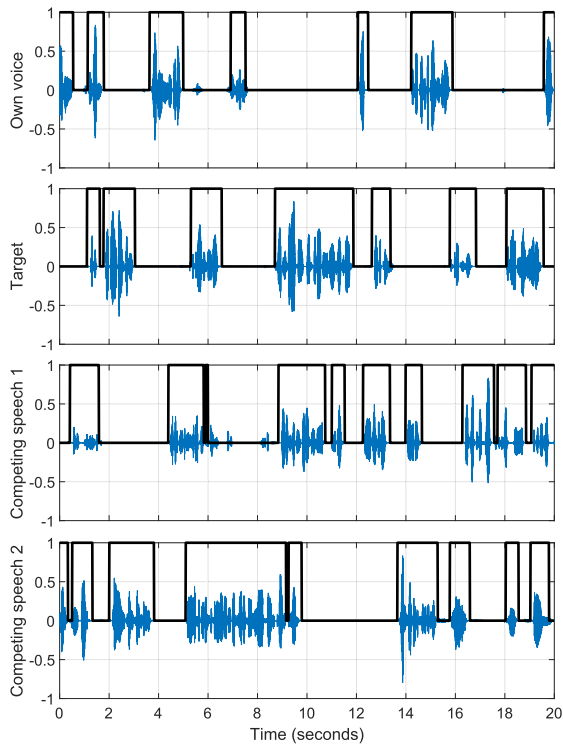


**FIGURE 2.** Speech signals with VAD outputs plotted on top of real conversations between the own voice and the conversational partner. The two top plots are conversations between a user and a conversational partner (target). The two bottom plots are competing speakers unrelated to the conversation between the user and the conversational partner.

## A. MINIMUM PROBABILITY OF SPEECH OVERLAP AND GAP

The paradigm presented in Fig. 1 ranks the candidate speakers using their voice activity patterns, prior to the enhancement. The proposed algorithm selects the speaker with minimum probability of speech overlap and gap related to the user's own voice as the talker-of-interest. We refer to this method as the Minimum Overlap-Gap (MOG) algorithm. Let $A_i(n)$, $i = 0, 1, \ldots, I$ be Bernoulli random variables of the VADs and let $\alpha_i(n)$, $i = 0, 1, \ldots, I$ be their corresponding realizations. The probability of a speech overlap and speech gap between the user's own voice and candidate speaker $i$, is denoted as $P_{A_0A_i}(\alpha_0(n) = 1, \alpha_i(n) = 1)$ and $P_{A_0A_i}(\alpha_0(n) = 0, \alpha_i(n) = 0)$, respectively. The MOG algorithm selects the speaker with minimum probability of overlaps and gaps:

$$\hat{i}_{\text{MOG}}(n) = \arg \min_{i \in \{1, \ldots, I\}} \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = k), \quad (2)$$

where $\hat{i}_{\text{MOG}}(n)$ is the estimated conversational partner channel index and minimizing the cost in (2) is equivalent to minimizing the occurrences of the states $\mathcal{S}_3(n, i)$ and $\mathcal{S}_4(n, i)$, i.e. gaps and overlaps, respectively. Alternatively, the optimization problem may also be formulated as maximizing the probability of mutual exclusion between the binary sequences $\alpha_0(n)$ and $\alpha_i(n)$ (see Appendix A) i.e.

$$\hat{i}_{\text{MOG}}(n) = \arg \max_{i \in \{1, \ldots, I\}} \sum_{k=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = 1 - k).$$

$$(3)$$

Furthermore, as shown in Appendix B, solving (3) is also equivalent to finding the candidate speaker index, which maximizes the mean-square-error (MSE) between the user own-voice VAD (OVAD) and candidate speaker's VAD, i.e.,

$$\hat{i}_{\text{MOG}}(n) = \arg \max_{i \in \{1, \ldots, I\}} \mathbb{E}\left[ (A_0(n) - A_i(n))^2 \right]. \quad (4)$$

Note that the optimization problem is bounded in [0, 1] as $A_0(n)$ and $A_i(n)$ are binary values. The definition of the MOG algorithm in (4) is a maximization of the MSE between two binary sequences and is thus computationally simple.

### B. BAYESIAN MOG FOR PROBABILITY-BASED SPEAKER RANKING

Probability-based ranking of the candidate speakers can provide additional insights compared to the MOG algorithm in (4) which only identifies a single talker-of-interest. In this approach, a posterior probability is estimated for each candidate speaker which quantifies the uncertainty of a candidate speaker being the talker-of-interest. This information can be particularly useful for a speech enhancement system, for example, to adjust the level of noise suppression.

#### 1) STATISTICAL MODELS OF THE SUM OF SQUARED ERROR

One approach to derive posterior probabilities for each candidate speaker, is to statistically model the distribution of overlaps and gaps between 1) a user and a conversational partner, and 2) a user and a competing speaker, and then use Bayes theorem to estimate the probabilities. To model the statistical distribution of overlaps and gaps, we introduce

the random variable $Z_i(n)$, which represents the squared error between the own voice VAD and the candidate speaker VAD:

$$Z_i(n) = (A_0(n) - A_i(n))^2, \tag{5}$$

where $Z_i(n)$, $A_0(n)$, and $A_i(n)$ are Bernoulli random variables. The random variable $Z_i(n)$ quantifies if $A_0(n)$ and $A_i(n)$ are overlapping or gapping, i.e., when $Z_i(n) = 0$, or not. We define the sum of squared errors (SSEs) as

$$\Phi_i(n) = \sum_{k=n-N+1}^{n} Z_i(k), \tag{6}$$

where $N$ is the number of past observations of $Z_i(n)$ upon which the decision will be based. The SSE quantifies the total amount of observed overlaps and gaps within $N$ observations. Low SSEs indicate large amounts of overlaps and gaps between $A_0(n)$ and $A_i(n)$, whereas high SSEs indicate small amounts of overlaps and gaps. It is also worth noting that $N$ is related to the *integration time*, which we define as

$$T_{\text{int}} = N \cdot f_{\text{s,vad}}, \tag{7}$$

where $f_{\text{s,vad}}$ is the sampling frequency of the VADs. The integration time $T_{\text{int}}$, is easier interpreted than $N$ as it also accounts for the sampling frequency of the VADs.

In order to model the distribution of $\Phi_i(n)$, we use that $\Phi_i(n)$ is a sum of $N$ Bernoulli distributed random variables. For independently and identically distributed $Z_i(n)$, then $\Phi_i(n)$ follows a binomial distribution. However, preliminary experiments with natural conversations have shown that observations of $\Phi_i(n)$ have a higher dispersion than a binomial distribution, hence the binomial distribution is too restrictive to explain the observations. Instead, we have found that a beta-binomial distribution provides a significantly better fit than the binomial distribution. The beta-binomial distribution is parameterized by $N$ and two shaping parameters $\gamma$ and $\beta$ and its probability mass function (PMF) is given as

$$p_{\Phi_i}(\phi_i; \gamma, \beta, N) = \binom{N}{\phi_i} \frac{B(\phi_i + \gamma, N - \phi_i + \beta)}{B(\gamma, \beta)}. \tag{8}$$

$B(\cdot, \cdot)$ is the Beta-function parameterized by $\gamma$ and $\beta$, and

$$\binom{N}{\phi_i} = \frac{N!}{\phi_i!(N - \phi_i)!}, \tag{9}$$

denotes the binomial coefficient. In the remaining part of the paper, we use the PMF notation $p_{\Phi_i}(\phi_i; \gamma, \beta, N) \triangleq p(\Phi_i = \phi_i; \gamma, \beta, N)$ for brevity. First, we statistical model $\Phi_i$ when the user is engaged in a conversation and afterwards model $\Phi_i$ for the interaction between the user and a competing speaker. Hence, the first statistical distribution $p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N)$ is fitted to observations of SSEs between a user and conversational partner engaged in a conversation, where the subscript $t$ denotes that the shaping parameters are related to the true conversational partner. The second distribution $p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N)$ is fitted to observations of SSEs between a user and competing speakers, where the user and competing speaker are engaged in different conversations.

### 2) HYPOTHESIS TESTING

In order to estimate probabilities for each candidate speaker, we define $I$ hypotheses

$\mathcal{H}_i$: Candidate speaker $i$ is the conversational partner, and the remaining $I-1$ speakers are competing speakers for $i = 1, \ldots, I$.

Under $\mathcal{H}_i$, it follows that $\Phi_i$ is distributed according to $p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N)$ and $\Phi_j$ for $j \neq i$, is distributed according to $p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N)$, i.e.

$$\Phi_i \sim p_{\Phi_i}(\phi_i|\mathcal{H}_i) \triangleq p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N)$$
$$\Phi_j \sim p_{\Phi_j}(\phi_j|\mathcal{H}_i) \triangleq p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N) \text{ for } j \neq i. \tag{10}$$

For each time $n$, we observe realizations, $\phi_k$, of $\Phi_k$ for all $k = 1, \ldots, I$. Assuming that $\Phi_k$, are statistically independent, the likelihood function conditioned on $\mathcal{H}_i$ is given by

$$p_{\Phi_1, \ldots, \Phi_I}(\phi_1, \ldots, \phi_I|\mathcal{H}_i)$$
$$= \prod_{k=1}^{I} p_{\Phi_k}(\phi_k|\mathcal{H}_i)$$
$$= p_{\Phi_i}(\phi_i|\mathcal{H}_i) \prod_{j \in \mathcal{I} \setminus i} p_{\Phi_j}(\phi_j|\mathcal{H}_i)$$
$$= p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N) \prod_{j \in \mathcal{I} \setminus i} p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N), \tag{11}$$

where $\mathcal{I} = \{1, \ldots, I\}$ is the set of candidate speaker indices, and $\mathcal{I} \setminus i$ denotes the set of competing speakers under hypothesis $\mathcal{H}_i$, i.e. $\mathcal{I}$ excluding the element $i$. Using Bayes theorem, the posterior probability of $\mathcal{H}_i$ is given by

$$P(\mathcal{H}_i|\phi_1, \ldots, \phi_I)$$
$$= \frac{P(\mathcal{H}_i) p_{\Phi_1, \ldots, \Phi_I}(\phi_1, \ldots, \phi_I|\mathcal{H}_i)}{p_{\Phi_1, \ldots, \Phi_I}(\phi_1, \ldots, \phi_I)}$$
$$= \frac{P(\mathcal{H}_i) p_{\Phi_i}(\phi_i; \gamma_t, \beta_t, N) \prod_{j \in \mathcal{I} \setminus i} p_{\Phi_j}(\phi_j; \gamma_v, \beta_v, N)}{\sum_{k=1}^{I} P(\mathcal{H}_k) p_{\Phi_k}(\phi_k; \gamma_t, \beta_t, N) \prod_{l \in \mathcal{I} \setminus k} p_{\Phi_l}(\phi_l; \gamma_v, \beta_v, N)}, \tag{12}$$

where $P(\mathcal{H}_i)$ is the prior probability of the conversational partner being channel $i$. This method of estimating the posterior probability is referred to as the Bayesian MOG algorithm.

## IV. PARAMETER ESTIMATION FROM CONVERSATIONAL SPEECH DATABASE

To implement the Bayesian MOG (BMOG) algorithm in (12), the shaping parameters $\gamma_t$, $\beta_t$, $\gamma_v$, and $\beta_v$ for the statistical models $p_\Phi(\phi; \gamma_t, \beta_t, N)$ and $p_\Phi(\phi; \gamma_v, \beta_v, N)$ are estimated from speech databases containing real conversations. Next, using the estimated statistical models, we analyze the theoretical speaker ranking performance of the MOG algorithm in terms of misclassification rate.
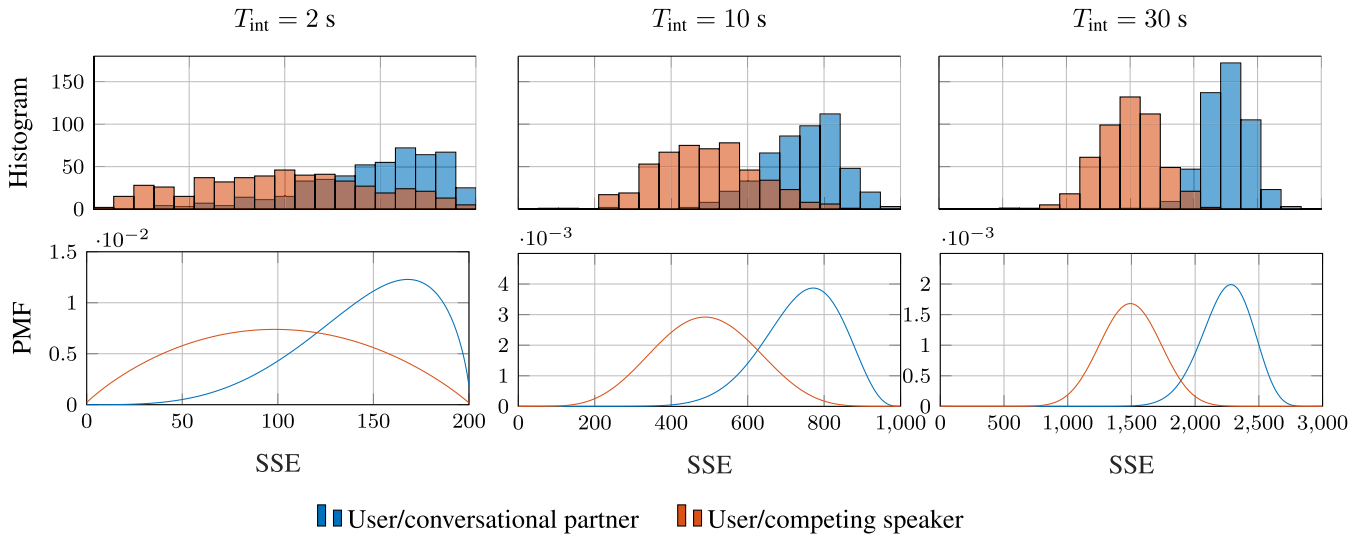
**FIGURE 3.** Histograms of $K = 500$ observations of SSEs from real conversations are shown in the upper plots. The bottom plots show beta-binomial distributions where the parameters are found from the observations. In all plots the blue color denotes the conversational partner and the red color denotes the competing speaker.

## A. SETUP AND SPEECH DATABASE

### 1) CONVERSATIONAL SPEECH DATABASE

In order to estimate the shaping parameters $\gamma_t$, $\beta_t$, $\gamma_v$, and $\beta_v$, we use the speech database in [27] which contains dialogues between 19 pairs of native-Danish talkers recorded during a task dialog experiment. The participants had normal hearing and were coupled into pairs to collaborate solving DiapixUK tasks [28]. DiapixUK is *spot-the-difference* tasks where partners were given two almost identical cartoon pictures with a few differences. The participants were not allowed to view each others pictures, but had to solve the DiapixUK task by exchanging descriptions of their picture through verbal communication. The partners were placed in different sound booths and communicated through headphones and head-worn microphones. The experiment had four test conditions: 1) native language (Danish) and no noise, 2) native language (Danish) and babble noise, 3) second language (English) and no noise, and 4) native language (English) and babble noise.

### 2) VOICE ACTIVITY DETECTION

The presence of speech in the signal $s_j(n)$ is determined by a binary VAD which produces an output sequence $\alpha_j(n) = \{0, 1\}$ for either of the speakers in the dialogue. For voice activity detection, we used the robust voice activity detector (rVAD) proposed in [29] applied to the essentially noise-free dialogue recordings. The input to rVAD is $L$ consecutive samples of $s_i(l)$ with sampling frequency $f_s$. The output of rVAD is a sequence of $N$ voice activity decisions $\alpha_i(n)$ at sampling frequency $f_{s,vad} = 100$ Hz. Version rVAD2.0 was used in this paper and can be found in [30].

## B. PARAMETER ESTIMATION FOR THE BETA-BINOMIAL DISTRIBUTION

We used the speech data set recorded in a quiet condition and in Danish language for parameter estimation. The speech

signals are sampled at 22.05 kHz but downsampled to 16 kHz for compatibility with rVAD. In order to collect observations of the SSEs for a user and a conversational partner, we used the following procedure:

1) Select an integration time $T_{int}$, e.g. $T_{int} = 10$ seconds, where the integration time is related to $N$ by $N = \frac{T_{int}}{f_{s,vad}}$.
2) Divide the speech signals into non-overlapping segments with length $T_{int}$.
3) Apply the rVAD on the speech signals of conversational partners.
4) Compute the SSE from the VAD outputs using (6).

To gather observations of the SSE between the user and a competing speaker, we perform a similar procedure, but instead of choosing a matching conversational pair, we randomly choose two non-conversational speakers to form a pair and compute the SSE. Histograms and fitted beta-binomial distribution of SSEs between a user and a conversational partner, as well as a user and a competing speaker are shown in Fig. 3 for different integration times. Clearly and as expected, the separability between $p_\Phi(\phi; \gamma_t, \beta_t, N)$ and $p_\Phi(\phi; \gamma_v, \beta_v, N)$ becomes greater as $T_{int}$ becomes larger. The dispersion of SSE becomes smaller for both distributions as $T_{int}$ increases. The shaping parameters $\gamma_t$, $\beta_t$, $\gamma_v$, and $\beta_v$ are functions of $T_{int}$.

### 1) PARAMETER ESTIMATION OF $\gamma_t$, $\beta_t$, $\gamma_v$, AND $\beta_v$ GIVEN $T_{int}$

For each $T_{int}$, the parameters $\gamma_t$, $\beta_t$, $\gamma_v$, and $\beta_v$ are estimated using observations of the SSEs. The observations of SSEs are denoted as $\phi_t^{(k)}$ and $\phi_v^{(k)}$, $k = 1, \ldots, K$, respectively, where the subscript $t$ denotes the SSE between the user and conversational partner, $v$ is the SSE between the user and a competing speaker, and $K$ is the total number of observations. Each observation of $\phi_t^{(k)}$ and $\phi_v^{(k)}$ are assumed independent. The parameters are found numerically using maximum

likelihood estimation such that

$$\hat{\gamma}_t(T_{\text{int}}), \hat{\beta}_t(T_{\text{int}}) = \arg\max_{\gamma_t, \beta_t} \prod_{k=1}^{K} p_\Phi(\phi_t^{(k)}; \gamma_t, \beta_t, N)$$

and

$$\hat{\gamma}_v(T_{\text{int}}), \hat{\beta}_v(T_{\text{int}}) = \arg\max_{\gamma_v, \beta_v} \prod_{k=1}^{K} p_\Phi(\phi_v^{(k)}; \gamma_v, \beta_v, N),$$

where $T_{\text{int}} = N \cdot f_{s,\text{vad}}$. In order to provide simple models of $\gamma_t$, $\beta_t$, $\gamma_v$, and $\beta_v$, scatter plots of estimated shaping parameters for different $T_{\text{int}}$ are shown in Fig. 4. We choose to describe the shaping parameters using a power model. Let $\tilde{h}(T_{\text{int}}; a, b)$ be the general form of a power model with parameters $a$ and $b$:

$$\tilde{h}(T_{\text{int}}; a, b) = a \cdot T_{\text{int}}^{b}. \tag{13}$$

This model can be useful for implementation of the BMOG algorithm for any $T_{\text{int}}$, and to facilitate the theoretical performance evaluation of the MOG algorithm in Sec. IV-C. To estimate the parameters $a$ and $b$ of the power model, we use a non-linear least squares procedure with the general form of

$$\hat{a}, \hat{b} = \arg\min_{a,b} \sum_{j=1}^{J} \left[\tilde{h}(T_{\text{int}}^{(j)}; a, b) - \hat{h}(T_{\text{int}}^{(j)})\right]^2, \tag{14}$$

where $\hat{h}(T_{\text{int}}^{(j)})$ is an estimated shaping parameter, i.e., either $\hat{\gamma}_t(T_{\text{int}})$, $\hat{\beta}_t(T_{\text{int}})$, $\hat{\gamma}_v(T_{\text{int}})$, or $\hat{\beta}_v(T_{\text{int}})$, and $J$ is the total number of data points for each ML estimated shaping parameter. We minimize (14) numerically. The estimated power model parameters are summarized in Table 1. Fig. 4 shows that the fitted power models provide an excellent fit to the ML estimated shaping parameters as a function of $T_{\text{int}}$.

**TABLE 1.** Power model parameters for modeling the estimated shaping parameters of the beta-binomial distributions.

|  | $\tilde{\gamma}_t(\cdot; \hat{a}, \hat{b})$ | $\tilde{\beta}_t(\cdot; \hat{a}, \hat{b})$ | $\tilde{\gamma}_v(\cdot; \hat{a}, \hat{b})$ | $\tilde{\beta}_v(\cdot; \hat{a}, \hat{b})$ |
|---|---|---|---|---|
| $\hat{a}$ | 2.5091 | 0.8522 | 0.7736 | 0.8057 |
| $\hat{b}$ | 0.7879 | 0.7817 | 0.9727 | 0.9681 |

## C. THEORETICAL PERFORMANCE OF THE MOG ALGORITHM

In this section, we analyze the theoretical performance of the MOG algorithm and compare it with performance achieved through simulations. Two quantities that have a significant impact on the performance of the MOG algorithm, are the number of candidate speakers, I, and the integration time $T_{\text{int}}$. Increasing the number of candidate speakers will increase the solution search space, hence increase the a priori risk of choosing a wrong candidate as the target speaker. Decreasing the integration time $T_{\text{int}}$ will lead to higher variance in the estimation of the SSEs in (6).

The misclassification rate is used to measure the performance of the MOG algorithm and is defined as the
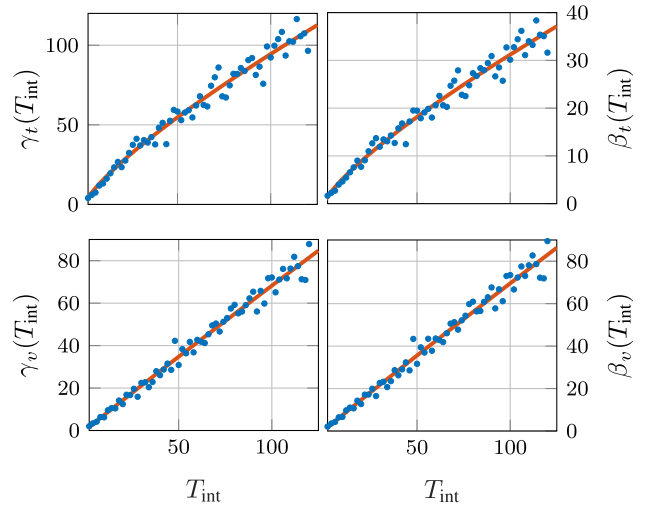


● MLE of shaping parameters —— Fitted power models

**FIGURE 4.** Shaping parameters of the beta-binomial distribution as a function of $T_{\text{int}}$. The blue data points are obtained from maximum likelihood estimation for different $T_{\text{int}}$. The red curves are fitted power models on the blue data points.

probability of classifying a competing speaker as the conversational partner. We denote the misclassification rate as $P(E = 1; I, T_{\text{int}})$ where $E \in \{0, 1\}$ is a Bernoulli random variable with $E = 1$ representing a misclassification. To derive an expression for the misclassification rate, we define $P(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1})$ as the probability of correct classification, where $\Phi_t$ denotes the SSE between the user and conversational partner, and $\Phi_{v,j}$ is the SSE between the user and the $j$'th competing speaker. The misclassification rate is then given by

$$P(E = 1; I, T_{\text{int}}) = 1 - P(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1}).$$

In Appendix C, we show that the misclassification rate of the MOG algorithm can be expressed as

$$P(E = 1; I, T_{\text{int}})$$
$$= 1 - \sum_{\phi=1}^{N} p_\Phi(\phi; \gamma_t, \beta_t, N) P_\Phi^{I-1}(\phi - 1; \gamma_v, \beta_v, N),$$
$$\tag{15}$$

where $P_\Phi(\phi - 1; \gamma_v, \beta_v, N)$ is the cumulative distribution of $p_\Phi(\phi - 1; \gamma_v, \beta_v, N)$ which is given by

$$P_\Phi(\phi - 1; \gamma_v, \beta_v, N) = \sum_{\kappa=0}^{\phi-1} p_\Phi(\kappa; \gamma_v, \beta_v, N). \tag{16}$$

For verification, we compare the theoretical misclassification rate given by (15) with the misclassification rate achieved with the MOG algorithm in simulations as seen in Fig. 5. From Fig. 5b, we clearly see a close match between the theoretical and simulated misclassification rates, where the conversational partners are speaking in Danish without any
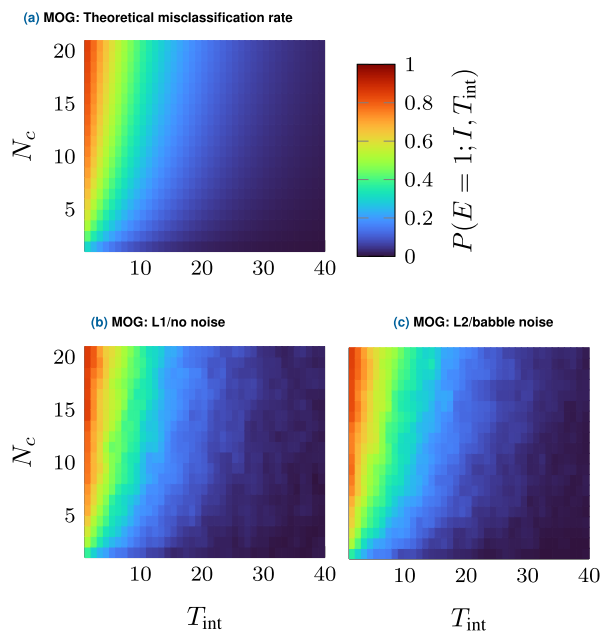
**FIGURE 5.** Probability of misclassifying the conversational partner as a function of the number of competing speakers $N_c$ and integration time $T_{\text{int}}$. The theoretical performance of the MOG algorithm is shown in Fig. 5a. The simulated MOG performance using the datasets L1/no noise and L2/babble noise is shown in Fig. 5b and Fig. 5c.

noise stimuli. Likewise, a close match between the theoretical and simulated misclassification rate can be seen in Fig. 5c, where the conversational partners are speaking in English (second language) with babble noise as noise stimuli. The close match indicates that the fitted statistical models are able to generalize to unseen conditions.

## V. EVALUATION IN SPEECH ENHANCEMENT APPLICATIONS

In this section, we demonstrate the use of MOG and BMOG for solving problem of enhancing a conversational partner in a multi-talker environment, using the speech enhancement paradigm of Fig. 1. In particular, we use MOG/BMOG to rank the candidate speakers according to how likely they are to be the conversational partner. In Secs. V-A and V-B, we outline the practical implementation of the MOG and BMOG algorithms and in Sec. V-C, we present the reference/baseline speaker ranking methods that will be used in our experiments. In Secs. V-D and V-E, we demonstrate the use of the proposed speech enhancement systems in two different applications for HADs.

### A. SPEECH ENHANCEMENT SYSTEM USING SPEAKER RANKING

Fig. 6 shows an example of the speech enhancement paradigm of Fig. 1 employing multiple microphones. In many situations, the microphone signals consist of a mixture of speech signals (including target and potential competing speakers) and noise from the environment. The unprocessed microphone signals are denoted as $x_m(n)$ for $m = 1, \ldots, M$,

where $M$ is the number of microphones and $n$ is the discrete-time index. Let $\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ be the noisy microphone signals stacked in a vector, which is processed by a speech separation system. The speech separation system separates the microphone signals into estimated speech signals $\hat{\boldsymbol{s}}(n) = [\hat{s}_0(n), \hat{s}_1(n), \ldots, \hat{s}_I(n)]^T$. Next, voice activity detection is applied to each of the separated signals, $\hat{s}_i(n), i = 1, \ldots, I$. A speaker ranking algorithm, e.g., MOG or BMOG, ranks the conversational partner by assigning a ranking score to each candidate speaker. Finally, in the example system in Fig. 6, the enhancement of the conversational partner is achieved simply as a linear combination of the separated speech signals $\hat{\boldsymbol{s}}(n)$. The weights are found using a gain function which maps the ranking score to a gain value for each separated speech signal. A straightforward gain function for the MOG algorithm, is to set the gain to a value of '1' to the estimated conversational partner channel, and a value of $0 < g_{\min} < 1$ for the remaining channels, i.e.,

$$g_j(n) = \begin{cases} 1, & \text{if } j = \hat{i} \\ g_{\min}, & \text{otherwise.} \end{cases} \qquad (17)$$

where $\hat{i}$ is the estimated channel of the conversational partner. It might occur that a competing speaker is estimated as being the conversational partner which can lead to severe loss in speech enhancement performance. It can also disrupt an ongoing conversational between a user and a conversational partner if the speaker ranking algorithm suddenly changes the estimated conversational partner. To increase the robustness, a minimum gain $g_{\min}$ can be applied such that a small amount of speech from all candidate speakers are always let through. Likewise, $g_{\min}$ can be made as a function of $n$, such that $g_{\min} = 1$ in the initial phase of a conversation, and gradually decreases towards a minimum value when the conversation has been established.

Another approach, specifically for the BMOG algorithm, is to use the estimated posterior probabilities as weights for the linear combination such that

$$g_j(n) = \max\left(g_{\min}, P(\mathcal{H}_i|\phi_1, \ldots, \phi_I)\right). \qquad (18)$$

A potential advantage of the posterior probability as a gain function is similar to that of introducing $g_{\min} > 0$ in (17): It reduces perceptual switching artifacts and limits the effect of target loss in case of misclassification. For both approaches, the estimated conversational partner signal is

$$\hat{s}_t(n) = \sum_{i=1}^{I} g_i(n)\hat{s}_i(n), \qquad (19)$$

where $\hat{s}_t(n)$ is the estimated speech signal of the conversational partner.

### B. IMPLEMENTATION OF THE MOG AND BMOG ALGORITHMS

In order to implement the MOG algorithm in (4), we estimate the MSE as the average square-error between $\alpha_0(n)$ and $\alpha_i(n)$
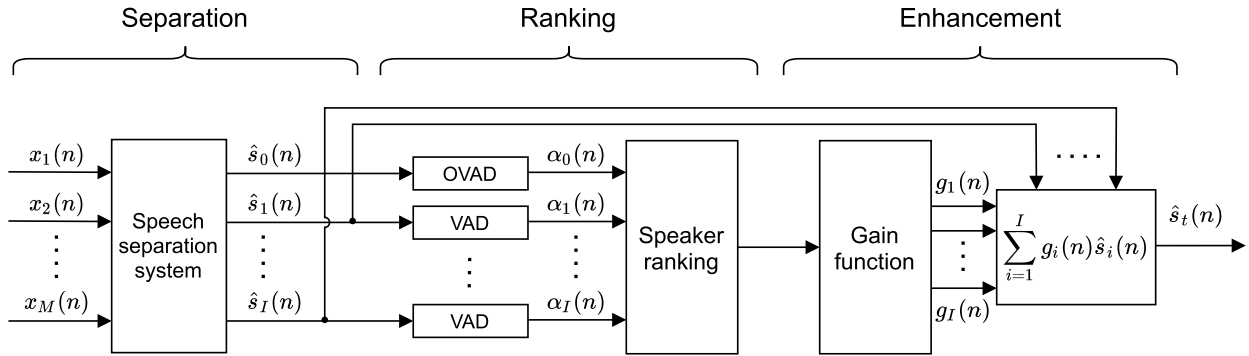
**FIGURE 6.** The proposed speech enhancement system consists of speech separation, speaker ranking, and enhancement. The output of the speaker ranking algorithm is for example, $\hat{i}_{\text{MOG}}(n)$ or $P(\mathcal{H}_i|\phi_1,\ldots,\phi_I)$.

over integration time $T_{\text{int}}$. The MOG estimate of the conversational partner index then becomes

$$\hat{i} = \arg\max_{i\in\{1,\ldots,I\}} \sum_{k=n-N+1}^{n} (\alpha_0(k) - \alpha_i(k))^2. \quad (20)$$

Implementation of the BMOG algorithm is a two-step procedure. First, the shaping parameters are computed for beta-binomial distributions given $T_{\text{int}}$ using (13) and TABLE 1, which may be done offline. Secondly, the posterior probabilities $P(\mathcal{H}_i|\phi_1,\ldots,\phi_I)$ are computed. To do so, the likelihood function in (11) is computed in the logarithmic domain for numerical stability. For this purpose, we first define the variable $\psi_i$ as:

$$\psi_i \triangleq P(\mathcal{H}_i)p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N) \prod_{j\in\mathcal{I}\setminus i} p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N). \quad (21)$$

The natural logarithm of $\psi_i$ is

$$\ln\psi_i = \ln P(\mathcal{H}_i) + \ln p_{\Phi_i}(\phi_i;\gamma_t,\beta_t,N)$$
$$+ \sum_{j\in\mathcal{I}\setminus i} \ln p_{\Phi_j}(\phi_j;\gamma_v,\beta_v,N). \quad (22)$$

Substituting (21) into (12) gives

$$P(\mathcal{H}_i|\phi_1,\ldots,\phi_I) = \frac{\psi_i}{\sum_{k=1}^{I}\psi_k}. \quad (23)$$

Using the logarithm function on both sides yields

$$\ln P(\mathcal{H}_i|\phi_1,\ldots,\phi_I) = \ln\psi_i - \ln\left(\sum_{k=1}^{I}\psi_k\right), \quad (24)$$

where

$$\ln\left(\sum_{k=1}^{I}\psi_k\right) = \ln\psi_1 + \ln\left(1+\sum_{j=2}^{I}e^{(\ln\psi_j-\ln\psi_1)}\right). \quad (25)$$

The posterior probability can be found by inserting (21), (22), and (25) into (24) and applying the exponential function $\exp(\cdot)$ to (24). The implementation of the BMOG algorithm is summarized in Algorithm 1.

---

**Algorithm 1** Implementation of the BMOG algorithm.

**Input:** $\alpha_i(n)$ for $i = 0, 1, \ldots, I$. Set the parameters $T_{\text{int}}, N = \left\lfloor \frac{T_{\text{int}}}{f_{\text{s,vad}}} \right\rfloor$, and $P(\mathcal{H}_i)\forall i$.

1: Compute the shaping parameters $\hat{\gamma}_t(T_{\text{int}})$, $\hat{\beta}_t(T_{\text{int}})$, $\hat{\gamma}_v(T_{\text{int}})$, $\hat{\beta}_v(T_{\text{int}})$ using (13) and TABLE 1.
2: Compute the SSEs using (6) to obtain $\phi_i(n)$ for all $i$.
3: Compute the log-likelihoods

$$\ln p_{\Phi_i}(\phi_i(n); \hat{\gamma}_t(T_{\text{int}}), \hat{\beta}_t(T_{\text{int}}), N),$$
$$\ln p_{\Phi_i}(\phi_i(n); \hat{\gamma}_v(T_{\text{int}}), \hat{\beta}_v(T_{\text{int}}), N),$$

   for all $i$ using (8).
4: Compute $\ln\psi_i(n)$ in (22) for all $i$.
5: Compute the log posterior probabilities from (24).
6: Use the exponential function $\exp(\cdot)$ on (24) to obtain the posterior probability in (23).

---

### C. STATE-OF-THE-ART METHODS FOR SPEAKER RANKING

The idea of using turn-taking to detect conversations between two speakers has been explored in [31]–[33] but was not used in the context of enhancing a conversational partner of a user as presented in Fig. 1. In [31], the presence of a conversation between two speakers was quantified using mutual information between the user's and candidate speakers' voice activity sequences. The normalized cross-correlation function was later proposed as a quantifier of conversations in [32]. Both methods can be compared to the MOG/BMOG algorithms in a fair manner, since all methods require access to VAD sequences for each speaker and they return a cost that can be used for ranking the candidate speakers.

#### 1) MAXIMUM MUTUAL information [31]

The mutual information method is based on finding the candidate speaker that maximizes the mutual information between the user's and candidate speaker's voice activity sequences

$$\hat{i}_{\text{MMI}} = \arg\max_{i\in\{1,\ldots,I\}} \sum_{k=0}^{1}\sum_{j=0}^{1} P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = j)$$
$$\times \log\frac{P_{A_0A_i}(\alpha_0(n) = k, \alpha_i(n) = j)}{P_{A_0}(\alpha_0(n) = k)P_{A_i}(\alpha_i(n) = k)},$$

where all joint and marginal probabilities are sample estimates obtained from $\alpha_i(n)$ over integration time $T_{\text{int}}$. One problem with the MMI algorithm is situations where the numerator or denominator of the logarithmic function becomes zero. These situations might occur if the integration time is short, e.g., 2 seconds, as there is a risk that the user or candidate speaker $i$ might be silent within the period of time. In the evaluations, we removed results where the numerator or denominator of the MMI algorithm becomes zero.

### 2) NORMALIZED CROSS-correlation [32]

Similarly, the normalized cross-correlation (NCC) method is here used to detect the presence of a conversational partner. The optimization problem of NCC is formulated as

$$\hat{i}_{\text{NCC}} = \arg \max_{i \in \{1,...,I\}} \frac{1 - \min_{p \in [r1,r2]} R_{0,i}(p)}{2}, \qquad (26)$$

where $R_{0,i}(p)$ is the normalized cross-correlation between $A_0$ and $A_i$ at lag $p$. $r1$ and $r2$ are search region bounds for the lag $p$. We set $p$ equal to zero in our evaluation.

### 3) SPEAKER RANKING PERFORMANCE

We examine the speaker ranking performance between the proposed MOG algorithm against MMI and NCC. The performance is reported in terms of misclassification rate as a function of the number of competing speakers $N_c$ and integration time $T_{\text{int}}$. We use speech signals from [27] for the performance evaluation. Specifically, we use the subset of the data set containing 2-person conversations in second language English (L2) in babble noise. The speech signals are segmented into segments of length $T_{\text{int}}$. For each $T_{\text{int}}$, one 2-person conversation is randomly selected to constitute the user's own voice and the user's conversational partner. A number of $N_c$ arbitrarily chosen speakers from the data set are selected to constitute the competing speakers. Fig. 7 shows the misclassification rate $P(E = 1; I, T_{\text{int}})$ as a function of $T_{\text{int}}$ and the number of competing speakers $N_c = I - 1$ for each ranking algorithm. A comparison between MOG, MMI, and NCC shows that the misclassification rate is significantly lower for the MOG algorithm compared to the MMI and NCC, particularly, when 1) the integration time is short, and/or 2) there is a large number of competing speakers. At long integration times, e.g. 40 sec, the difference between the algorithms is smaller. However, the MOG algorithm consistently performs better than the MMI and NCC algorithms.

### D. APPLICATION 1: WIRELESS HEARING AID NETWORK

In this section, we demonstrate the use of the proposed (B)MOG based speech enhancement paradigm, cf. Fig. 1 in a hearing aid (HA) application, in which the HAs of several users are wirelessly connected. The basic idea is that multiple HA users can distribute their own voice signal to the other users' HAs through a wireless network. This can be useful,
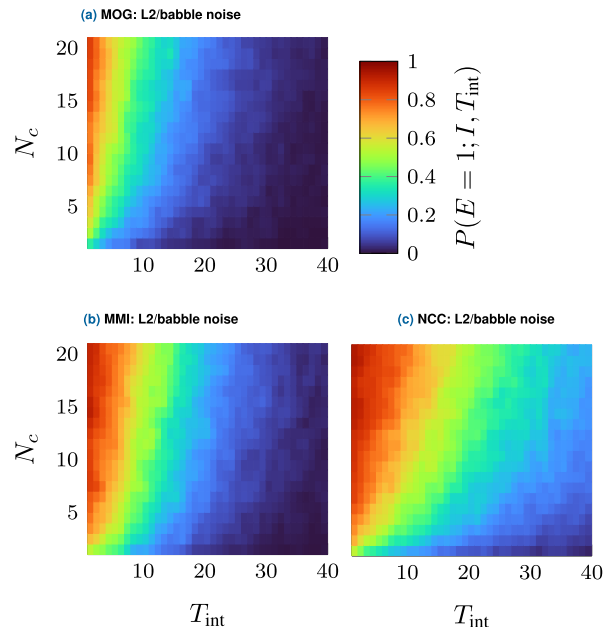


**FIGURE 7.** Misclassification rate as a function of the number of competing speakers $N_c$ and integration time $T_{\text{int}}$ for MOG (proposed), MMI, and NCC. The database used for this evaluation is from [27] with the subset where all speakers were talking in their second language (L2) in babble noise.

e.g., in acoustically challenging social gatherings with multiple HA users. The proposed speech enhancement paradigm can in this situation assist the HA user by first ranking and then enhancing the estimated conversational partner amongst the users. [1]

The signal model of the sound picked-up by the user's HA microphone can be described as

$$x_i(n) = s_i(n), \quad i = 0, \dots, I, \qquad (27)$$

where $s_0(n)$ is the HA user's own voice signal as picked-up by the user's microphone while $s_i(n)$ for $i = 1, \dots, I$ are the clean speech signals picked up by the microphones located at the candidate speakers.

### 1) SIMULATION SETUP

We reuse the speech database presented in Sec. V-C3 for the candidate speakers and own voice signals. We use the data set with conversations in second language and babble noise, which was not used for estimation of the shaping parameters. Two conversational partners are randomly chosen from the data set, where one is randomly chosen as the HA user and the other as the conversational partner for each signal realization. The competing speakers are chosen from the same data set, but are not conversing with the HA user. The HA user's conversational partner is unknown to the speech enhancement systems. We use rVAD 2.0 [29], [30] for voice

---

[1]In this situation, the separation stage in Figs. 1 and 6 is obsolete – microphones are located on each candidate speaker and allow direct estimation of their voice activity pattern $\alpha_i(n)$; hence, the separation stage is unnecessary.
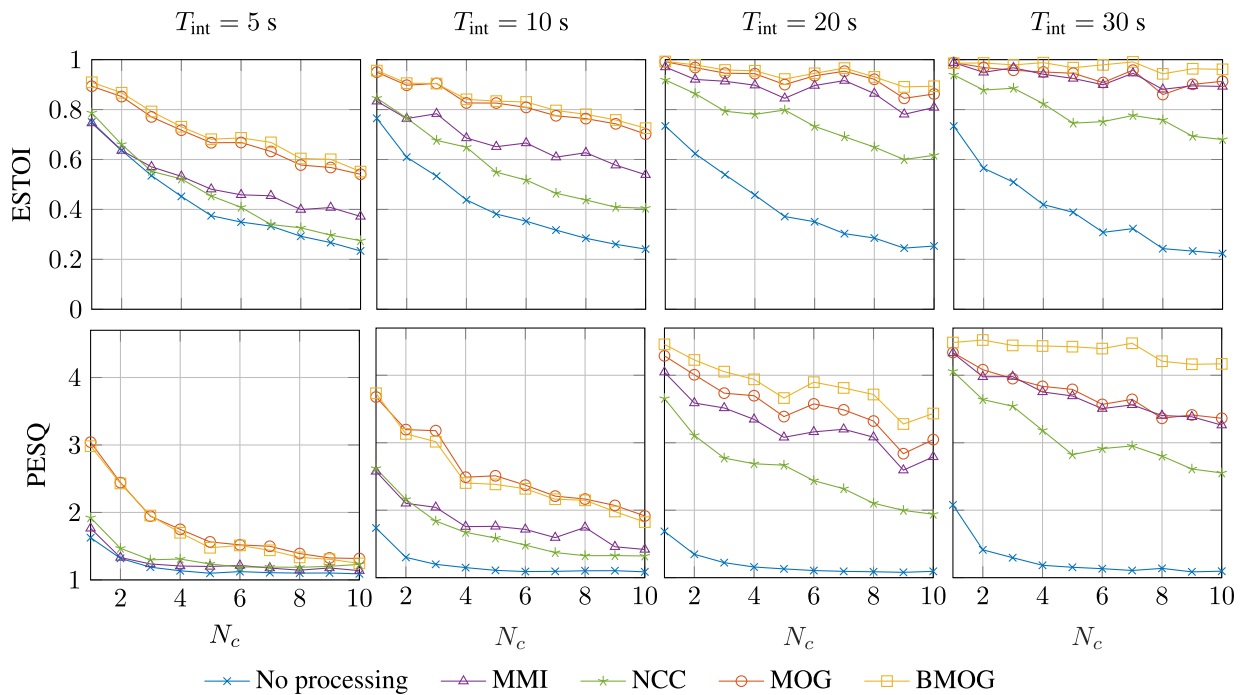
**FIGURE 8.** Averaged ESTOI and PESQ scores as a function of the number of competing speakers $N_c$ and integration time $T_{int}$.

activity detection and the sampling frequency of the VAD output is $f_{s,vad} = 100$ Hz. The integration time needed for the speaker ranking algorithms, is implemented as sliding windows with length $T_{int}$ and with a hop size of 1 sample at sampling frequency $f_{s,vad}$. The speech enhancement systems used in the evaluation are referred to as:

- **No processing**: The speech enhancement system does not apply any speaker ranking algorithms and simply outputs the sum of all candidate speakers.
- **MMI, NCC**, and **MOG**: The MMI, NCC, and the proposed MOG algorithms are used as speaker ranking. The gain function is implemented as in (17).
- **BMOG**: Posterior probabilities of the conversational partner are estimated using BMOG and used as a gain function for enhancement, cf. (19). The prior probability distribution $P(\mathcal{H}_i)$ in (12) is set uniform.

### 2) RESULTS: WIRELESS HEARING AID NETWORK

We evaluate the speech enhancement performance by comparing the enhanced conversational partner with the clean speech signal of the conversational partner in terms of ESTOI [34], [35], PESQ [36], and segmental SNR [37]. We evaluate the speech enhancement systems for $T_{int} = \{5, 10, 20, 30\}$ s and the following number of competing speakers $N_c = \{1, 2, \ldots, 10\}$ [s]. The minimum gain for MMI, NCC, and MOG is set to $g_{min} = 0.01$. A minimum gain of $g_{min} > 0$ is necessary for the MMI, NCC, and MOG enhancement systems to avoid rare situations with a complete suppression of the conversational partner. These situations typically arise at low $T_{int}$ and results in undefined PESQ and

segmental SNR scores. The minimum gain for BMOG was set to $g_{min} = 0$ as it did not experience similar problems. The results are shown in Fig. 8 and each score is averaged over 100 realizations of conversations. Generally, we see a significant improvement in terms of both ESTOI and PESQ when using MOG and BMOG compared to NCC and MMI. The improvement is particularly notably at low integration time such as $T_{int} = 5$ s and $T_{int} = 10$ s. At higher integration times, the improvements become less prominent with the exception of NCC, which seems to perform the worst. We note that BMOG seems to perform much better than MOG in terms of PESQ at $T_{int} = 30$. This is due to the minimum gain which is set to 0.01 for MOG but 0 for BMOG. From our experiments, we have observed that setting the minimum gain to be above 0 can help NCC, MMI, and MOG perform better on average at low integration times, e.g., $T_{int} = 5$. However, the trade-off is slightly degraded performance at high integration times as shown in the results.

From these results, it is clear that speech enhancement systems that use MOG and BMOG generally outperform the NCC and MMI methods for this particular application.

### E. APPLICATION 2: BEAMFORMING SYSTEM IN HEARING AIDS

In this section, we demonstrate the use of the proposed speech enhancement paradigm in another hearing aid application. Modern hearing aids are equipped with multiple microphones which allow for implementation of acoustic beamformers to enhance the speech signal of a conversational partner of a HA user. However, retrieving the speech signal can be

particularly difficult in situations with multiple competing speakers, because it is hard to decide who is the conversational partner. Hence, in this application the proposed (B)MOG speech enhancement paradigm is used to efficiently retrieve the speech signal of the conversational partner amongst several competing speakers.

First, we model the received signal at the microphones of the HAs. The user's and candidate speakers' speech signals propagate to the microphones and are simulated using acoustic impulse responses (AIRs). The AIR from the $i$'th speaker to the $m$'th microphone is denoted as $h_{i,m}(n)$ where $i = 0, 1, \ldots, I$ is the speaker index, and $m = 1, \ldots, M$ is the microphone index. The index value $i = 0$ is used to denote the user's index. The AIRs can be decomposed into $h_{i,m}(n) = h_{i,m'}(n) * d_{i,m}(n)$ where $*$ denotes the convolution operator, $h_{i,m'}(n)$ is the AIR from the $i$'th speaker to a preselected reference microphone $m' \in \{1, \ldots, M\}$, and $d_{i,m}(n)$ is the impulse response from the reference microphone to the $m$'th microphone also referred to as the relative impulse response. Let $s_i'(n)$ be the received signal of the $i$'th speaker at the reference microphone, $m'$, i.e. $s_i'(n) = s_i(n) * h_{i,m'}(n)$. Then the received signal of the $i$'th speaker at the $m$'th microphone is $s_{i,m}'(n) = s_i'(n) * d_{i,m}(n)$. We denote $v_m(n)$ as being the noise vector (e.g. ambient noise and microphone self-noise) as received at the $m$'th microphone. The noisy signal at the $m$'th microphone is then modeled as

$$x_m(n) = \sum_{i=0}^{I} s_{i,m}'(n) + v_m(n). \tag{28}$$

### 1) SPEECH SEPARATION USING BEAMFORMERS

The received microphone signal, $x_m(n)$, is a mixture of clean user and candidate speaker signals received at microphone $m$, $s_{i,m}'(n)$, plus noise $v_m(n)$. Following the speech enhancement paradigm in Fig. 6, the microphone signals are first separated into user and candidate speaker signals before applying speaker ranking. We use the minimum power distortionless response (MPDR) beamformer to separate the speech signals. The MPDR beamformers are implemented in the time-frequency domain using the short-time Fourier transform (STFT) and are for each time-frequency tile computed as [38]

$$\boldsymbol{W}_i(k, l) = \frac{C_x^{-1}(k, l)\boldsymbol{D}_i(k)}{\boldsymbol{D}_i^H(k)C_x^{-1}(k, l)\boldsymbol{D}_i(k)}, \quad \boldsymbol{W}_i(k, l) \in \mathbb{C}^M, \tag{29}$$

where $k$ and $l$ denote the frequency and frame indices, respectively. $C_x(k, l) = \mathbb{E}\{X(k, l)X^H(k, l)\}$ is the cross power spectral density (CPSD) matrix of the noisy microphone signals and $\boldsymbol{D}_i(k) = [D_{i,1}(k), \ldots, D_{i,M}(k)]^T$, $i = 0, 1, \ldots, I$, $k = 0, 1, \ldots, K$ denotes the relative acoustic transfer function (RATF) vector for the $i$'th speaker and $k$'th frequency bin [39], [40]. The $m$'th element of the RATF vector is the frequency domain representation of $d_{i,m}(n)$ under the assumption that the STFT window length is longer than the relative impulse response. Unfortunately, the number of candidate speakers and their RATF vectors are seldomly

known in practice. Instead, we use $I$ to denote the number of MPDR beamformers steered towards a set of $I$ unique and fixed directions in the acoustic environment. In other words, the spatial filter bank is implemented using a dictionary of RATF vectors $\mathcal{D}(k) = \{\boldsymbol{D}_0(k), \boldsymbol{D}_1(k), \ldots, \boldsymbol{D}_I(k)\}$, $k = 0, 1, \ldots, K$, where we assume that the dictionary is given in advance. Assuming that each beam contains a maximum of one candidate speaker (i.e., that candidate sources are sufficiently spatially separated), each beamformer output, $\hat{s}_i'(n)$, is treated as a candidate speaker signal. The output of each beamformer is

$$\hat{S}_i'(k, l) = \boldsymbol{W}_i^H(k, l)\boldsymbol{X}(k, l), \tag{30}$$

where $\hat{S}_i'(k, l)$ is the enhanced signal from direction $i$, and is treated as a speech signal from a candidate speaker. The beamformer outputs $\hat{S}_i'(k, l)$ are transformed back to the time-domain using the inverse STFT to obtain $\hat{s}_i'(n)$. The remaining part of the speech enhancement system, i.e., ranking and enhancement, follows the same procedure as in application 1 in Sec. V-D.

### 2) SIMULATION OF THE ACOUSTIC SCENE

We simulate the acoustic scenes to resemble a cocktail party-like scenario with a HA user engaged in a conversation with a conversational partner. Such a situation involves the presence of speech signals from the HA user, the conversational partner, and competing speakers, and the presence of noise from the environment.

To simulate the received signals at the microphones, we use a database of AIRs measured in a sound studio where room reverberation has been removed [41]. The measurement setup consists of a spherical loudspeaker array with a HA user seated in the center of the array. The HA user is wearing a behind-the-ear (BTE) hearing aid on each ear. Each BTE hearing aid has three microphones where two are placed in a front/rear configuration on the HA and the third is placed in the ear canal. The microphones are used in a binaural HA configuration where we assume wireless, simultaneous, and error-free signal exchange between the left and right HAs. Hence, beamformers are implemented using a the total number of $M = 6$ microphones. The AIRs are measured from uniformly spaced positions in the horizontal plane with respect to the head of the HA user and with a resolution of 7.5° resulting in AIRs for 48 different angles. We define 0° as the frontal direction from the user's point-of-view. The own voice AIRs are measured using a mouth reference microphone placed in front of the HA user's mouth.

We use the conversational speech database in [27], as in Application 1, as speech material in our simulation. Realistic noise measured in a canteen is used in our simulation. The noise is measured using a spherical microphone array to accurately capture the noise field [42]. The noise recordings are transformed and convolved with the AIRs to reproduce the same noise field as would have been experienced by a HA user in the canteen.

Competing speakers are added to the acoustic scenes. The speech material for the competing speaker are from the same speech database as in Sec. V-D [27]. The speech of the competing speakers is unrelated to the conversation between the user and conversational partner. We experiment with $N_c = 3$ and $N_c = 5$ competing speakers in our evaluation. Increasing the number of competing speaker to much larger than $N_c = 5$, results in poorer speech separation as the beamformers cannot sufficiently suppress the speakers from other directions. The purpose here is mainly to demonstrate the feasibility of using (B)MOG ranking in a beamforming context and for a larger number of competing speakers, other better performing speech separation systems could be used, e.g., (Conv-)TasNet [12], [13] or Wavesplit [43].

For $N_c = 3$, the conversational partner is placed randomly in the positions $\{0°, 90°, 180°, 270°\}$, and the competing speakers are placed at the remaining 3 positions. Similarly for $N_c = 5$, the conversational partner is placed randomly in the positions $\{0°, 60°, 120°, 180°, 240°, 300°\}$, and the competing speakers are placed at the remaining 5 positions. The positions of the speakers are fixed for the whole duration of a realization of an acoustic scene. We do not simulate head-movements of the HA user but these movement can be compensated with other sensors, e.g., accelerometers in practice. To simulate the received signals of the speech sources at the microphones, we convolve the speech signals with the AIRs associated with the direction. The speech power of each competing speaker is approximately identical to the speech power of the conversational partner before convolving with the AIRs. Canteen noise is added to the acoustic scenes and the SNR is defined as the ratio between the clean speech power of the conversational partner at the source location and the power of the background noise. The SNR is set to 12 dB.

The search region of the beamformers, $\boldsymbol{W}_i(k, l)$, is $0°, 7.5°, \ldots, 352.5°$ in the azimuth angle. The RATF dictionary is given as $\mathcal{D}(k) = \{\boldsymbol{D}_{0,m}(k), \boldsymbol{D}_{1,m}(k), \ldots, \boldsymbol{D}_{I,m}(k)\}$ where $\boldsymbol{D}_{0,m}(n)$ is the own voice RATF vector. The elements $\boldsymbol{D}_{i,m}(n)$ $i = 1, \ldots, I$ are RATF vectors associated with sound sources impinging from direction $\theta = (i - 1) \cdot 7.5°$ in the horizontal plane where $\theta = 0°$ is the frontal direction with respect to the HA user.

To implement the OVAD/VAD blocks in Fig. 6, rVAD 2.0 [30] is used for voice activity detection on the separated speech signals $\hat{s}_i(n)$ and the own voice signal $\hat{s}_0(n)$.

The sampling frequency of the received microphone signals is set to 16 kHz. We use a square-root Hann window with a window size of 256 samples for the STFT and inverse STFT. The hop-size is 128 samples.

The beamforming system is summarized in Algorithm 2 in pseudo-code.

### 3) EVALUATION OF THE SPEECH ENHANCEMENT PARADIGM IN BEAMFORMING SYSTEMS

We evaluate the performance in terms of 1) speaker ranking performance in Sec. V-E4 and 2) speech enhancement performance in Sec. V-E5. First, the speaker ranking in this

---

**Algorithm 2** Beamforming system for application 2.

**Input:** $\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^T$, $\mathcal{D}_k = \{\boldsymbol{D}_1(k), \ldots, \boldsymbol{D}_I(k)\}$

1: Apply STFT to $\boldsymbol{x}(n)$ to obtain $\boldsymbol{X}(k, l)$ for all $k$ and $l$.
2: **for** all $i \in \Theta$ **do**
3:     **for** $k = 1, 2, \ldots, K$ **do**
4:         Estimate the noisy CPSD matrix:

$$\hat{C}_x(k, l) = \frac{1}{L}\underline{\boldsymbol{X}}(k, l)\underline{\boldsymbol{X}}^H(k, l)$$

        where $\underline{\boldsymbol{X}}(k, j) = [\boldsymbol{X}(k, l - L + 1), \ldots, \boldsymbol{X}(k, l)]$.
5:         Compute the MPDR beamformer weights, $\boldsymbol{W}_i(k, l)$, using (29).
6:         Enhance the signal from direction $i$ using (30).
7:     **end for**
8:     Inverse STFT $\hat{S}_i'(k, l)$ to obtain $\hat{s}_i'(n)$.
9: **end for**
10: Estimate voice activity of each candidate speaker $\alpha_i(n) = $ VAD $\left(\hat{s}_i'(n)\right)$.
11: Use a speaker ranking algorithm e.g. Algorithm 1 and compute the gain function to obtain $g_i(n)$.
12: Enhance the conversational partner using (19).

---

application is closely related to direction-of-arrival (DOA) estimation. DOA estimation often arises in beamforming applications where the goal is to estimate the direction of the talker-of-interest in order to steer a beamformer. In our context, DOA estimation is related to estimating the channel of the conversational partner. Hence, the MOG algorithm is in fact a DOA estimator in this context. Secondly, the speech enhancement performance will quantify the potential benefit of using the proposed speech enhancement paradigm in a beamforming context for HAs. The reported performance scores are averaged from simulations of 40 realizations of the acoustic scenes for the results in Sec. V-E4 and Sec. V-E5.

**TABLE 2.** DOA estimation accuracy as a function of integration time $T_{\text{int}}$ and number of competing speakers $N_c$.

| MOG | $T_{\text{int}} = 5$ | $T_{\text{int}} = 10$ | $T_{\text{int}} = 20$ | $T_{\text{int}} = 30$ |
|---|---|---|---|---|
| $N_c = 3$ | 31.60% | 46.23% | 65.94% | 75.03% |
| $N_c = 5$ | 35.54% | 48.23% | 63.83% | 70.97% |
| **MMI** | | | | |
| $N_c = 3$ | 16.91% | 32.69% | 51.54% | 61.26% |
| $N_c = 5$ | 20.00% | 31.77% | 49.66% | 58.91% |
| **SRP-PHAT** | | | | |
| $N_c = 3$ | 19.72% | 19.72% | 19.72% | 19.72% |
| $N_c = 5$ | 23.39% | 23.39% | 23.39% | 23.39% |

To evaluate the speaker ranking performance, we evaluate the DOA accuracy and the mean-absolute-error (MAE) between the estimated DOA $\hat{\theta}_n$ and the true DOA $\theta_n$ of the conversational partner. The DOA accuracy is the probability of estimating the correct DOA of the conversational partner

**TABLE 3.** Mean-absolute-error of estimated DOA as a function of integration time $T_{\text{int}}$ and number of competing speakers $N_c$.

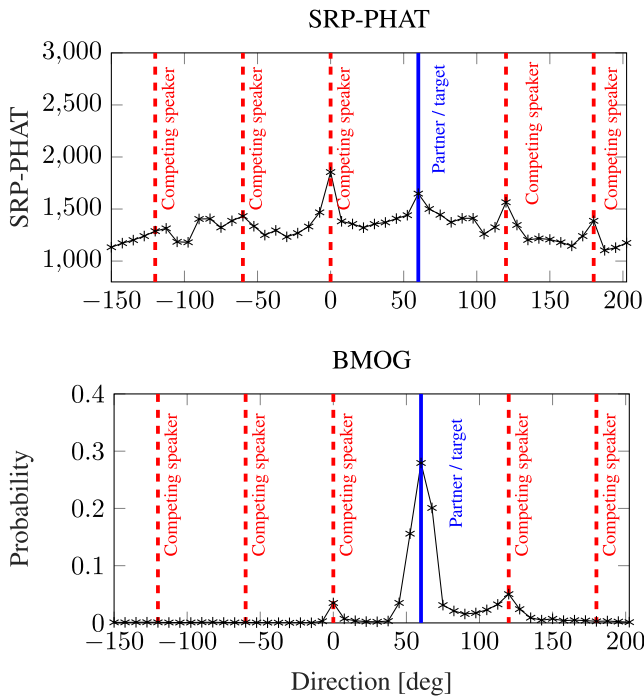| MOG | $T_{\text{int}} = 5$ | $T_{\text{int}} = 10$ | $T_{\text{int}} = 20$ | $T_{\text{int}} = 30$ |
|---|---|---|---|---|
| $N_c = 3$ | 60.55° | 45.41° | 29.81° | 22.30° |
| $N_c = 5$ | 58.58° | 46.67° | 30.81° | 22.39° |
| **MMI** | | | | |
| $N_c = 3$ | 79.68° | 60.35° | 42.27° | 33.77° |
| $N_c = 5$ | 72.90° | 59.71° | 43.16° | 33.17° |
| **SRP-PHAT** | | | | |
| $N_c = 3$ | 79.62° | 79.62° | 79.62° | 79.62° |
| $N_c = 5$ | 77.25° | 77.25° | 77.25° | 77.25° |



**FIGURE 9.** Example of the average output over 24 seconds of the SRP-PHAT algorithm and BMOG algorithm as a function of direction. The acoustic scene consist of 1 conversational partner and 5 competing speakers in canteen noise. The SRP-PHAT algorithm is not able to distinguish between the conversational partner and competing speakers, however, the proposed BMOG algorithm ($T_{\text{int}} = 10$ s) is however effective at locating the conversational partner.

and the MAE is estimated as the average absolute error:

$$\hat{\text{MAE}} = \frac{1}{N} \sum_{n=1}^{N} |\arg\left(\exp(j(\theta_n - \hat{\theta}_n))\right)|, \qquad (31)$$

where $\theta_n$ and $\hat{\theta}_n$ are in radians, $\sqrt{j} = -1$, and $\arg(\cdot)$ is the argument of a complex number. The $\hat{\text{MAE}}$ is averaged over The speech enhancement performance is reported in terms of ESTOI, PESQ, and segmental SNR scores to estimate the speech intelligibility, speech quality, and noise suppression performance of the proposed speech enhancement paradigm, respectively. The ESTOI, PESQ, and segmental SNR scores are computed using the output of the

enhancement system $\hat{s}_t(n)$ and the clean conversational partner speech signal received at the reference microphone index $s'_{t,m'}(n)$.

Our evaluation includes four beamforming systems which are based on the speech enhancement paradigm in Fig. 6. All systems use the same spatial filter bank of MPDR beamformers for speech separation. We use the rVAD 2.0 for voice activity detection for all systems. We refer the beamforming systems to as:

- **SE-Oracle**: The beamforming system, SE-Oracle, is used as a reference system to indicate the upper bound performance if the direction of the conversational partner is known in advance.
- **SE-MMI**: The beamforming system, SE-MMI, uses the MMI algorithm to find the direction of the conversational partner. The output at time $n$ of SE-MMI is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{MMI}}(n)}(n)$ where $\hat{i}_{\text{MMI}}(n)$ is the DOA estimate of the conversational partner at time $n$.
- **SE-MOG**: The beamforming system, SE-MOG, uses the MOG algorithm to find the direction of the conversational partner. The output at time $n$ of SE-MOG is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{MOG}}(n)}(n)$ where $\hat{i}_{\text{MOG}}(n)$ is the DOA estimate of the conversational partner at time $n$.
- **SE-SRP-PHAT**: The beamforming system, SE-SRP-PHAT, uses the well-known SRP-PHAT algorithm [6] to estimate the DOA of the conversational partner. In contrast to the speaker ranking algorithms NCC, MMI, and MOG, the SRP-PHAT algorithm does not utilize turn-taking to the candidate speakers related to conversations but instead searches for the most dominant speaker. The output at time $n$ of SE-SRP-PHAT is $\hat{s}_t(n) = \hat{s}'_{\hat{i}_{\text{SRP-PHAT}}(n)}(n)$ where $\hat{i}_{\text{SRP-PHAT}}(n)$ is the DOA estimate of the conversational partner at time $n$.
- **SE-BMOG**: The beamforming system, SE-BMOG, uses the BMOG algorithm to compute a posterior probability distribution of the direction of the conversational partner. The output of SE-BMOG at time $n$ is a linear combination of the separated candidate speakers using the posterior probabilities as weights, i.e., $\hat{s}_t(n) = \sum_{i=1}^{I} P(\mathcal{H}_i | \phi_1, \ldots, \phi_I) \hat{s}'_i(n)$. The prior probability distribution for BMOG was set to be a uniform prior probability distribution.

We did not include a beamforming system with a NCC-based speaker ranking algorithm as it performed significantly poorer than the other algorithms in preliminary experiments.

### 4) RESULTS: DOA ESTIMATION PERFORMANCE IN BEAMFORMING SYSTEMS

This section focuses on speaker ranking/DOA performance and not speech enhancement performance of the complete beamforming system, which is treated in Sec. V-E5. Therefore, BMOG is not included since the output of BMOG is a probability distribution and not an estimate of the conversational partner as the MOG algorithm.
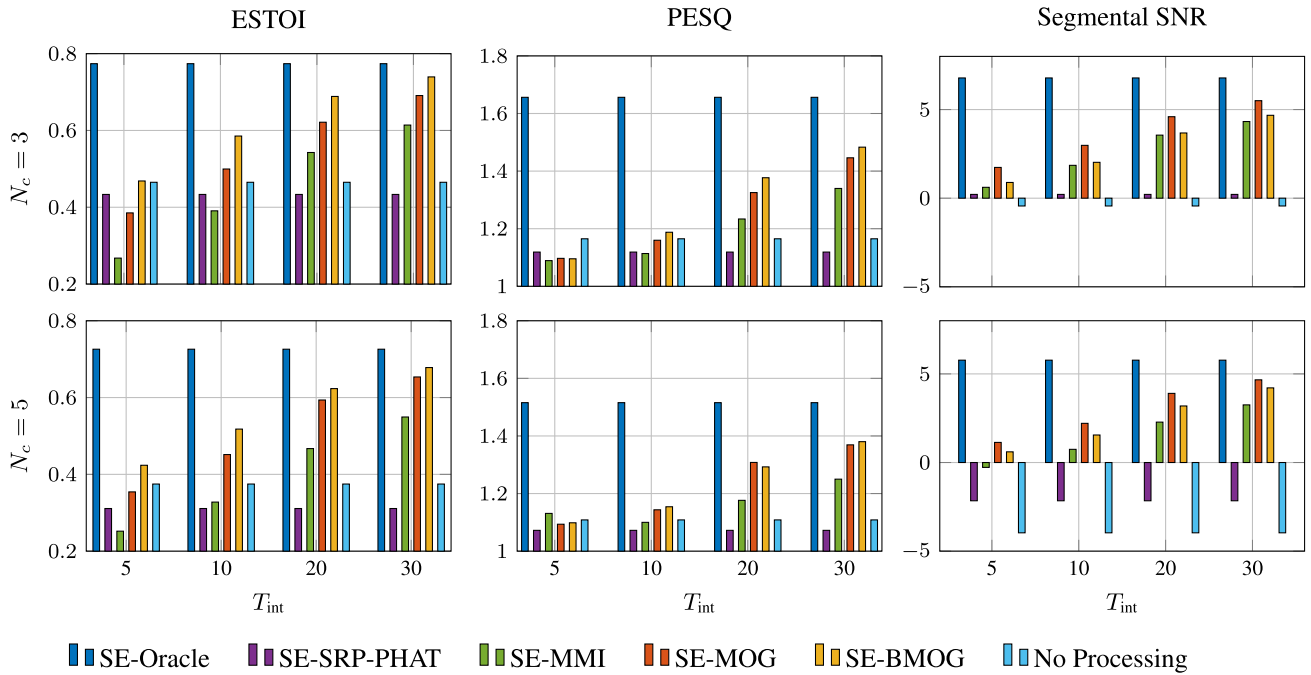
**FIGURE 10.** Averaged beamforming performance as function of the number of competing speakers $N_c$ and integration time $T_{int}$.

The results for DOA estimation performance in terms of DOA accuracy and MAE are shown in table 2 and 3, respectively. Each score in the table is an average over 40 realizations of the acoustic scenes.

The MOG algorithm seems to outperform the MMI algorithm consistently by approximately 15%-points. Similarly, the MAE for the MOG algorithm is lower than the MAE for MMI and SRP-PHAT for all $T_{int}$ and $N_c$. It is also clear, that the SRP-PHAT algorithm in general struggles in estimating the conversational partner DOA in a multi-speaker situation which is demonstrated in Fig. 9. Essentially, the SRP-PHAT algorithm constantly switches between the candidate speakers as the estimate of the conversational partner. The MOG algorithm, however, effectively exploits the turn-taking mechanism in conversations and is able to detect the conversational partner.

An interesting observation is that the DOA estimation accuracy is slightly higher for $N_c = 5$ compared to $N_c = 3$ at low integration times, e.g., $T_{int} = 5$ s. Likewise, the MAE is lower for $N_c = 5$ compared to $N_c = 3$ at low integration times. However, note that the angular distance between the conversational partner and competing speakers becomes larger at $N_c = 3$ compared to $N_c = 5$. That is, for $N_c = 3$ the speakers are located at $\{0°, 90°, 180°, 270°\}$ whereas for $N_c = 5$ the speakers are located at $\{0°, 60°, 120°, 180°, 240°, 300°\}$. Therefore, possible explanations of these observations at low integration times are that 1) the DOA estimates of MMI and MOG become more biased for $N_c = 3$, which results in a lower accuracy and 2) MMI and MOG are more likely to return a

higher absolute error for $N_c = 3$ than for $N_c = 5$ in case of a DOA estimation error. However, it is evident from the results, that the MOG algorithm has significantly higher accuracy compared to MMI and SRP-PHAT for all combinations of $T_{int}$ and $N_c$.

### 5) RESULTS: SPEECH ENHANCEMENT PERFORMANCE IN BEAMFORMING SYSTEMS

The results for beamforming performance are shown in Fig. 10, which plots performance scores ESTOI, PESQ, and segmental-SNR as a function of integration time $T_{int}$ for different beamforming systems. Cleacly, the MOG algorithm outperforms the MMI and SRP-PHAT algorithms significantly in most situations. The results also indicate that the SRP-PHAT algorithm performs slightly worse than the unprocessed signal in multi-speaker environments unless additional knowledge on the conversational partner is given. The MMI algorithm also performs slightly worse than the unprocessed signal in terms of ESTOI at $T_{int} = 5$ s and $T_{int} = 10$ s as the MMI algorithm can erroneously estimate a competing speaker as being the conversational partner for low integration times. The speech enhancement system using the BMOG algorithm, however, performs best on average across all scores, especially in terms of ESTOI and PESQ. This is likely due to a softer gain function based on the estimated posterior probability, which is less aggressive compared to the gain function used in the MOG algorithm. The softer gain function translates to higher ESTOI and PESQ scores, but a slightly lower segmental SNR score. With long integration times, both speech enhancement systems using MOG and

BMOG are extremely effective at retrieving a conversational partner in a multi-speaker situation as they perform close to the oracle beamformer. However, long integration times also require that the conversational partner stays within the same beam for longer duration, e.g. in a restaurant where the speakers are seated.

## VI. CONCLUSION

In this paper, we have proposed a speech enhancement paradigm using a speaker ranking algorithm which can effectively retrieve a desired speech signal in a multi-talker environment. Specifically, the proposed speech enhancement paradigm exploits turn-taking behavior to determine the conversational partner amongst a set of candidate talker of a user by finding the talker with minimum probability of speech overlaps and gaps. The proposed algorithm only requires access to microphone signals, which is in contrast to existing methods which require additional sensor inputs, e.g. EEG, cameras, etc. We  demonstrated the proposed speech enhancement paradigm in two applications, where retrieval of a conversational partner's speech signal in a multi-talker environment, is desired. We compared the proposed systems to current state-of-the-art speech enhancement systems, and results indicate that the proposed systems significantly outperform the state-of-the-art systems.

## APPENDIX A PROOF OF MINIMIZING SPEECH OVERLAP AND GAP, AND MAXIMIZING MUTUAL EXCLUSION

In this Appendix, it is shown that minimizing the probability of speech overlap and gap is equivalent to maximizing the probability of mutual speech exclusion between the user's own voice VAD and candidate speaker's VAD. Specifically, we show that

$$
\arg\min_i \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0 = k, \alpha_i = k)
$$

$$
= \arg\max_i \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0 = k, \alpha_i = 1-k). \quad (32)
$$

*Proof:* The sum of the support of $P_{A_0 A_i}(\alpha_0 = k, \alpha_i = j)$ is equal to one such that

$$
\sum_{k=0}^{1} \sum_{j=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = j) = 1. \quad (33)
$$

The probabilities are split into the probability of speech overlap and gap, and the probability of mutual speech exclusion

$$
\sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = k)
$$

$$
= 1 - \sum_{j=0}^{1} P_{A_0 A_i}(\alpha_0(n) = j, \alpha_i(n) = 1-j). \quad (34)
$$

where the left-hand side is the probability of speech overlap and gap and the right-hand side is '1' subtracted by the

probability of mutual speech exclusion. Hence, minimizing the probability of speech overlap and gaps is equivalent to:

$$
\hat{i}_{\mathrm{MOG}}(n) = \arg\min_i \ 1 - \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = 1-k),
$$

$$
(35)
$$

or maximizing the probability of mutual speech exclusion:

$$
\hat{i}_{\mathrm{MOG}}(n) = \arg\max_i \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = 1-k),
$$

$$
(36)
$$

hence proving the equivalence in (32).

## APPENDIX B PROOF OF MINIMIZING SPEECH OVERLAP AND GAP, AND MAXIMIZING MEAN-SQUARE-ERROR

In this Appendix, we show that minimizing the probability of speech overlap and gap is identical to maximizing the mean-square-error between the own voice VAD and the candidate speaker VAD i.e.

$$
\arg\min_i \sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0(n) = k, \alpha_i(n) = k)
$$

$$
= \arg\max_i \mathbb{E}\left[ (A_0(n) - A_i(n))^2 \right]. \quad (37)
$$

*Proof:* The probability of speech overlap and gap is

$$
\sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0 = k, \alpha_i = k) = P_{A_0 A_i}(\alpha_0 = 1, \alpha_i = 1)
$$

$$
+ P_{A_0 A_i}(\alpha_0 = 0, \alpha_i = 0). \quad (38)
$$

We may then write

$$
P_{A_0 A_i}(\alpha_0 = 1, \alpha_i = 1) + P_{A_0 A_i}(\alpha_0 = 0, \alpha_i = 0)
$$

$$
= \sum_{k=0}^{1} \sum_{j=0}^{1} kj P_{A_0 A_i}(\alpha_0 = k, \alpha_i = j)
$$

$$
+ \sum_{m=0}^{1} \sum_{n=0}^{1} (1-m)(1-n) P_{A_0 A_i}(\alpha_0 = m, \alpha_i = n) \quad (39)
$$

and using the expectation operator, we have that

$$
\mathbb{E}[A_0 A_i] = \sum_{k=0}^{1} \sum_{j=0}^{1} kj P_{A_0 A_i}(\alpha_0 = k, \alpha_i = j)
$$

$$
\mathbb{E}[(1-A_0)(1-A_i)] = \sum_{m=0}^{1} \sum_{n=0}^{1} (1-m)(1-n)
$$

$$
\times P_{A_0 A_i}(\alpha_0 = m, \alpha_i = n). \quad (40)
$$

Hence, the probability of speech overlap and gap is

$$
\sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0 = k, \alpha_i = k) = \mathbb{E}[A_0 A_i] + \mathbb{E}[(1-A_0)(1-A_i)]
$$

$$
= 1 - \mathbb{E}[A_0] - \mathbb{E}[A_i] + 2\mathbb{E}[A_0 A_i]. \quad (41)
$$

Since $A_0$ and $A_i$ are Bernoulli random variables, then $\mathbb{E}[A_0] = \mathbb{E}[A_0^2]$ and $\mathbb{E}[A_i] = \mathbb{E}[A_i^2]$, such that the probability of speech overlap and gap is

$$\sum_{k=0}^{1} P_{A_0 A_i}(\alpha_0 = k, \alpha_i = k) = 1 - \mathbb{E}\left[(A_0 - A_i)^2\right], \quad (42)$$

where $\mathbb{E}\left[(A_0 - A_i)^2\right]$ is the mean-square-error (MSE) between $A_0$ and $A_i$. We see that the probability of speech overlap and gap is equivalent to $1 - \mathbb{E}\left[(A_0 - A_i)^2\right]$. Hence, the optimization problem for the MOG algorithm is

$$\begin{aligned}
\hat{i}_{\text{MOG}}(n) &= \arg\min_i \; 1 - \mathbb{E}\left[(A_0(n) - A_i(n))^2\right] \\
&= \arg\max_i \; \mathbb{E}\left[(A_0(n) - A_i(n))^2\right], \quad (43)
\end{aligned}$$

which is a maximization of the MSE between the own voice VAD and a candidate speaker VAD.

## APPENDIX C EXPECTED MISCLASSIFICATION RATE FOR MOG

The speaker misclassification rate is defined as the probability of classifying a wrong candidate speaker as the conversational partner. Using the MOG algorithm, we consider a misclassification as when $\Phi_t$ is equal to or smaller than $\Phi_v$. For a number $I$ of candidate speakers and integration time $T_{\text{int}}$, the misclassification rate $P(E = 1; I, T_{\text{int}})$ is given by

$$P(E = 1; I, T_{\text{int}}) = 1 - P\left(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1}\right),$$

where

$$\begin{aligned}
&P\left(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1}\right) \\
&= \sum_{\phi=1}^{N} p_{\Phi}(\phi; \gamma_t, \beta_t, N) P_{\Phi}^{I-1}(\phi - 1; \gamma_v, \beta_v, N) \quad (44)
\end{aligned}$$

denotes the correct classification rate, and $P_{\Phi}(\phi - 1; \gamma_v, \beta_v, N)$ is the cumulative distribution function of $p_{\Phi}(\phi - 1; \gamma_v, \beta_v, N)$,

$$P_{\Phi}(\phi - 1; \gamma_v, \beta_v, N) = \sum_{\kappa=0}^{\phi-1} p_{\Phi}(\kappa; \gamma_v, \beta_v, N). \quad (45)$$

*Proof:* First we consider the probability of correct classification under the assumption that $\Phi_{v,j}$ for all $j$'s are independent:

$$\begin{aligned}
&P\left(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1}\right) \\
&= \sum_{\phi=1}^{N} \sum_{\kappa_1=0}^{\phi-1} \cdots \sum_{\kappa_{I-1}=0}^{\phi-1} p_{\Phi}(\phi; \gamma_t, \beta_t, N) \\
&\quad \times p_{\Phi}(\kappa_1; \gamma_{v,1}, \beta_{v,1}, N) \\
&\quad \times \cdots \times p_{\Phi}(\kappa_{I-1}; \gamma_{v,I-1}, \beta_{v,I-1}, N)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{\phi=1}^{N} p_{\Phi}(\phi; \gamma_t, \beta_t, N) \sum_{\kappa_1=0}^{\phi-1} p_{\Phi}(\kappa_1; \gamma_{v,1}, \beta_{v,1}, N) \\
&\quad \times \cdots \times \sum_{\kappa_{I-1}=0}^{\phi-1} p_{\Phi}(\kappa_{I-1}; \gamma_{v,I-1}, \beta_{v,I-1}, N). \quad (46)
\end{aligned}$$

To simplify the expression, we define following cumulative distribution function

$$P_{\Phi}(\phi - 1; \gamma_{v,j}, \beta_{v,j}, N) = \sum_{\kappa_j=0}^{\phi-1} p_{\Phi}(\kappa_j; \gamma_{v,j}, \beta_{v,j}, N), \quad (47)$$

for all $j = 1, \ldots, I - 1$. Inserting (47) into (46), we have

$$\begin{aligned}
&P\left(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t > \Phi_{v,I-1}\right) \\
&= \sum_{\phi=1}^{N} p_{\Phi}(\phi; \gamma_t, \beta_t, N) P_{\Phi}(\phi - 1; \gamma_{v,1}, \beta_{v,1}, N) \\
&\quad \times \cdots \times P_{\Phi}(\phi - 1; \gamma_{v,I-1}, \beta_{v,I-1}, N). \quad (48)
\end{aligned}$$

Assuming that $\Phi_{v,j}$ are independent and identically distributed such that $\gamma_{v,j} = \gamma_v$, $\beta_{v,j} = \beta_v$ for all $j$, we may simplify to

$$\begin{aligned}
&P\left(\Phi_t > \Phi_{v,1}, \ldots, \Phi_t \geq \Phi_{v,I-1}\right) \\
&= \sum_{\phi=1}^{N} p_{\Phi}(\phi; \gamma_t, \beta_t, N) P_{\Phi}^{I-1}(\phi - 1; \gamma_v, \beta_v, N). \quad (49)
\end{aligned}$$

As the misclassification rate is

$$P(E = 1; I, T_{\text{int}}) = 1 - P\left(\Phi_t > \Phi_{v,1} \ldots, \Phi_t > \Phi_{v,I-1}\right), \quad (50)$$

then inserting (49) into (50) yields the derived result

$$\begin{aligned}
&P(E = 1; I, T_{\text{int}}) \\
&= 1 - \sum_{\phi=1}^{N} p_{\Phi}(\phi; \gamma_t, \beta_t, N) P_{\Phi}^{I-1}(\phi - 1; \gamma_v, \beta_v, N).
\end{aligned}$$

## REFERENCES

[1] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[2] J. Jensen and U. Kjems, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 295–299.

[3] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5728–5732.

[4] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[5] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 151–155.

[6] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.

[7] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[8] P. Hoang, Z.-H. Tan, J. M. de Haan, T. Lunner, and J. Jensen, "Robust Bayesian and maximum *a posteriori* beamforming for hearing assistive devices," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Ottawa, ON, Canada, Nov. 2019, pp. 1–5.

[9] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," 2017, *arXiv:1705.00919*.

[10] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 241–245.

[11] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[12] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 696–700.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[14] A. A. Nair, A. Reiter, C. Zheng, and S. Nayar, "Audiovisual zooming: What you see is what you hear," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, Oct. 2019, pp. 1107–1118.

[15] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers Neurosci.*, vol. 13, p. 153, Mar. 2019.

[16] A. Favre-Félix, C. Graversen, R. K. Hietkamp, T. Dau, and T. Lunner, "Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment," *Trends Hearing*, vol. 22, Jan. 2018, Art. no. 233121651881438.

[17] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066017.

[18] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Banff, AB, Canada Oct. 2017, pp. 3042–3047.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[20] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[21] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4. Hong Kong, Apr. 2003, pp. 1–4.

[22] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[23] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed-memory Markov process," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. Cambridge, MA, USA: MIT Press, 2005.

[24] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers Psychol.*, vol. 6, p. 731, Jun. 2015.

[25] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *J. Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010.

[26] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Sci. Rep.*, vol. 9, no. 1, p. 10451, Dec. 2019.

[27] A. J. Sørensen, M. Fereczkowski, and E. N. MacDonald, *Task Dialog by Native-Danish Talkers in Danish and English in Both Quiet and Noise*. Switzerland, U.K.: Zenodo, Mar. 2018.

[28] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behav. Res. Methods*, vol. 43, no. 3, pp. 761–770, Sep. 2011.

[29] Z.-H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, Jan. 2020.

[30] Z.-H. Tan. (Jul. 2020). *rVAD2.0*. [Online]. Available: https://github.com/zhenghuatan/rVAD

[31] S. Basu, "Conversational scene analysis," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, Sep. 2002.

[32] A. Harma and K. Pham, "Conversation detection in ambient telephony," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 4641–4644.

[33] Y. Kawaguchi, M. Togami, and Y. Obuchi, "Turn taking-based conversation detection by using DOA estimation," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, Makuhari, Japan, Sep. 2010, pp. 3134–3137.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio. Speech. Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Dec. 2011.

[35] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2. Salt Lake City, UT, USA, May 2001, pp. 749–752.

[37] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, 1998, pp. 2819–2822.

[38] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2002.

[39] 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, Aug. 2001.

[40] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[41] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. 2971–2981, May 2019.

[42] P. Minnaar, S. F. Albeck, C. S. Simonsen, B. Søndersted, S. D. Oakley, and J. Bennedbæk, "Reproducing real-life listening situations in the laboratory for testing hearing aids," Oct. 2013, Paper 8951.

[43] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

**POUL HOANG** received the M.Sc. degree in electrical engineering with specialization in signal processing and computing from Aalborg University, Aalborg, Denmark, in 2018. He is currently pursuing the industrial Ph.D. degree with Oticon A/S in collaboration with Aalborg University. His industrial Ph.D. is partly funded by the Innovation Fund Denmark. Since 2021, he has been employed as a DSP Specialist at Oticon A/S. His main research interests include speech enhancement and acoustic signal processing.

**ZHENG-HUA TAN** (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 1999. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA, an Associate Professor at the Department of Electronic Engineering, SJTU, and a Postdoctoral Fellow at the AI Laboratory, KAIST, Daejeon, South Korea. He is currently a Professor with the Department of Electronic Systems and the Co-Head of the Centre for Acoustic Signal Processing Research, Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He has (co)authored over 200 refereed publications. His research interests include machine learning, deep learning, pattern recognition, speech and speaker recognition, noise-robust speech processing, multimodal signal processing, and social robotics. He is the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC). He is an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He has served as an associate/guest editor for several other journals. He was the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

**JESPER JENSEN** received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. Student and an Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and an Assistant Professor with the Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. He is currently a Senior Principal Scientist with Oticon A/S, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, Aalborg University. He is also a Co-Founder of the Centre for Acoustic Signal Processing Research (CASPR), Aalborg University. His main research interests include acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

• • •

**JAN MARK DE HAAN** received the M.Sc. degree in electrical engineering from the Blekinge Institute of Technology, Karlskrona, Sweden, in 1998, and the Ph.D. degree in applied signal processing from the Department of Applied Signal Processing, Blekinge Institute of Technology, in 2004. In 2003, he was a Visiting Researcher at the Western Australia Telecommunication Research Institute, Perth, WA, Australia. Since 2004, he has been employed at Oticon A/S, Copenhagen, Denmark. His main research interests include acoustic signal processing and signal processing applications in hearing aids.