Generalized Turn-Taking–Based Conversational Partner Identification in Multi-Talker Scenarios using Voice Activity

Leila Eslami[1,3], *Lars Dalskov Mosgaard*[2], *Jesper Rindom Jensen*[3], *Bolaji Adesokan*[1], *Eline Borch Petersen*[2]
[1] WS Audiology, SIP Department, Lynge, Denmark
[2] ORCA Labs, Lynge, Denmark
[3] Audio Analysis Lab, Aalborg University, Denmark
Emails: {leila.eslami, larsdalskov.mosgaard, bolaji.adesokan, eline.petersen}@wsa.com, jrj@es.aau.dk

***Abstract*—Identifying conversational partners in multi-talker environments is essential for hearing-aid systems but remains challenging. Although traditional speech enhancement methods improve signal quality, they do not determine which talker the hearing-aid user (HAU) intends to engage with. The Bayesian Minimum Overlap-Gap (BMOG) algorithm is an existing turn-taking-based method for conversational partner inference using voice activity (VA), but it is restricted to dyadic conversations. This paper proposes the Multi-Talker BMOG (MT-BMOG) framework, which extends BMOG to multi-talker scenarios using only per-talker VA streams. The framework introduces a Conversational Alignment Score (CAS) to quantify turn-taking coordination across groups of arbitrary size. This score is modeled within a probabilistic framework to infer conversational partner groups and distinguish them from competing talkers. Evaluation on conversational datasets of varying sizes shows that MT-BMOG outperforms extended BMOG baselines. In real 2-, 3-, and 4-talker conversations, the proposed framework achieves mean partner identification accuracies of 97.1%, 90.1%, and 90.4%, respectively.**

***Index Terms*—Conversational partner identification, Turn-taking, Multi-talker conversations, Bayesian inference**

I. Introduction

Identifying conversational partners in multi-talker environments remains a fundamental challenge for speech enhancement systems operating in co-located source scenarios, including hearing aids. Beyond enhancing speech in noise, these systems must determine which talker the hearing-aid user (HAU) intends to engage with. Reliable conversational partner inference is therefore crucial for improving communication and supporting social interaction.

Traditional speech enhancement approaches rely on temporal filtering for single microphones and beamforming based on Direction-of-Arrival (DOA) estimation or spatial statistics for microphone arrays [1], [2]. However, in multi-talker scenarios, these methods may suppress the intended partner while preserving competing speech. Likewise, DOA-based estimators, including maximum-likelihood [3], [4] and deep-learning approaches [5], often degrade without reliable prior knowledge of the target talker's location or activity, leading to unstable speaker selection and degraded signal quality. Auxiliary-sensor approaches, such as EEG [6]–[8], eye-movement tracking [9], [10], and audiovisual fusion [11], can improve partner detection in controlled environments but are generally impractical for hearing aids due to cost and hardware complexity. This motivates microphone-only approaches based on conversational dynamics.

Turn-taking is a defining feature of human conversation, characterized by short gaps (approximately 200 ms) [12]–[14]. These temporal patterns provide cues for distinguishing conversational partners from competing talkers. The Bayesian Minimum Overlap-Gap algorithm (BMOG) [15] exploits turn-taking structure, but it was originally formulated for dyadic interactions and does not naturally extend to multi-talker settings where group-level coordination matters. While we do not address front-end processing here, methods for speaker separation, diarization [16]–[18], and voice activity detection methods [19]–[21] have been studied.

To address this limitation, we propose the Multi-Talker BMOG (MT-BMOG) framework. The proposed method introduces the Conversational Alignment Score (CAS), a generalized measure of turn-taking coordination for groups of arbitrary size, enabling principled partner inference in multi-talker conversations. We evaluated MT-BMOG on real 2-, 3- and 4-talker datasets and demonstrated consistent improvements over extended BMOG baselines.

II. Problem Formulation

In this work, we assume that Voice Activity Detection (VAD) has been performed and that per-talker binary activity streams, as well as the number of talkers, are available. This work focuses exclusively on the partner-inference stage.

We consider a conversational scenario in which the HAU interacts with $M-1$ conversational partners selected from K surrounding talkers, while the remaining $K-M+1$ talkers are considered competing talkers. The set of all non-HAU talkers is denoted by $\mathcal{T} = \{t_1, \ldots, t_K\}$. A binary voice-activity sequence $A_{t_k}(n) \in \{0, 1\}$ is defined for each talker $t_k \in \mathcal{T}$, where $A_{t_k}(n) = 1$ denotes speech and $A_{t_k}(n) = 0$ denotes silence at time index n. The HAU's voice activity is represented by $A_0(n) \in \{0, 1\}$.

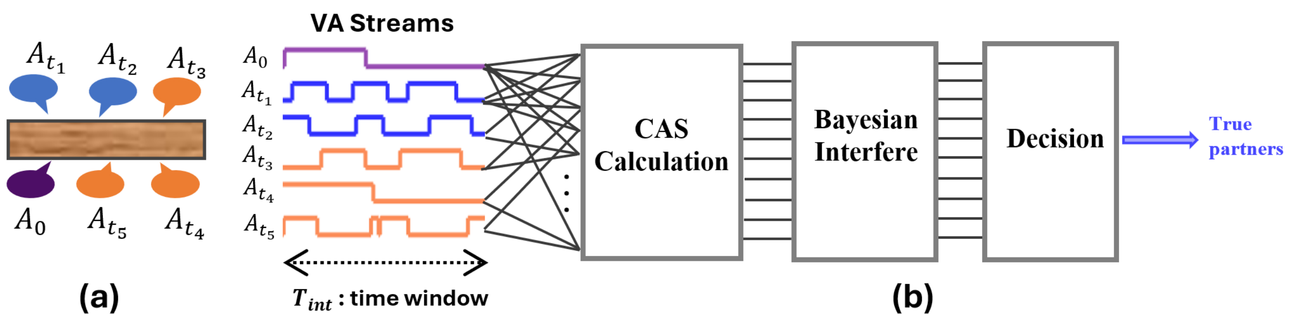

Fig. 1. Overview of the MT-BMOG framework. (a) Example 3-talker scenario with one HAU (A_0), two conversational partners (A_{t_1}, A_{t_2}), and three competing talkers ($A_{t_3}, A_{t_4}, A_{t_5}$). (b) System pipeline: within each time window T_{int}, per-talker VA streams are used to form all possible groups, including the HAU (A_0). For each candidate group (10 in the 3-talker case), the CAS is computed, followed by Bayesian inference and a decision to identify the group containing conversational partners.

We construct I candidate partner groups by pairing the HAU with all possible subsets of $M-1$ talkers selected from $\mathcal{T}$.

The goal is to identify the subset of candidates that form the conversational partners of the HAU. Let H_i denote the hypothesis that candidate group $i \in \{1, \dots, I\}$ forms the true conversational partner group (TG). Conversational partner identification is then formulated as a multi-hypothesis decision problem,

$$\hat{TG} = \arg \max_{i \in \{1,\dots,I\}} P\big(H_i \mid f_1^{(M)}, f_2^{(M)}, \dots, f_I^{(M)}\big), \quad (1)$$

where $f_i^{(M)}(\cdot)$ denotes a turn-taking coordination function defined for the M-talker case and candidate group i. In the BMOG algorithm, the coordination function is defined for the dyadic case between the HAU and a candidate talker k as $f_k^{(2)}(n) = \big(A_0(n) - A_{t_k}(n)\big)^2$, minimizing overlap and gap frames in 2-talker conversations [15]. However, this dyadic formulation does not generalize reliably to group conversations because it does not capture higher-order interactions among multiple talkers. Therefore, a generalized formulation of the BMOG algorithm is required for multi-talker conversations.

III. Proposed Algorithm

The proposed framework, MT-BMOG, identifies conversational partners in multi-talker conversations by introducing the CAS. The method is first introduced for a 3-talker scenario ($M = 3$) to build intuition and is subsequently generalized to arbitrary group sizes. Fig. 1(a) illustrates an example 3-talker interaction involving one HAU, two conversational partners, and three competing talkers. The corresponding MT-BMOG processing pipeline is shown in Fig. 1(b), including candidate group formation, CAS computation, Bayesian inference, and the final decision step.

TABLE I
Joint VA states of the HAU (A_0) and two talkers (t_x, t_y) in a 3-talker conversation scenario.

States	VA Configuration	Description
S_1	$A_0 = 0,\ A_{t_x} = 0,\ A_{t_y} = 0$	All talkers silent
S_2	$A_0 = 1,\ A_{t_x} = 0,\ A_{t_y} = 0$	Only HAU active
S_3	$A_0 = 0,\ A_{t_x} = 1,\ A_{t_y} = 0$	Only partner t_x active
S_4	$A_0 = 0,\ A_{t_x} = 0,\ A_{t_y} = 1$	Only partner t_y active
S_5	$A_0 = 1,\ A_{t_x} = 1,\ A_{t_y} = 0$	HAU and t_x active
S_6	$A_0 = 1,\ A_{t_x} = 0,\ A_{t_y} = 1$	HAU and t_y active
S_7	$A_0 = 0,\ A_{t_x} = 1,\ A_{t_y} = 1$	t_x and t_y active
S_8	$A_0 = 1,\ A_{t_x} = 1,\ A_{t_y} = 1$	All talkers active

A. Conversational Alignment Score (CAS)

Table I shows the possible joint VA states of the HAU and two partner talkers in a 3-talker scenario, presented here for illustration. Afterward, a generalization to M talkers is introduced. States S_2–S_4 correspond to ideal turn-taking, where exactly one talker is active at a time, whereas S_5–S_8 represent overlapping speech and S_1 corresponds to a gap in which all talkers are silent.

Turn-taking coordination is quantified by emphasizing the ideal turn-taking states. Among the I candidate partner groups, the group i corresponding to the conversational partners (TG) is expected to maximize the likelihood of observing states S_2–S_4 over time. For the 3-talker case, with candidate group i containing talkers $t_x, t_y \in \mathcal{T}$ together with the HAU, this leads to the following closed-form coordination function:

$$\begin{aligned} f_i^{(3)}(n) = &(A_0(n) - A_{t_x}(n))^2 + (A_0(n) - A_{t_y}(n))^2 \\ &+ (A_{t_x}(n) - A_{t_y}(n))^2 + 3A_0(n)A_{t_x}(n)A_{t_y}(n) \\ &- A_0(n) - A_{t_x}(n) - A_{t_y}(n). \end{aligned} \quad (2)$$

Generalization beyond three talkers follows the same principles using pairwise differences and group interaction terms. Let candidate group i contain $M-1$ non-HAU talkers, denoted by $G_i = \{t_{k_1}, \dots, t_{k_{M-1}}\} \subset \mathcal{T}$, and define the set of talkers in the group (including the HAU) as $\mathcal{Q}_i \triangleq \{0\} \cup G_i$. The M-talker coordination function is defined as:

$$\begin{aligned} f_i^{(M)}(n) \triangleq &\sum_{\substack{p,q \in \mathcal{Q}_i \\ p<q}} \big(A_p(n) - A_q(n)\big)^2 - (M-2) \sum_{z \in \mathcal{Q}_i} A_z(n) \\ &+ \sum_{m=3}^{M} m(-1)^{m+1} \sum_{\substack{U \subseteq \mathcal{Q}_i \\ |U|=m}} \prod_{v \in U} A_v(n). \end{aligned} \quad (3)$$

where U denotes a subset of $\mathcal{Q}_i$ with cardinality $|U| = m$. The CAS value can be estimated by averaging $f^{(M)}$ over N

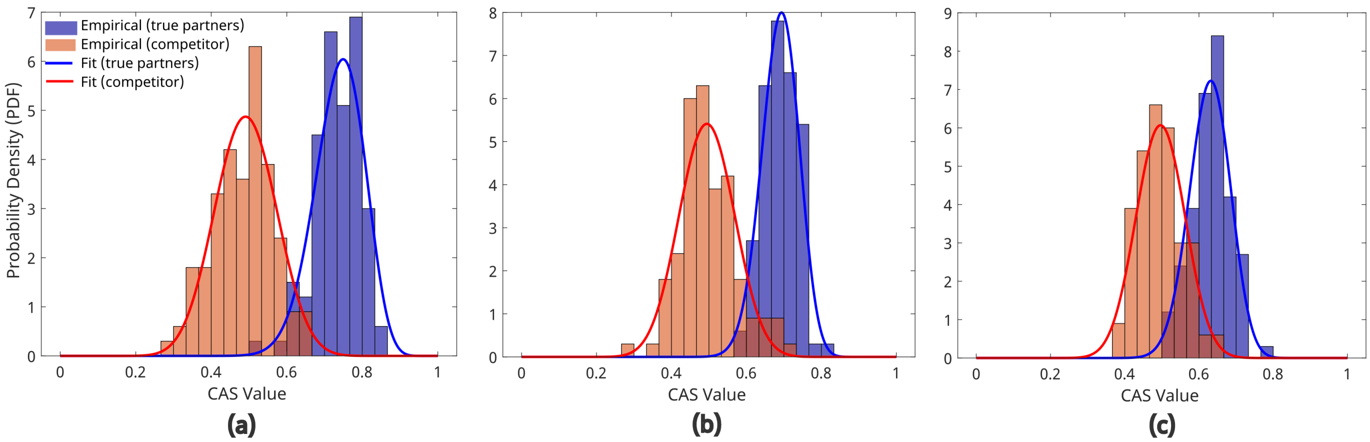

Fig. 2. Normalized CAS distributions for (a) 2-talker at $T_{\text{int}} = 23\,\text{s}$, (b) 3-talker at $T_{\text{int}} = 91\,\text{s}$, and (c) 4-talker at $T_{\text{int}} = 240\,\text{s}$. Histograms display the CAS values for conversational partners (blue) and competing talkers (red), and fitted curves show the Beta–Binomial modeling used for Bayesian inference.

consecutive VA time steps. The CAS value for the candidate group i in the M-talker case is obtained as:

$$\widehat{\mathrm{CAS}}_i^{(M)}(n;N) = \sum_{r=n-N+1}^{n} f_i^{(M)}(r). \tag{4}$$

where $f_i^{(M)}(n)$ is obtained from Eq. (3) for M-talkers at time index n, with integration time $T_{\text{int}} = N/f_s$ and f_s denoting the VA sampling rate. Then, the normalized CAS is obtained by dividing the CAS value by the number of samples within the integration window, N, yielding a score bounded between 0 and 1.

B. Probabilistic Modeling of CAS

We model partner inference as a hypothesis test, where H_i denotes the hypothesis that candidate group i is the conversational partner set. The CAS_i value quantifies the amount of well-timed turn-taking within an integration window of duration T_{int} (N samples). Following [15] and supported by the empirical distributions shown in Fig. 2, CAS_i is modeled using a Beta–Binomial distribution:

$$p(\mathrm{CAS}_i;N,\gamma,\beta) = \binom{N}{\mathrm{CAS}_i}\frac{\mathrm{B}(\mathrm{CAS}_i+\gamma,\ N-\mathrm{CAS}_i+\beta)}{\mathrm{B}(\gamma,\beta)}, \tag{5}$$

TABLE II
ESTIMATED BETA–BINOMIAL PARAMETERS FOR 2-, 3-, AND 4-TALKER CONVERSATIONAL DATASETS.

	2-Talker		3-Talker		4-Talker	
	$\tilde{\gamma}_t$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\beta}_t$
$\hat{a}$	2.5091	0.8522	2.5128	1.1434	2.1020	1.2110
$\hat{b}$	0.7879	0.7817	0.6363	0.6344	0.5662	0.5709
	$\tilde{\gamma}_c$	$\tilde{\beta}_c$	$\tilde{\gamma}_c$	$\tilde{\beta}_c$	$\tilde{\gamma}_c$	$\tilde{\beta}_c$
$\hat{a}$	0.7736	0.8057	0.3074	0.3382	0.6225	0.6083
$\hat{b}$	0.9727	0.9681	0.8677	0.8528	0.6934	0.7000

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function and (γ,β) are the Beta prior parameters. We define two Beta–Binomial models with separate parameter sets (γ_t,β_t) and (γ_c,β_c) corresponding to conversational partners and competing talkers, respectively. Under hypothesis H_i, the likelihoods are modeled as $\mathrm{CAS}_i \sim \mathrm{BetaBinomial}(N;\gamma_t,\beta_t)$ for the conversational partner group, and $\mathrm{CAS}_j \sim \mathrm{BetaBinomial}(N;\gamma_c,\beta_c)$ for all $j \neq i$ (competing talkers).

C. Bayesian Inference and Decision

Assuming conditional independence of CAS scores across candidate groups given hypothesis H_i, the likelihood under hypothesis H_i is given by:

$$P(\mathrm{CAS}_1,...,\mathrm{CAS}_I;H_i) = p(\mathrm{CAS}_i;N,\gamma_t,\beta_t)\prod_{j\in\mathcal{I}_{\setminus i}} p(\mathrm{CAS}_j;N,\gamma_c,\beta_c). \tag{6}$$

where $\mathcal{I}_{\setminus i}$ denotes the set of competing speakers under hypothesis H_i, i.e., $\mathcal{I}$ excluding the element i.

Assuming a uniform prior, $P(H_i) = 1/I$, the posterior probability is

$$P(H_i \mid \mathrm{CAS}_1,...,\mathrm{CAS}_I) = \frac{P(H_i)p(\mathrm{CAS}_i;N,\gamma_t,\beta_t)\prod_{j\in\mathcal{I}_{\setminus i}} p(\mathrm{CAS}_j;N,\gamma_c,\beta_c)}{\sum_{k=1}^{I} P(H_k)p(\mathrm{CAS}_k;N,\gamma_t,\beta_t)\prod_{q\in\mathcal{I}_{\setminus k}} p(\mathrm{CAS}_q;N,\gamma_c,\beta_c)} \tag{7}$$

Finally, the conversational partner set is selected by maximizing the posterior probability as follows:

$$\hat{TG} = \arg\max_{i\in\{1,...,I\}} P(H_i \mid \mathrm{CAS}_1,...,\mathrm{CAS}_I). \tag{8}$$

IV. EXPERIMENTAL SETUP AND METHODS

A. Datasets and Evaluation

All methods were evaluated on real 2-, 3-, and 4-talker conversational datasets [22]–[24], covering diverse conditions including low/high noise levels, normal/impaired hearing status, and aided/unaided hearing-aid use. The Beta–Binomial

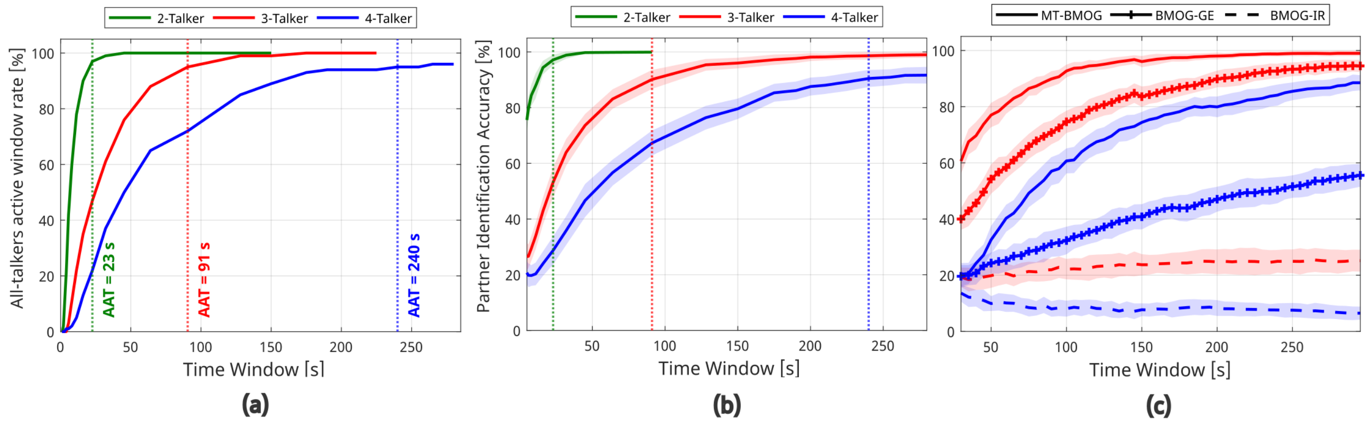

Fig. 3. Performance analysis as a function of the integration time for 2-, 3-, and 4-talker conversations. (a) Percentage of time windows in which all talkers are active as a function of window length. Vertical dotted lines indicate the All-Talker Activity Threshold Time ($\mathrm{ATT}_{0.95}$) for each scenario. (b) Partner identification accuracy of the proposed MT-BMOG framework versus integration time. Vertical dotted lines show $\mathrm{ATT}_{0.95}$. (c) Comparison of MT-BMOG with extended BMOG baselines (BMOG-IR and BMOG-GE) for 3- and 4-talker scenarios. Lines denote the mean partner identification accuracy, and shaded regions indicate the standard deviation across 50 repeated random trials.

shaping parameters of the CAS distributions for conversational partners and competing talkers were estimated using the power model in [15], with the resulting parameters summarized in Table II.

Conversational activity was analyzed as a function of window length by computing the percentage of sampled windows in which all talkers were active (Fig. 3(a)). A talker was considered active if the window contained at least one continuous speech segment of duration ≥ 1 s. We define the All-Talker Activity Threshold Time (ATT_α) as the minimum window length for which at least an α fraction of sampled windows contains activity from all talkers. For each window length, 100 windows were randomly sampled. In this study, $\alpha = 0.95$, such that $\mathrm{ATT}_{0.95}$ denotes the window length at which 95% of the sampled windows contain activity from all talkers.

Conversational and competing groups were generated using 100 random realizations per dataset. In each realization, two distinct conversational groups were selected. A talker from the first group was randomly selected as the HAU; the remaining members of that group were treated as conversational partners, and all talkers from the second group were treated as competing talkers.

Partner identification accuracy was used as the performance metric. For a given integration time and group size, 100 conversational scenarios were randomly generated. In each scenario, the proposed method estimated the conversational partner group. Accuracy was computed as the percentage of conversational partners correctly identified across the 100 trials. To reduce the influence of random sampling, the entire evaluation procedure was repeated 50 times using different random realizations. The mean of the resulting accuracies was subsequently reported as a function of integration time.

B. Extended BMOG Methods

To enable a fair comparison with MT-BMOG and the original BMOG approach, we implement two BMOG-based extensions for multi-talker scenarios.

The first method, BMOG-Individual Ranking (BMOG-IR), applies BMOG independently between the HAU and each candidate talker, forming all possible user–candidate dyads. BMOG posterior probabilities are then computed for each dyad, and the top $M-1$ ranked candidates are selected as conversational partners in the M-talker case. The second method, BMOG-Group Evaluation (BMOG-GE), evaluates candidate partner groups jointly. For each candidate group, the VA streams of the non-HAU talkers are merged into a single binary activity sequence by summing and binarizing the VA streams. This representation captures whether at least one group member is active, but discards overlap information within the group. BMOG is then applied between the HAU and the merged group activity sequence, and the group with the highest posterior probability is selected as the conversational partner set.

V. Results and Discussion

A. CAS Distributions Across Group Sizes

Fig. 2 shows the CAS distributions for conversational partners and competing talkers, together with fitted Beta–Binomial likelihoods, for 2-, 3-, and 4-talker conversations evaluated at $\mathrm{ATT}_{0.95}$. Conversational partners consistently exhibit higher CAS values than competing talkers, supporting the use of conversational alignment as a cue for partner inference. The separation is strongest in the 2-talker case, remains clear in the 3-talker case, and becomes narrower in the 4-talker case. This trend reflects the increased overlap and higher-order interaction complexity in larger conversational groups. Nevertheless, the fitted models still capture distinct differences between conversational partners and competing talkers, supporting probabilistic partner identification even in 4-talker scenarios.

B. MT-BMOG Performance

Fig. 3 summarizes the effect of integration window length on conversational activity and partner identification performance. Fig. 3(a) shows the percentage of sampled windows in which all talkers are active as a function of window length. The $ATT_{0.95}$ increases substantially with the number of talkers, reaching 23 s, 91 s, and 240 s for the 2-, 3-, and 4-talker cases, respectively. This trend reflects the reduced likelihood of observing active participation from every participant in larger conversational groups over short time windows.

Fig. 3(b) shows that MT-BMOG accuracy increases with integration time and follows the corresponding activity trends in Fig. 3(a). At $ATT_{0.95}$, the mean partner identification accuracy reaches 97.1% for 2 talkers, 90.1% for 3 talkers, and 90.4% for 4 talkers. Thus, when the observation window is long enough for 95% of sampled windows to contain activity from all talkers, MT-BMOG achieves high partner identification accuracy across multi-talker scenarios. This indicates that the performance of MT-BMOG depends strongly on whether each individual talker contributes sufficient activity within the observation window.

Fig. 3(c) compares MT-BMOG with the two extended BMOG baselines. BMOG-IR exhibits low partner identification accuracy in both the 3- and 4-talker scenarios because its purely pairwise formulation does not capture group-level conversational dynamics. Consequently, its performance does not improve substantially with increasing integration time and does not follow the activity trends shown in Fig. 3(a). BMOG-GE improves performance through group-level aggregation but still underperforms MT-BMOG, particularly in the 4-talker case. This is because BMOG-GE merges the activity streams of candidate partners and therefore discards within-group overlap information. In contrast, MT-BMOG explicitly models multi-talker turn-taking coordination and consistently achieves higher accuracy.

Finally, the reduced performance observed with an increasing number of conversational partners, particularly in the 4-talker condition, is caused by the inherent dynamics of the interaction, where fewer talkers are active as the group size increases (Fig. 3(a)). Future work will integrate complementary cues and more robust speaker separation and VAD methods to improve partner inference and practical feasibility in complex multi-talker scenarios.

VI. Conclusion

We proposed MT-BMOG, a multi-talker extension of BMOG that uses the Conversational Alignment Score (CAS) for partner inference. Evaluation on real 2-, 3-, and 4-talker datasets showed that MT-BMOG consistently outperformed extended BMOG baselines. The results indicate that larger groups require longer observation windows to capture sufficient turn-taking activity, but high identification accuracy is achieved once sufficient activity is observed. Future work will focus on reducing integration time by incorporating complementary cues and improving robustness to speaker separation and VAD errors.

References

[1] A. Kuklasiński, 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, 2016.

[2] O. Schwartz, S. Gannot, and E. A. 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.* IEEE, 2016, pp. 151–155.

[3] 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, 2017.

[4] P. Hoang *et al.*, "Robust bayesian and maximum a posteriori beamforming for hearing assistive devices," in *Proc. IEEE Global Conf. Signal and Information Process.* IEEE, 2019, pp. 1–5.

[5] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2017, pp. 136–140.

[6] N. Das, A. Bertrand, and T. Francart, "Eeg-based auditory attention detection: boundary conditions for background noise and speaker positions," *J. Neural Engineering*, vol. 15, no. 6, p. 066017, 2018.

[7] J. Hjortkjær *et al.*, "Real-time control of a hearing instrument with eeg-based attention decoding," *J. Neural Engineering*, vol. 22, no. 1, p. 016027, 2025.

[8] M. Jaeger, B. Mirkovic, M. G. Bleichner, and S. Debener, "Decoding the attended speaker from eeg using adaptive evaluation intervals captures fluctuations in attentional listening," *Frontiers in Neuroscience*, vol. 14, p. 603, 2020.

[9] A. Favre-Felix *et al.*, "Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment," *Trends in Hearing*, vol. 22, p. 2331216518814388, 2018.

[10] M. A. Skoglund *et al.*, "Comparing in-ear eog for eye-movement estimation with eye-tracking: Accuracy, calibration, and speech comprehension," *Frontiers in Neuroscience*, vol. 16, p. 873201, 2022.

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

[12] R. E. Corps, C. Gambi, and M. J. Pickering, "Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation," *Discourse Processes*, vol. 55, no. 2, pp. 230–240, 2018.

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

[14] M. Barthel, S. Sauppe, S. C. Levinson, and A. S. Meyer, "The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm," *Frontiers in Psychology*, vol. 7, p. 1858, 2016.

[15] P. Hoang, Z.-H. Tan, J. M. De Haan, and J. Jensen, "The minimum overlap-gap algorithm for speech enhancement," *IEEE Access*, vol. 10, pp. 14 698–14 716, 2022.

[16] Z.-Q. Wang *et al.*, "TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[17] C. Subakan *et al.*, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 21–25.

[18] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: State-space model is all you need in speech separation," 2024. [Online]. Available: https://arxiv.org/abs/2404.02063

[19] H. Dinkel *et al.*, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Trans. Audio., Speech, Lang. Process.*, vol. 29, pp. 1542–1555, 2021.

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

[21] S. Tao *et al.*, "Single channel speech presence probability estimation based on hybrid global-local information," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, 2023, pp. 1–5.

[22] A. J. Sørensen, M. Fereczkowski, and E. MacDonald, "Task dialog by native-danish talkers in danish and english in both quiet and noise," *Zenodo*, vol. 10, 2018. [Online]. Available: https://zenodo.org/records/1204951

[23] E. B. Petersen, "Investigating conversational dynamics in triads: Effects of noise, hearing impairment, and hearing aids," *Frontiers in Psychology*, vol. 15, p. 1289637, 2024.

[24] R. Nicoras *et al.*, "Effective design for experiments on small-group conversation: Insights from an example study," *American J. Audiology*, vol. 34, no. 2, pp. 305–320, 2025.