*Original Article*

# Sound Source Selection Based on Head Movements in Natural Group Conversation

Trends in Hearing
Volume 26: 1–8

DOI: 10.1177/23312165221097789
journals.sagepub.com/home/tia
SAGE

**Hao Lu[1] and W. Owen Brimijoin[2]**

**Abstract**

To optimally improve signal-to-noise ratio in noisy environments, a hearing assistance device must correctly identify what is signal and what is noise. Many of the biosignal-based approaches to solving this question are themselves subject to noise, but head angle is an overt behavior that may be possible to capture in practical devices in the real world. Previous orientation studies have demonstrated that head angle is systematically related to listening target; our study aimed to examine whether this relationship is sufficiently reliable to be used in group conversations where participants may be seated in different layouts and the listener is free to turn their body as well as their head. In addition to this simple method, we developed a source-selection algorithm based on a hidden Markov model (HMM) trained on listeners' head movement. The performance of this model and the simple head-steering method was evaluated using publicly available behavioral data. Head angle during group conversation was predictive of active talker, exhibiting an undershoot with a slope consistent with that found in simple orientation studies, but the intercept of the linear relationship was different for different talker layouts, suggesting it would be problematic to rely exclusively on this information to predict the location of auditory attention. Provided the location of all target talkers is known, the HMM source selection model implemented here, however, showed significantly lower error in identifying listeners' auditory attention than the linear head-steering method.

**Keywords**

head tracking, wearable device, natural group conversation, sound source selection

Received 16 November 2021; Revised 12 April 2022; accepted 13 April 2022

## Introduction

### *Source Selection Problem in Group Conversation*

Group conversation is an important form of daily social interaction, but it is also commonly conducted in noisy environments such as restaurants and classrooms, which can greatly affect ease of communication. Older and hearing-impaired listeners tend to rate group conversations as a particularly difficult situation (Gatehouse & Noble, 2004; Strawbridge et al., 2000), and normal-hearing listeners' ease of conversation is also affected by background noise in group conversation (Maruyama et al., 2020; McKellin et al., 2007). One approach to improving speech intelligibility and ease of conversation in such environments is to increase the signal-to-noise ratio. A typical way to achieve this is via beamforming, which is frequently applied in modern hearing aids (e.g., Moore et al., 2019), and was designed to preserve signals from one direction while attenuating signals from other directions. The beamforming mode on hearing aids has frequently been shown to provide better speech understanding over omnidirectional mode (Bentler, 2005). However, to get the most benefit from beamforming requires two important assumptions to be true: first that the sound sources are spatially separated, and second that the beam is pointed correctly at the sound source to which the user is listening. The first assumption is often true in real-life situations when the sound sources are talkers in group conversation, but for the second assumption to always hold would require listeners to point their heads in particular ways or, alternatively, would require a source selection model that is capable of accurately identifying the target of the user's auditory attention on a moment-to-moment basis and steering the beam in this

[1]Department of Psychology, University of Minnesota, Minneapolis, MN, USA
[2]Reality Labs Research, Redmond, WA, USA

**Corresponding author:**
Hao Lu, University of Minnesota, N218 Elliott Hall, 75 East River Parkway, Minneapolis, MN 55455, USA.
Email: luxx0489@umn.edu

direction. When the second assumption is violated, i.e., when the direction of the beamformer does not align well with the user's direction of attention, the user will not receive optimal SNR benefit and may even experience difficulty when trying to orient towards the desired sound source (Brimijoin et al., 2014; Hládek et al., 2019).

Enhancing the correct sounds for a person would require a source selection model that correctly reflects the user's auditory attention; the ideal construction of such a model is an unsolved research question. Selection system can be explicit, e.g., manual selection of sound sources via button presses, or implicit, e.g., based on a statistical model of natural behavior in conversation. Explicit models are reliable but inflexible, and implicit models based solely on behavior cannot yet fully predict attention. That said, even the simple implicit models are a potentially valuable approach; compared to source selection methods using buttons or pointing, source selection based on eye gaze was found to be faster and was rated as feeling more natural by hearing-aids users in simulated multi-talker situations (Hart et al., 2009), suggesting that implicit behaviors can be a promising method of source selection. Yes, sophisticated attention models might take advantage of the multiple behavioral and neural signals that have been shown to be correlated with auditory attention, such as EEG (e.g., Biesmans et al., 2017; Bleichner et al., 2016) and eye gaze (e.g., Favre-Félix et al., 2018; Kidd, 2017), but these approaches require expensive sensors, significant power consumption, careful calibration, and may be vulnerable to noise, limiting their current useability on devices to be used in daily social interactions on moving users. We argue that head movement (Bentler, 2005) is a potentially more pragmatic choice, as it can be conveniently estimated with inertial measurement units (IMU) and/or cameras on wearable devices (e.g., Nützi et al., 2011).

## *The Undershoot Problem of Head-Steering Beamformers*

Although more accessible and lower-noise than other measurements, head orientation has been shown to undershoot the true location of the target (Guitton & Volle, 1987). This undershoot poses a problem for purely head-steered selection systems because a listener's head may be pointed at one target, but their eyes (and attentional focus) are focused on a target further off their midline. The orientation undershoot, however, appears to be consistent, and thus could nonetheless provide information on the attended sound. When torso midline is used as a basis, the relative angle between head orientation and target location has been shown to follow an approximately linear relationship (Brimijoin et al., 2010). Similar undershoot trends have been found in natural three-party group conversation (Lu et al., 2021), four-party group conversation (Stiefelhagen & Zhu, 2002), and in more complicated virtual-reality environments (Hendrikse et al., 2019). Thus, while there is an undershoot, the linear relationship found in lab settings suggests it may be possible to predict the true location of an auditory target using head orientation alone. We hypothesized however, that such a linear model may be unstable due to different room layouts, different talker positions, and variation caused by individual differences. Therefore, it was necessary to examine whether such a linear relationship is stable enough across room layouts and participants to be applied as the sole basis of a source selection model or whether additional statistical modeling would be appropriate.

## *Alternative Source Selection Method Based on Hidden Markov Model*

Arguably any source selection method, whether it is driven by head angle alone or by more sophisticated statistical modeling, would perform better if the source-selection approach were simplified to be a classification model (selecting one from N candidate sources), rather than a regression model (predicting the exact angle of attention). Luckily, in group conversation situations, all talkers involved in conversation are usually sitting or standing at relatively fixed locations, so the locations of all potential target talkers are spatially separated and approximately fixed unless talkers join or leave group conversation, even in a busy cafeteria. As the locations of talkers are usually bounded during group conversation, they can be registered through cameras (e.g., Aghaei et al., 2016) or a head-mounted microphone array (e.g., Tourbabin et al., 2019). Once the number and locations of possible target talkers are identified, source selection can be simplified as a classification problem (Grimm et al., 2020), where continuous head movements are used to predict discrete auditory attention states rather than angle of attention. A Hidden Markov model (HMM) is a suitable unsupervised model for this type of classification problem (Eddy, 2004), and it has been widely applied in the analysis of eye tracking data to reveal how eye gaze switches across multiple regions of interest (e.g., Chuk et al., 2020; Kim et al., 2020; Simola et al., 2008). In the analysis of head movements and utterances of four participants in group conversation, an HMM was shown to be able to achieve over 80% accuracy in estimating auditory attention of all participants (Otsuka et al., 2005). To our knowledge, however, the performance of HMMs in predicting auditory attention targets with only listener's head movement has not yet been evaluated in natural conversation.

The current study aimed to address the above questions and in the process examined two types of source selection models, both based on listeners' head movements during group conversation. The first was based solely on the linear relationship between target location and head orientation, and the second was based on an HMM with known target

locations. The performance and stability of both models were examined using typical conversational layouts.

## Materials and Methods

### Experiment Data

The group conversation data used in this study were drawn from a public dataset called EasyCom (Donley et al., 2021). The EasyCom dataset includes multiple types of behavior data gathered during natural group conversations in 71 dB SPL background noise. The participants were seated and were asked to engage in conversation during several tasks, including group introductions, ordering food, solving puzzles, playing games, and reading sentences out loud. The dataset was originally collected for a machine learning project on egocentric video, so demographic data were not collected from participants, unfortunately precluding us from examining the effect of gender, age, and hearing threshold. While it has been found that age and hearing level may not affect head movements and eye gaze in group conversation (Lu et al., 2021), future work will need to incorporate these data. For the current study we used data from 9 sessions of 3 participants sitting around a table plus a standing host (4-people session), and 2 sessions of 4 sitting participants plus a standing host (5-people session). The room layout is shown in Figure 1. From the two types of sessions in the EasyCom dataset, we extracted head movements recorded by an Optitrack system and manually labeled speech activity using the egocentric video and audio recording. The head location and orientation of all sitting participants (ID 2, 3, 4, 5, 6, 7 in Figure 1) was recorded at 20-Hz sample frequency. The average length of the sessions used in this study is 28.9 min (min: 27.2 min, max: 29.6 min).

### Preprocessing

The quaternion rotations from the EasyCom dataset were converted to Euler angles following the rotation order YXZ (from the view of participant 2 in Figure 1, positive X points left; positive Y points upwards; positive Z points forward) to obtain the yaw, pitch, and roll movements of the head during conversation. As the head location of all participants were approximately on the same horizontal plane, only yaw was used in the subsequent analyses. When analyzing yaw head movement for each participant, the zero-yaw angle was arbitrarily defined as the cardinal direction closest to the direction of other participants, e.g., when all other participants were on the north side of one participant, the zero-yaw movement angle for that participant was defined as south.

The target of auditory attention was defined by manually labeled speech activity segments, further validated based on the audio and video recording. Because egocentric video was only available for the participants at location 2 in the EasyCom dataset, only the auditory target of this participant

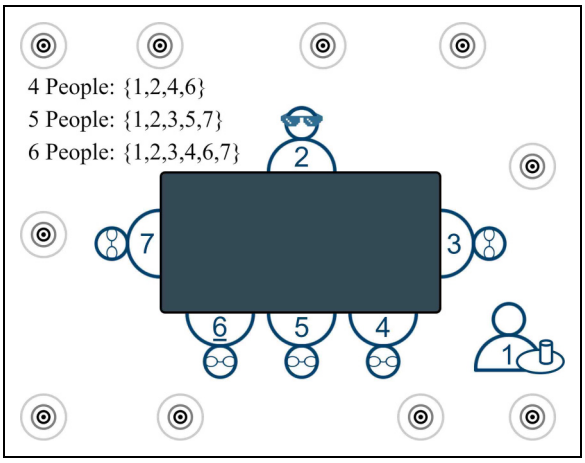


**Figure 1.** Scenario layout for the participants. The participants sat around a table with Participant IDs based on their role or seated location. The standing host was assigned ID 1, and all sitting participants were assigned ID 2-7. The circular rings indicate noise sound sources. The sets of Participant IDs are shown for each total number of participants. This study used 4-people sessions with sitting participant 2, 4, 6 and a host 1, and 5-people sessions with sitting participants at 2, 3, 5, 7 and a host 1. The AR glasses with egocentric camera and microphone array are indicated by the color-filled glasses. This figure is directly from Figure 3 of the EasyCom dataset description (Donley et al., 2021).

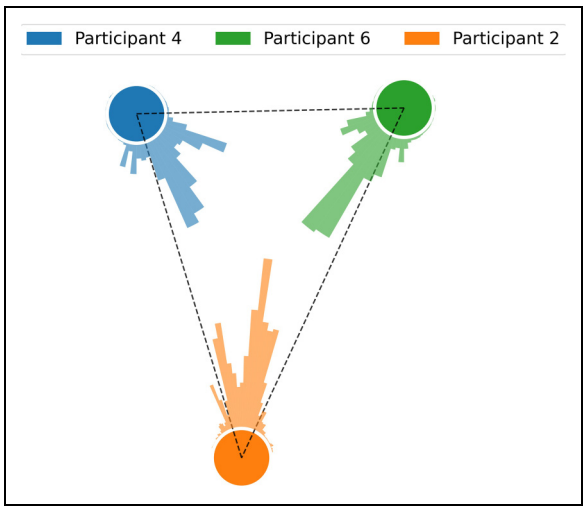


**Figure 2.** Distribution of head orientation of 3 participants in one representative 4-people session. The one standing host is not shown in this figure. Three circles represent the true location of three participants, and the dashed lines connecting them represent the direction of one participant's true location in the view of another participant. The polar histograms represent the distribution of each participant's head orientation during the entire conversation session.

was manually annotated. Five graders were instructed to independently label the most likely auditory target, based on the egocentric video and audio recording. When only a single person was speaking, it was assumed that all other participants' auditory attention was focused on that active talker. When there were competing talkers and/or graders gave inconsistent labels, the auditory target was defined as the most frequently identified speaker. One 4-people session was dropped due to inconsistent attention annotation across graders. In total, eight 4-people session and two 5-people sessions were analyzed.

## HMM Fitting and Analysis

For each participant in each session, the number of hidden discrete states for the HMM was set as the number of sitting target talkers (2 target talkers for 4-people sessions, and 3 target talkers for 5-people sessions). By the basic assumption of HMMs, the observed continuous variables followed distributions defined through emission functions corresponding to each hidden state. In this case, we assumed that the observed head movements behaved according to a normal distribution with a mean and variance that depended on the target talker. All other parameters were initialized using the defaults of the hmmlearn package (https://github.com/hmmlearn/hmmlearn). The HMMs designed following the described protocol were then used in the three following analyses.

First, the relationship between head movement distribution and the true location of target talkers was analyzed. The HMMs were individually fitted on the head movements of all three participants in the eight 4-people sessions. The means of the emission functions in the fitted HMMs were used to represent the overall head orientation when listening. The 5-people sessions were excluded from this analysis because two sessions were insufficient for statistical analysis.

Second, the performance of HMMs in identifying the attended target talker was evaluated against the linear model. The HMMs were individually fitted on the head yaw data of participant 2 (location shown in Figure 1) of all 4-people sessions and 5-people sessions. The decoded hidden states were paired with the target talker through the estimated means of the emission functions and the average of true locations of the entire session. The performance of the fitted HMM and head orientation was quantified as the yaw angle error between the predicted target direction and the true location of the attended target talker. The average locations of target talkers through the entire session were calculated and used to convert the hidden states of HMM to a yaw angle estimate. All participants other than participant 2 were excluded from this analysis because auditory target annotation was only available for participants sitting at location 2 (participant 2 in Figure 1) as described in the preprocessing subsection.

Third, since the performance of an HMM depends on the quantity of data available – and data acquisition in real environments may be problematic, we also evaluated the accuracy of the HMM fitted on truncated data. For each 4-people session, HMMs were fitted on head movements limited to the first x minutes of group conversation, with x

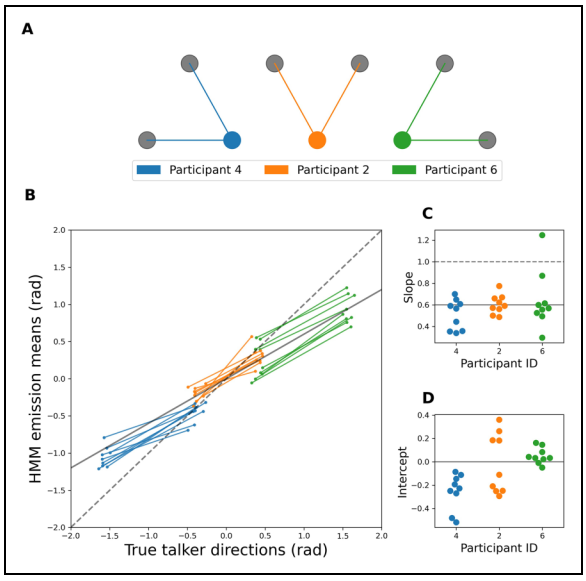


**Figure 3.** The true talker directions and the means of HMM emission functions. Panel A shows the relative location from three participants involved in the 4-people session. Panel B shows the relationship between true talker directions and HMM emission means. Each line represents one participant. The dashed line represents a perfect linear relationship without undershoot, and the solid gray line represents the linear relationship with undershoot from previous studies. Panel C and D show the slope and intercept of the linear relationship in panel B.

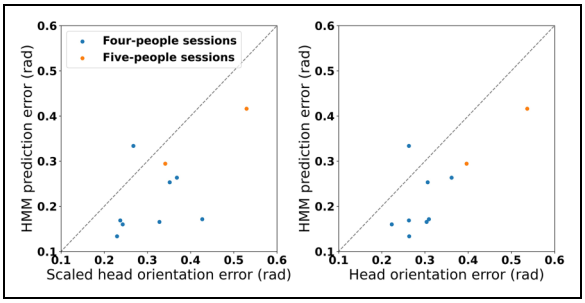


**Figure 4.** The relationship between average error of HMM predictions and head orientation. Each dot represents one participant at location 2 in Figures 1 & 2. The left panel shows the comparison between HMM prediction and head orientation scaled by 1.6, and the right panel shows the comparison between HMM prediction and raw head orientation. There were two possible target talkers in 4-people sessions and three possible target talkers in 5-people sessions.

ranging from 1 to 25 min. The model fitted on truncated data was then used to predict the location of the target talker through the entire session, and the prediction was evaluated against ground truth.

All pre-processing and analysis were conducted in Python 3.9.4. The hidden Markov models were constructed with the Python library hmmlearn (https://github.com/hmmlearn/hmmlearn). Statistical analysis was conducted in R 4.1.0 (Team R Core, 2018). The linear mixed model was fitted with the nlme R package (Pinheiro et al., 2019) and pairwise comparisons were conducted with the lsmeans R package (Lenth, 2016).

## Results

### *True Location and Head Orientation*

The head movements of 3 participants involved in the 4-people sessions were separately analyzed and their distribution from one representative session is shown in Figure 2. All three participants' head orientations mostly fell between the true locations of the other two participants, suggesting the previously described undershoot trend. The distribution of head movements approximated a bimodal distribution.

The means of HMM emission functions were computed for each participant in 4-people sessions and are shown alongside true talker directions in Figure 3. To test if these natural conversation undershoot trends were consistent with those found in lab settings, the slope and intercept of the linear relationship between head orientation and talker direction were extracted. This linear relationship for participants sitting in different locations (participants with ID 2, 4, and 6 in Figure 1) was compared across talker configurations to evaluate its consistency. A one-way ANOVA on the slopes showed no significant main effect of participant location ($F_{2,24} = 1.85$, $p = .32$). The slopes for all participant locations were significantly smaller than 1 ($p < .0001$ for all), suggesting significant undershoot of head orientation. There was no significant difference between the slopes and the previously published measured slope of 0.6 in lab settings ($p > .16$ for all). A one-way ANOVA on the intercepts, however, showed significant main effect of participant locations ($F_{2,24} = 7.39$, $p = .003$). The intercept of participants at location 4 was significantly different from 0 ($t_{24} = 4.31$, $p = .0002$), while no significant difference from 0 was found for participants at location 2 and 6 ($p > .4$ for both). Pairwise comparison showed that the intercept of participants at location 4 is significantly lower than that of participants at location 2 ($t_{24} = 3.65$, $p = .004$, p value corrected with Tukey's method) and that of participants at location 6 ($t_{24} = 2.87$, $p = .022$, p value was corrected with Tukey's method). No significant difference was found between the intercept of participants at location 2 and location 6 ($t_{24} = 0.78$, $p = .72$, p value corrected with Tukey's method).

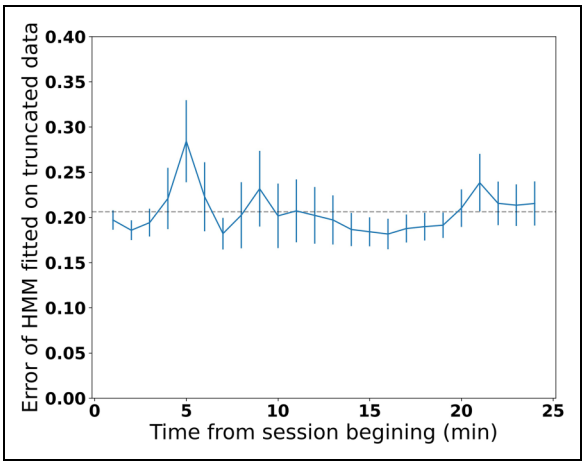


**Figure 5.** The group mean of error of HMMs fitted on first x-min of head movement in predicting target talker location through the entire 4-people recording sessions. The error bar represents the standard error of the group mean. The horizontal dashed line represents the group average of error of HMMs fitted on the entire session.

To evaluate if the fitted HMMs provide benefit over the traditional head-steering method, the error in predicting target location was analyzed for HMM and two versions of head-steering methods. The first version of the head-steering method used the head orientation scaled by a constant 1.6 as previously reported in lab settings (Brimijoin et al., 2010), and the second version of the head-steering method used raw head orientation. When there were two possible target talkers (4-people sessions), the error of HMM prediction was found to be significantly lower than that of the head orientation scaled by 1.6 ($t_7 = 3.13$, $p = .016$) and lower than that of raw head orientation ($t_7 = 3.28$, $p = .013$). Both of the two 5-people sessions suggested that HMM prediction had lower error than head-steering methods, but no formal statistical conclusion can be drawn due to the small sample size (n = 2).

To test the usability of HMMs in real-time situations, the error of HMM fitted on truncated head movement data from 4-people sessions was calculated as shown in Figure 5. As shown in the figure, the error remained flat as the size of training data increased. A peak was seen at approximately 5 min after the beginning of recording sessions. The potential reason can be that participants were asked to read a printed menu and order food at that time, and their head movements may not have been well correlated with the target of their auditory attention. The HMMs fitted on head movement during the first few minutes of group conversation showed performance similar to the HMMs fitted on the data from entire sessions, suggesting that a simple HMM approach may be a viable method of source selection over the course of a conversation in realistic situations.

## Discussion

### Relationship between Auditory Target Location and Head Orientation

People tend to turn their heads towards the active talker in a conversation. The relationship between head angle and active talker was shown here to be approximately linear, with a slope consistent with that found previously in lab-based head orientation studies (Brimijoin et al., 2010). All things being equal therefore, one could propose a sound source selection system that operated simply by choosing the source closest to 1.6 times the listener's head angle. The problem with this slope-based approach is that the intercept of the linear relationship varies across different target talker layouts (Figure 3). It is probable that this intercept is at least partially related to the orientation of the torso, since the undershoot observed in our dataset is similar to that found in fixed seated orienting tasks (e.g., Brimijoin et al., 2010) where the torso was fixed in the same orientation as the chair and all sound sources are distributed evenly around the listener. Previous work has showed that a conversing group naturally forms an empty space surrounded by the participants involved, termed the o-space (Kendon, 1990), so the EasyCom participants could also naturally change their torso direction towards the location of other participants involved in the conversation. Unfortunately, torso orientation was not captured in the EasyCom database, and previous studies (e.g., Brimijoin et al., 2010) did not systematically alter seating / torso angle, so we are unable to disentangle whether the observed difference in intercepts were due to talker layout, differences in torso orientation, or some interaction between the two. Further study on head and body motion tracking data in group conversation may help to better profile head movement behavior in conversation and shed light on this puzzle. If the torso position is indeed the cause of inconsistent intercepts, the performance of the linear model could be greatly improved using other sensors, e.g., egocentric camera (Jiang & Grauman, 2017), to detect head-on-torso angle. For our head-angle-only approach, however, we can say that although the pattern of head undershoot appears to be consistent and to remain largely unchanged from seated orientation experiments to natural conversation, dependence of the intercept on talker layout and/or torso orientation leads us to conclude that head movement alone is unsuitable to be used to predict the true location of the auditory target.

### Source Selection Based on a Hidden Markov model

Since predicting real-time target location purely based on head orientation is challenging, a more reasonable choice may be a two-step model that includes source registration and source selection. First the possible target locations are established based on information from sensors, e.g., camera and mic array, and then the current attended target is selected based on measured user behavior, e.g., head movement. We proposed here a source selection model based on an HMM that converts real-time head movements to a prediction of the target of a listener's auditory attention. The proposed model performed better than pure head-steering methods and appears to be able to generalize from 3 to 4-people group conversations (Figure 4). Also, it was shown that HMMs fitted on head movement quickly converged within the first few minutes of group conversation (Figure 5), meaning that such a model implemented on an actual assistance device could quickly become reliable over the course of even a short conversation. As the target location in group conversation is relatively stationary, a general estimate of the target location is sufficient to convert the decoded hidden states to target direction.

In addition to group conversation, our HMM-based source selection method can also be generalized to other situations. The model only requires that the locations of targets be relatively fixed, so any type of fixed discrete sound source can be selected. Furthermore, the output from our proposed model could also be combined with other information to provide a more nuanced prediction of users' auditory attention. For example, eye tracking data could be combined with head tracking data to provide extra information, as head movement and eye gaze have been shown to be only weakly correlate in group conversation (Vrzakova et al., 2016), suggesting there is additional information that could be utilized. The input from other sensors could also be fused to cross-validate each other, so the HMM could be better tuned. For example, estimating the number of potential auditory targets could be greatly improved by including face tracking data, estimates of the number of clusters of head orientation, and an analysis of eye fixation clusters across a group.

### Caveats

As the ground truth of auditory attention in our experiment was defined through manual annotation, it may not always perfectly reflect the intent of participants during experiment. Even the identification of the onsets and offsets of speech segments are prone to annotator error. Furthermore, when there was only one active talker, it was assumed that the participant was always listening to the only active talker, which may also not be true in practice if, for example, the participant was distracted or attending to a non-auditory target. Another possible situation where the HMM is likely to fail is when head movement was correlated with other tasks instead of group conversation. This can be seen in Figure 5, where the error increased at approximately 5-min after the beginning of group conversation when participants were asked to read menu and order food. In practice, the Markov assumption does not hold for human motion due to the continuous nature of body movement. Involving head movement in other axes and adding an autoregressive component to represent this continuity may further improve

the accuracy of HMM but will also add extra parameters that must be estimated. Previous research has shown that head movement in a one-on-one interaction can be described through a vector autoregressive model including yaw, pitch and roll movement (Chen et al., 2020). In addition, in each session only one participant's manually annotated attention was used in evaluation of model performance, as the EasyCom dataset only included egocentric video from one participant. The generalizability of the HMM on participants at different locations remains to be evaluated. Finally, it should be noted that the behavioral data was collected from participants who were not provided with beamforming hearing aids or simulated beamforming audio. Since it has been demonstrated that participant's behavior in group conversation is affected by the use of simulated beamformers (Hládek et al., 2019), the accuracy of a source selection algorithm that relies on this behavior will inevitably be impacted by the assistance itself. Future studies must tackle the behavioral interactions that will result when the loop is closed between source selection and source enhancement.

## Conclusion

Although in lab settings the relationship between head orientation and target location can be reasonably well-described using only a linear model with fixed parameters, our analysis on head movement in natural group conversation showed that the intercept of this linear relationship depended on the location of other talkers (and potentially the orientation of the listener's torso). This dependency of intercept on talker layout causes problems for a hearing assistance device employing a linear model of source selection that is based solely on head movements, likely incorrectly identifying the attended talker when the talkers are not spaced evenly around the listener. The extension we proposed - using an HMM with approximate locations of all talkers estimated from other sensors - was shown to be significantly more capable of correctly identifying an attended talker than linear approaches and may be a promising means of performing source selection on real devices in real world settings.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

This research was funded by Meta.

### ORCID iD

Hao Lu https://orcid.org/0000-0001-5478-9244

### References

Aghaei M., Dimiccoli M., & Radeva P. (2016). With whom do I interact? Detecting social interactions in egocentric photo-streams. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2959–2964. https://doi.org/10.1109/ICPR.2016.7900087

Bentler R. A. (2005). Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, *16*(07), 473–484. https://doi.org/10.3766/jaaa.16.7.7

Biesmans W., Das N., Francart T., & Bertrand A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412. https://doi.org/10.1109/TNSRE.2016.2571900

Bleichner M. G., Mirkovic B., & Debener S. (2016). Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. *Journal of Neural Engineering*, *13*(6), 066004. https://doi.org/10.1088/1741-2560/13/6/066004

Brimijoin W. O., McShefferty D., & Akeroyd M. A. (2010). Auditory and visual orienting responses in listeners with and without hearing-impairment. *The Journal of the Acoustical Society of America*, *127*(6), 3678–3688. https://doi.org/10.1121/1.3409488

Brimijoin W. O., Whitmer W. M., McShefferty D., & Akeroyd M. A. (2014). The effect of hearing aid microphone mode on performance in an auditory orienting task. *Ear & Hearing*, *35*(5), e204–e212. https://doi.org/10.1097/AUD.0000000000000053

Chen M., Chow S.-M., Hammal Z., Messinger D. S., & Cohn J. F. (2021). A person-and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics. *Multivariate Behavioral Research*, *56*(5), 739–767. https://doi.org/10.1080/00273171.2020.1762065

Chuk T., Chan A. B., Shimojo S., & Hsiao J. H. (2020). Eye movement analysis with switching hidden markov models. *Behavior Research Methods*, *52*(3), 1026–1043. https://doi.org/10.3758/s13428-019-01298-y

Donley J., Tourbabin V., Lee J.-S., Broyles M., Jiang H., Shen J., Pantic M., Ithapu V. K., & Mehra R. (2021). EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments. https://arxiv.org/abs/2107.04174v1.

Eddy S. R. (2004). What is a hidden markov model? *Nature Biotechnology*, *22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Favre-Félix A., Graversen C., Hietkamp R. K., Dau T., & Lunner T. (2018). Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment. *Trends in Hearing*, *22*, 233121651881438. https://doi.org/10.1177/2331216518814388

Gatehouse S., & Noble W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, *43*(2), 85–99. https://doi.org/10.1080/14992020400050014

Grimm G., Kayser H., Hendrikse M., & Hohmann V. (2020). A gaze-based attention model for spatially-aware hearing aids. *Speech Communication - 13th ITG-Fachtagung Sprachkommunikation*, 231–235. https://ieeexplore.ieee.org/abstract/document/8578029.

Guitton D., & Volle M. (1987). Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of Neurophysiology*, *58*(3), 427–459. https://doi.org/10.1152/jn.1987.58.3.427

Hart J., Onceanu D., Sohn C., Wightman D., & Vertegaal R. (2009). The attentive hearing aid: Eye selection of auditory sources for hearing impaired users. *Human-Computer Interaction - INTERACT 2009, 12th IFIP TC 13 International Conference*, *5726 LNCS(PART 1)*, 19–35. https://doi.org/10.1007/978-3-642-03655-2_4

Hendrikse M. M. E., Llorach G., Hohmann V., & Grimm G. (2019). Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. *Trends in Hearing*, *23*, 233121651987236. https://doi.org/10.1177/2331216519872362

Hládek Ľ, Porr B., Naylor G., Lunner T., & Owen Brimijoin W. (2019). On the interaction of head and gaze control with acoustic beam width of a simulated beamformer in a two-talker scenario. *Trends in Hearing*, *23*, 233121651987679. https://doi.org/10.1177/2331216519876795

Jiang H., & Grauman K. (2017). Seeing invisible poses: Estimating 3D body pose from egocentric video. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 3501–3509. https://doi.org/10.1109/CVPR.2017.373

Kendon A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press.

Kidd G. (2017). Enhancing auditory selective attention using a visually guided hearing aid. *Journal of Speech, Language, and Hearing Research*, *60*(10), 3027–3038. https://doi.org/10.1044/2017_JSLHR-H-17-0071

Kim J., Singh S., Thiessen E. D., & Fisher A. V. (2020). A hidden Markov model for analyzing eye-tracking of moving objects: Case study in a sustained attention paradigm. *Behavior Research Methods*, *52*(3), 1225–1243. https://doi.org/10.3758/s13428-019-01313-2

Lenth R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Lu H., McKinney M. F., Zhang T., & Oxenham A. J. (2021). Investigating age, hearing loss, and background noise effects on speaker-targeted head and eye movements in three-way conversations. *The Journal of the Acoustical Society of America*, *149*(3), 1889–1900. https://doi.org/10.1121/10.0003707

Maruyama N., Hiraguri Y., Kawai K., & Ueda M. (2020). Assessing the ease of conversation in multi-group conversation spaces: Effect of background music volume on acoustic comfort in a café. *Building Acoustics*, *27*(2), 137–153. https://doi.org/10.1177/1351010X19897232

McKellin W. H., Shahin K., Hodgson M., Jamieson J., & Pichora-Fuller K. (2007). Pragmatics of conversation and communication in noisy settings. *Journal of Pragmatics*, *39*(12), 2159–2184. https://doi.org/10.1016/j.pragma.2006.11.012

Moore A. H., de Haan J. M., Pedersen M. S., Naylor P. A., Brookes M., & Jensen J. (2019). Personalized signal-independent beamforming for binaural hearing aids. *The Journal of the Acoustical Society of America*, *145*(5), 2971–2981. https://doi.org/10.1121/1.5102173

Nützi G., Weiss S., Scaramuzza D., & Siegwart R. (2011). Fusion of IMU and vision for absolute scale estimation in monocular SLAM. *Journal of Intelligent & Robotic Systems*, *61*(1-4), 287–299. https://doi.org/10.1007/s10846-010-9490-z

Otsuka K., Yamato J., Takemae Y., Murase H., & Yamato J. (2005). A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. *Proceedings of the 7th International Conference on Multimodal Interfaces*, 191–198. https://doi.org/10.1145/1088463.1088497

Pinheiro J., Bates D., Debroy S., Sarkar D., & Authors E., & R-core. (2019). Linear and non-linear mixed effects models. Package "nlme", version: 3.1-141. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/nlme/index.html.

Simola J., Salojärvi J., & Kojo I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, *9*(4), 237–251. https://doi.org/10.1016/j.cogsys.2008.01.002

Stiefelhagen R., & Zhu J. (2002). Head orientation and gaze direction in meetings. *CHI, 02, Extended Abstracts on Human Factors in Computer Systems*, *1*, 858. https://doi.org/10.1145/506621.506634.

Strawbridge W. J., Wallhagen M. I., Shema S. J., & Kaplan G. A. (2000). Negative consequences of hearing impairment in old age: A longitudinal analysis. *The Gerontologist*, *40*(3), 320–326. https://doi.org/10.1093/geront/40.3.320

Team R Core (2018). R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Tourbabin V., Donley J., Rafaely B., & Mehra R. (2019). Direction of arrival estimation in highly reverberant environments using soft time-frequency mask. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, *2019-Octob*, 383–387. https://doi.org/10.1109/WASPAA.2019.8937233

Vrzakova H., Bednarik R., Nakano Y. I., & Nihei F. (2016). Speakers' head and gaze dynamics weakly correlate in group conversation. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, *14*, 77–84. https://doi.org/10.1145/2857491.2857522