Integrating Gaze and Speech for Enabling Implicit Interactions

Anam Ahmad Khan
The University of Melbourne
Melbourne, VIC, Australia
anamk@student.unimelb.edu.au

Joshua Newn
The University of Melbourne
Melbourne, VIC, Australia
joshua.newn@unimelb.edu.au

James Bailey
The University of Melbourne
Melbourne, VIC, Australia
baileyj@unimelb.edu.au

Eduardo Velloso
The University of Melbourne
Melbourne, VIC, Australia
eduardo.velloso@unimelb.edu.au

ABSTRACT

Gaze and speech are rich contextual sources of information that, when combined, can result in effective and rich multimodal interactions. This paper proposes a machine learning-based pipeline that leverages and combines users' natural gaze activity, the semantic knowledge from their vocal utterances and the synchronicity between gaze and speech data to facilitate users' interaction. We evaluated our proposed approach on an existing dataset, which involved 32 participants recording voice notes while reading an academic paper. Using a Logistic Regression classifier, we demonstrate that our proposed multimodal approach maps voice notes with accurate text passages with an average F_1-Score of 0.90. Our proposed pipeline motivates the design of multimodal interfaces that combines natural gaze and speech patterns to enable robust interactions.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **Sound-based input / output**.

KEYWORDS

implicit annotation, natural gaze, voice interfaces, natural language processing, semantic similarity

ACM Reference Format:
Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3502134

1 INTRODUCTION

"Of central interest is how voice and gesture can be made to interorchestrate, actions in one modality amplifying, modifying, disambiguating, actions in the other. The approach involves the significant use of pronouns, effectively as "temporary variables" to reference items on the display."

— R. Bolt, *Put-that-there: Voice and gesture at the graphics interface*, 1980 [7]

Research in multimodal human-computer interaction has long been aware of how the limitations of one input modality can be complemented by the strengths of another. In "Put-that-there", the seminal work quoted above, Richard Bolt's team demonstrated how pointing with a hand gesture can provide the deictic information necessary to resolve ambiguities in speech commands, while voice provides the semantic information that enriches the meaning of the pointing gesture.

With the increased availability of affordable eye trackers, gaze has also been explored as an alternative to hand gestures for complementing speech (e.g. [6, 15, 36]). These works usually employ the gaze vector to disambiguate the target of a speech command, either by implicitly inferring the object of the user's attention or by explicitly requiring the user to direct their eyes as a pointer. However, in these works, speech tends to be used as an explicit modality, in the form of clear voice commands. In this kind of interaction, users are more likely to exhibit behaviours that differ from their natural speech patterns [32]. For example, they are more likely to simplify what they say ("Play Beatles" vs "Can you please play The Beatles?"), hyperarticulate ("A-LE-XA, SET-A-TI-MER"), and pause when using deixis ("Put that <pause> there <pause>") [7, 28]. To enable truly natural multimodal interaction with speech, we need methods for integrating these modalities that are robust to natural speech patterns. Although natural speech creates challenges for voice interfaces—such as the extensive vocabulary, the lack of a pre-defined set of commands, and longer sentences—the extra information embedded in speech offers an additional opportunity for extracting meaning from users' utterances and enriching the interaction.

In a similar way, the role of gaze in multimodal applications that combine eye tracking and voice interfaces has also been limited. Because the eyes generally point at what draws users' attention, gaze is frequently used as a deictic input modality, to indicate the object in the interface to which the speech command refers [8, 40]. However, eye movements offer much more beyond pointing. They provide a window into the users' mind, allowing conclusions to be drawn regarding users' intentions, goals, and cognitive processes [4, 10]. For instance, work on context and activity recognition from gaze has demonstrated that it is possible to infer a wide variety

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502134

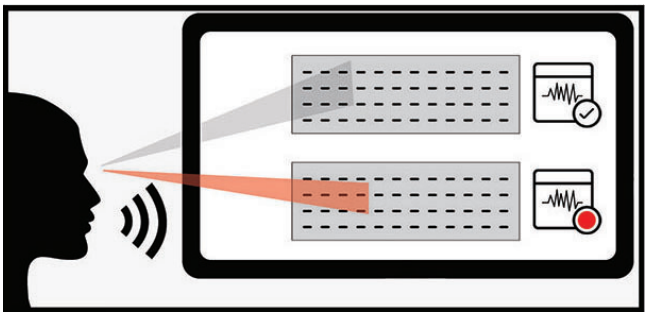

Figure 1: Implicit anchoring of voice notes to digital content by leveraging users' natural gaze and speech patterns

of contextual information from this data including users' decision-making strategies [14] and the covert aspects of users' states [10]. In the same way that the additional information offered by natural speech in comparison to voice commands can enrich interaction techniques, so can the contextual information available in natural gaze data in comparison to simple gaze pointing.

In this work, we explore how to integrate natural gaze and voice behaviours to support implicit interaction. We argue that the synchronicity between speech utterances and eye movements can be leveraged and combined with semantic knowledge extracted from the content being looked at to better understand the context of the interaction. We demonstrate our approach in the challenge of anchoring voice notes to text as the user voices their thoughts while reading it, as shown in Figure 1. This is a task that involves natural gaze behaviour (as the user reads the text), natural speech (as the user voices their comments), knowledge of the stimulus (as the content of the text influences both gaze and speech behaviours), and is representative of a broader class of tasks where the challenge lies in establishing the relationship between behaviours exhibited in different modalities. We propose a machine learning-based pipeline for integrating gaze and speech that extracts (1) *lexical* features that capture the semantic relevance of users' uttered notes with the read text passages, (2) *gaze-based* features that capture users' global gaze pattern while they recorded a voice note and (3) *reference-informed* features that leverage the synchronicity between gaze and speech data to that help identify items on display referred by the user while uttering a voice note. We evaluated our proposed approach on a dataset, which involved 32 participants recording voice notes while reading an academic paper. Using a Logistic Regression classifier, we demonstrate that our proposed multimodal approach maps voice notes with accurate text passages with an average F_1-Score of 0.90, outperforming existing anchoring approach [19].

In summary, this paper takes the first steps towards exploring the implicit design space of *gaze+speech* for anchoring voice notes on digital content. We contribute (1) a novel multimodal pipeline of combining gaze and speech input modality for the task of implicit annotations of digital content with vocal utterances; (2) demonstrate how both the user's visual attention and the content of the voice note reveal to which digital content a vocal utterance is related; (3) present a classifier trained on gaze and speech features of 32 participants that maps voice utterances to text regions with an F_1-Score of 0.90. Our findings lay a necessary and important foundation for designing systems that harness users' behaviours in a holistic and well-integrated manner.

2 RELATED WORK

Gaze and voice-based multimodal interactions have been the subject of research for many decades. We first position the related work by reviewing literature that leverages input from gaze and speech modalities to derive rich multimodal interaction techniques. We then briefly discuss prior work using machine learning techniques to model and understand users' gaze and speech interaction patterns. Finally, we describe the task of anchoring users' vocal utterance to relevant content during reading.

2.1 Multimodal interaction using gaze and speech

The combination of gaze and speech for enabling new interaction techniques has been well explored in HCI research. The underlying foundation of this research is that while both input modalities are prone to errors, their combination can resolve each other's weakness, resulting in robust and expressive interaction [26, 41]. Most multimodal interaction techniques explored to date use data from these modalities as explicit input [39, 41]. In such applications, interaction is often facilitated by using gaze as an explicit pointing device and voice as a means to communicate the users' intent. For instance, van der Kemp and Sundstedt proposed a hands-free drawing system where voice commands are used to control the interface and gaze is used for positioning the cursor [41]. Although such gaze and voice-enabled applications provide rapid interactions, they require conscious effort from the user to provide input, either by requiring them to direct their eyes at specific interface elements and/or by restricting their vocabulary to a limited set of commands.

To enable more natural interactions, a line of research has attempted to use gaze behaviour as an implicit source of information to facilitate voice-based interactions [15, 17, 38]. For instance, researchers have attempted to use natural gaze behaviour to overcome the challenges of reference resolution in spoken utterances [38]. Reference resolution is the process of identifying the item displayed on screen referred by the user in a spoken utterance [33]. Voice interfaces often fail to understand user's commands as they are unable to resolve ambiguous references (e.g., *"colour **this** green"*) made to objects displayed on the screen. As gaze offers a window into users' visual attention, it is often leveraged to provide the contextual information needed for reference resolution in spoken dialogues [17, 33, 42]. For instance, Hakaani et al. presented a conversational web system that employs gaze patterns to resolve ambiguous references in the spoken commands made by users while web browsing [15].

Even though prior work shows that the implicit use of gaze can enrich users' interactions experience with voice interfaces, these works still use voice as a means to communicate users' intent. However, speech is also a context-rich input modality that embeds acoustic and linguistic information regarding users' behaviour and their environment. Thus, the extra information embedded in speech can be leveraged and combined with natural gaze behaviour to make users' interactions more accurate with voice interfaces. Hence, our work contributes to this research line by exploring how visual

attention extracted from users' gaze behaviour can be combined with semantic information embedded in their vocal utterances to enable accurate interactions in a reference resolution challenge where natural gaze and speech patterns are unavoidable, namely, implicitly anchoring voice notes to digital content.

2.2 Multimodal classification using gaze and speech

Combining gaze and speech data streams to model and understand users' behaviour is not a trivial task. For this purpose, previous works have leveraged machine learning models for understanding users' interaction patterns for various scenarios [3, 16, 21]. For instance, in the field of affective computing, researchers have attempted to predict whether users suffer from depression by combining features from users' gaze and speech behaviour during natural communication [3]. In line with prior work, we also leverage machine learning techniques to combine natural gaze and speech behaviour to facilitate users' interactions. We contribute to this body of work by offering a pipeline that not only analyses users' natural behaviours but also shows how they relate to the stimulus. Specifically, we integrate the analysis of the gaze data time series, the natural speech patterns, and the content being looked by the user to build a classification model for facilitating users interaction with multimodal interfaces.

2.3 Anchoring of voice notes to digital content

Voice annotations are a means for users to express ideas, identify problems, and suggest edits while consuming digital content [18, 37]. Despite its benefits, voice annotations suffer from the challenge of *anchoring*, which relates to the mapping between voice notes and the digital regions to which they refer. As a result, applications that support voice annotation generally rely on users explicitly selecting the regions of the content to which the system should attach a voice note. Although the manual selection of regions for anchoring voice notes gives users complete control over the anchoring process, the task can pose challenges for the user, particularly if a recorded voice note is to be attached to more than one region. Not only this is procedurally difficult for users, but it also requires them to switch their attention form the main task of reading to the secondary task of manual anchoring, which results in breaking the task flow.

Implicit anchoring approaches attempt to overcome the challenges in manual selection by leveraging input from natural human communication channels (e.g., gaze, speech, gesture). Gaze—the modality that best reflects users' visual attention—has been explored in recent work for implicitly creating voice annotations on the digital text [19]. The authors leveraged user's natural gaze patterns and machine learning techniques to propose an approach that automatically maps users' uttered notes to relevant text regions. Although their work demonstrates the general feasibility of gaze for creating annotations, it highlights that the gaze-based approach does not work in specific, but common, scenarios, such as when the user skims the relevant text region while commenting on it and whenever similar gaze patterns are observed in two adjacent text regions. This observation highlights the need to increase the robustness of the gaze-based implicit anchoring to enhance users' interaction experience.

Further, the gaze-only approach fails to leverage the fact that the content of users' utterances is often semantically related to the content of the stimulus, which can further assist the disambiguation task. We argue that this fact can best be leveraged by combining both gaze and speech for two reasons. First, while uttering voice notes, users often look directly at the focus of their interest. Hence, their speech data often include referential terms (e.g. *"**This** is important for the analysis"*), which explicitly refer to the items of interest [15, 27]. Second, while uttering notes, users often tend to comment about the content they are reading [29, 31]. Hence, the note's content can have semantic and syntactic relevance to the text regions regarding which a note is made. Therefore, in our work, we perform a combined analyses of user's natural gaze and speech behaviour to provide a more thorough understanding of users' context, resulting in a robust implicit anchoring of voice notes to digital text.

3 IMPLICIT ANCHORING OF VOICE NOTES USING GAZE AND SPEECH

The main contribution of this work is a machine learning-based data processing pipeline that facilitates the implicit anchoring of voice notes. Our pipeline systematically combines gaze and speech data streams to build a prediction model that indicates the relevant region for anchoring notes. For this purpose, we employed the dataset collected by Khan et al. [19] that contains gaze and speech data (details described in Section 3.1) of users while reading and recording voice notes on digital documents.

This section first summarises the dataset used for building our multimodal pipeline. Following, we detail our pre-processing steps applied to the dataset to extract the relevant input to our proposed multimodal pipeline. Finally, we describe our exploratory analysis of the dataset to understand the different natural gaze and speech patterns observed during note-taking to inform our machine learning-based pipeline.

3.1 Dataset description

The dataset collected by Khan et al. [19] for the task of creating implicit annotation on digital documents contains data from 32 participants. Full details about the procedure can be found in the original ToCHI article [19]. All participants were PhD students aged between 27–42 years (M= 31, SD= 3.6). Eighteen of them were men, and fourteen were women. Participants in the dataset were diverse in terms of their native languages (3 Arabic, 3 Tamil, 2 Mandarin, 4 English, 2 Urdu, 2 Bengali, 7 Persian, 4 Sinhala, 2 Yoruba, 1 Bahasa speaker).

The data collection study was conducted in a controlled laboratory setting using a multisensor experimental desktop setup. The main task of the study consisted of reading a research paper in the custom-built PDF viewer as if it was part of a literature review while making voice notes about its content. Participants were further instructed only to speak when they wanted to make a voice note. Upon finishing the main task, the researcher collected the ground truth data by requesting participants to highlight the text passages in the research paper for each corresponding voice note they made.

The dataset contains participants' raw gaze coordinates, the raw audio recordings of their speech captured throughout the study,

the changes in the visual stimuli, depicted through the scrollbar value and page number of the document being read. Further, the dataset contained manually annotated reference text passages for all recorded voice notes, which later served as the ground truth of the proposed machine learning models.

3.2 Data pre-processing

We processed the raw data in two stages in order to extract the individual voice notes, the accompanying gaze trace and the textual passages that were read by the user while uttering a voice note.

3.2.1 Extracting voice notes. To extract voice notes from the raw audio data, we removed silent segments filtering out the audio below 26dB and discarding audio segments of less than 3s. As a result, we obtained 691 separate voice notes (mean: 22 per participant). Further, we used the Rev-AI python API [1], a speech-to-text recognition library, to obtain the automatic timestamped text transcription of the retrieved recorded voice notes. To better understand the effect of the speech-to-text accuracy on the anchoring performance, we also transcribed the audio notes manually using a transcription service.

3.2.2 Extracting candidate text passages. We extracted and pre-processed the visual stimuli to capture the text passages read by the participants. For this task, we defined a Region-of-Analysis (ROA) for each voice note which started from the end time of a voice note to the end time of the successive voice note under consideration. We mapped the voice notes to their reference passage by only analysing the gaze patterns and the textual data of the document that lies in the specified ROA of the voice note. For each voice note, we first extracted the image data frames which lie within the specified ROA by using the page navigation and scroll bar value recorded during data collection. We then used the PyTessBaseAPI [2] python library to segment the extracted images into text paragraph. We saved the textual content and the geometrical coordinates of the extracted paragraph for later analysis. The extracted text passages served as the candidate passages to which a voice note would be mapped.

This data pre-processing steps resulted in a set of 691 transcribed voice notes and their corresponding candidate text passages, along with the accompanying gaze data that lied within the specified ROA of the voice note.

3.3 Understanding gaze and speech patterns for voice notes

This work is motivated by the belief that the combined analysis of gaze data, speech data, and the task stimulus offers novel opportunities to support implicit interactions. To validate this assumption, we conducted an exploratory analysis of the dataset. Our goal was to understand the different kinds of behaviours that users engaged in while making voice notes. This insight would help us build a multimodal pipeline that best capture the relationships between the different modalities. From the exploratory data analysis, we observed that users make different kinds of voice notes which could be broadly split into three different categories: *content-related*, *reflective*, and *implicitly referenced*. We have included a representative example of each voice note category in the supplementary material.

(1) *Content-related*: These were long voice notes in which participants reflected on their understanding by summarising the content they had just read. When participants are creating this type of note, we observed that they often verbalised their understanding by using words present in the text passage to which the voice note referred. Hence, the textual content of these voice notes had high relevance to the text of the referent paragraph. Further, while making content-related notes, participants often fixated on the text passages that elicited the note. Therefore, the depicted gaze patterns were also indicative for inferring the relevant text passage for voice annotation.
(2) *Reflective*: These were short voice notes in which participants reflected upon the research paper by either expressing their own opinion regarding the content—*"I don't agree with what the author is proposing"* (P6)—or by linking the content to the material they had previously encountered. The content of these voice notes usually revealed participants' own sentiments and had no clear connection to what is written in the referent paragraph. Moreover, we observed that while recording this type of note, some participants looked at the content that had elicited the voice note, whereas others fixated on random sections of text. Thus, the depicted gaze patterns may or may not be relevant for predicting the referent text passage for anchoring the recorded note.
(3) *Implicitly referenced*: These were voice notes in which participants made an implicit reference to the content—*"Okay. So, this seems promising."*(P3). This occurred when participants were directly looking at the relevant text passage while recording the voice note. In several instances, instead of explicitly uttering the content to which they were referring, they made an implicit reference using a demonstrative pronoun, such as *these*, *this*, *those*. Depending on the content of the recorded note, the textual data of these voice notes may or may not have relevance with the text of the referent paragraph, but the gaze data is highly relevant in these cases as it could help resolve the ambiguous references in spoken utterances.

Implication: Based on the types of voice notes observed during the analysis, we concluded that both participants' natural gaze and speech patterns contain rich information about text passages that elicited their notes. Participant's' gaze patterns reflect their visual attention that could be used to identify the text passage being read while uttering a note. Similarly, the content of the speech voiced by them also has some semantic and syntactic relevance with the content of the referent text passage. Moreover, we observed that it is also necessary to leverage the interaction between gaze and speech modality as it can help identify important regions on display implicitly referenced by the user while uttering a voice note. Using these findings, we propose a machine learning-based multimodal pipeline that systematically extracts and combines features from both modalities to facilitate robust anchoring of voice notes.

[1] https://pypi.org/project/rev-ai/
[2] https://pypi.org/project/tesserocr/

4 MULTIMODAL PIPELINE

The findings of the exploratory data analysis informed a five-stage multimodal pipeline that extracted 15 region-based features (see Appendix B) and trained a machine learning model to predict text passages for anchoring voice notes implicitly. Figure 3 illustrates the pipeline. The fundamental novelty of this pipeline is the way users' gaze and speech data stream are combined with the stimulus for the prediction task. The pipeline not only extracts discriminative features from each data stream but also engineers features that jointly use the individual data streams to capture rich context information neither modality has in isolation. Using such an approach, we aim to better model users' context for robustly anchoring voice notes. Below we describe the five stages of building the multimodal pipeline.

4.1 Stage 1: Extracting lexical features

The first stage of the pipeline focuses on analyzing the speech data stream to extract lexical features that can capture the semantic relevance between a transcribed voice note and the read text passage. In order to capture these features, we used the *Bidirectional Encoder Representation from Transformers* (BERT) model and cosine similarity metrics [22]. BERT is an *word embedding*, which reflects the structure of the words in terms of their morphology so that semantically similar words are represented by similar vectors [45]. Our design decision of choosing BERT as a vector embedding model was motivated by the high performance of this model for text similarity tasks [34]. We employed the Sentence-BERT pretrained model to obtain vector embeddings for each voice note and candidate text passages. Each text paragraph and a recorded voice note was treated as a sentence and fed into the pre-trained model to obtain a 768-dimensional vector. We then used these vector embeddings to find the cosine similarity between the transcribed note and the candidate text passage. The obtained cosine similarity scores were used to engineer a *semantic relevance rank* feature for each text passage. To calculate this feature, we ordered each text passage based on their cosine similarity score in an ascending manner and used the obtained order as a semantic relevance rank feature for each text passage.

Our exploratory data analysis of the voice notes (Section 3.3) revealed that the notes belonging to the *Reflective* category were generic comments, and as such, their textual content is likely to be unrelated to the candidate text passages. Therefore, the semantic relevance rank feature value calculated for the candidate text passages of these notes would not be useful, and as such, should not influence the classification pipeline. To identify the patterns of these notes, we calculated a p-value for the semantic relevance rank feature of each text passage. The calculated p-value was used as another lexical feature and is engineered in two stages described below.

(1) **Creating a distribution of similarity score**: We created a sampling distribution of cosine scores, which were observed when similarity is measured between unrelated text passages and a transcribed voice note. We obtained the sampling distribution by randomly selecting 100 text passages from a corpus we had prepared. Then, we calculated the cosine similarity between these random text passages and the voice note under consideration. Resultingly, we got a distribution of scores when similarity is measured between unrelated text passage and user's note. An example of this distribution is shown in Figure 2 plot A and B. The intuition behind this is that if the note is just as relevant to the random corpus as it is to the passages in the stimulus, it is likely that the content of the note is not helpful for the classification task.
(2) **Calculating the p-value feature**: After building the distribution of similarity score, we fitted a cumulative distribution function on it to obtain the p-value feature for the observed cosine score of each text passage. The calculation of the p-value feature is detailed in the Appendix B. The p-value feature ranges from 0 to 1 and reflects the likelihood that the observed cosine similarity score is drawn from the score distribution for a user note that is unrelated to the text passage. The feature value closer to 0 indicates that the observed cosine score of a text passage is unlikely to have been drawn from the score distribution calculated for text passages unrelated to the voice note. In contrast, a feature value closer to 1 suggests it is highly likely that the observed cosine score is drawn from the score distribution for text passages unrelated to the transcribed note. This observation is reflected in Figure 2, in which we compare the p-value feature of the text passages for content-related and reflective voice notes. As shown in Figure 2 plot A, for reflective voice note, the cosine scores of the candidate text passages are concentrated around the mean of the obtained sampling distribution. Consequently, the area under the complementary cumulative distribution plot for all candidate text passage is quite large, which results in their p-value being closer to 1 (see Figure 2 plot C). On the other hand, as shown in Figure 2 plot B, for content-related voice notes, the cosine scores of the candidate text passages are concentrated around the right tail of the obtained sampling distribution. This results in candidate text passages having a relatively low area under the complementary cumulative distribution curve, and their p-values are closer to 0 (see Figure 2 plot D).

4.2 Stage 2: Extracting reference-informed features

We expected the lexical features to be useful only when the user makes an explicit reference to the content of the relevant text passage. However, in our exploratory analysis of the voice notes (see Section 3.3), we observed that users often made voice notes in which implicit references were made to data displayed on the screen. For these notes, the lexical features seemingly fail to capture the semantic relevance of the note in relation to the text passages. Therefore, this stage of the pipeline focuses on jointly analysing data from both input modalities to engineer features that can identify users' visual attention while making an implicit reference to content.

For this stage, we first analysed each user note to identify whether an implicit reference was made in the note. To accomplish this, we focused on three demonstrative pronouns; *this*, *these* and *those*. Whenever a note contained these pronouns, we expected that an implicit reference to the content displayed on the screen would be made. Hence, for each voice note, we first engineered a binary

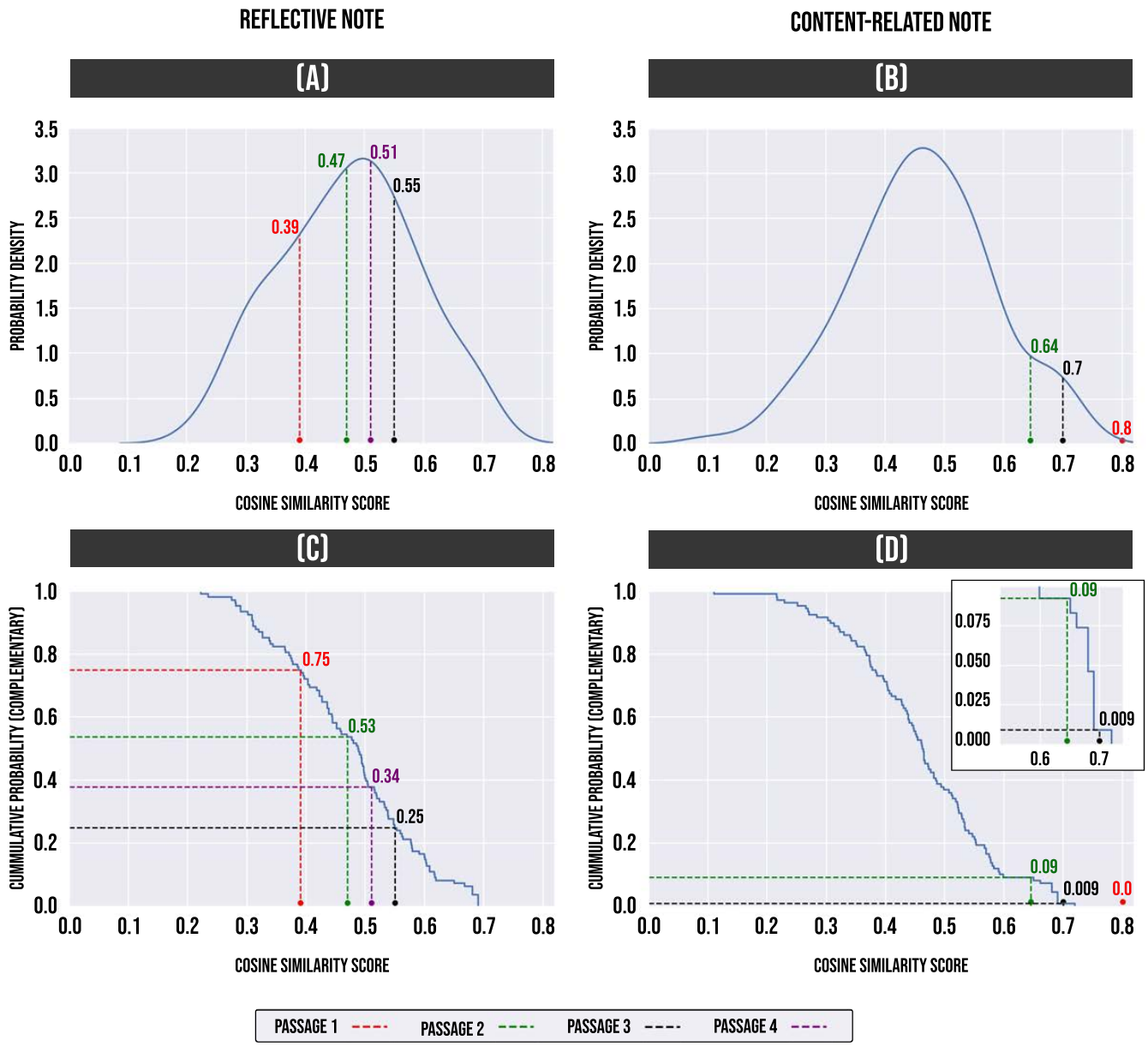

Figure 2: Visualisation of the p-value feature for content-related and reflective voice note. Plot A and B show the probability density curve for the score distribution when similarity is measured between unrelated text passage and a voice note. Plot C and D show the complementary cumulative density plot of the obtained distribution for reflective and content-related voice note, respectively.

feature that indicated whether the user uttered the pronoun or not. If one of these demonstrative pronouns occurred in the note, we analysed the gaze data observed during a window ranging from 2 seconds before the start of the user's utterance of the pronoun to 2 seconds after the end time of the utterance. We clustered the gaze points observed in this window into fixations using the Dispersion-Threshold Identification algorithm [35] by setting the value of dispersion and duration threshold parameter to 20px and 100ms, respectively. After obtaining fixations, we engineered the following three gaze features corresponding to each demonstrative pronoun for each text passage.

(1) Total number of fixations residing in a text passage. This feature reflects how often the user looked at the text passage during the utterance of an implicit reference.

(2) Sum of the duration of fixations residing in a text passage. This feature indicated how long the user looked at the text passage during the utterance.
(3) The minimum distance between the text passage and any gaze fixation point that occurred in the window of analysis. The calculated distance feature represents how far the user was looking from the text passage when uttering an implicit reference. The calculation of this feature is detailed in Appendix B.

4.3 Stage 3: Extracting gaze-based features

This stage of the pipeline focuses on analysing the gaze data stream to extract users' broader gaze behavior while composing voice notes. For this purpose, we leveraged the Annotation classifier proposed by Khan et al. [19], which predicted the text passage for annotating voice notes.

We replicated the proposed annotation classifier to compute a gaze feature for each candidate text passage that indicates the probability of being classified as *Annotated* based on the users' global reading patterns. To calculate this feature, for each voice note, we analysed the gaze data observed in the defined ROA to compute the set of gaze features proposed by Khan et al. [19]. We then fed the computed features into the proposed Annotation classifier to get the probability for the text passage being predicted as Annotated. This feature was fed into our proposed model and is reflective of the broader gaze patterns depicted while users read the digital text and recorded voice notes.

4.4 Stage 4: Model building and evaluation

After engineering various features, we combined them in a machine learning model that could predict referent text passages for anchoring voice notes. We modelled the task of anchoring voice notes as a binary classification problem. For each voice note, a feature vector (lexical, reference-informed, and gaze-based) was fed into a classifier to predict the label of each candidate text passage as either *Annotated* (the voice note corresponded to this text passage) or *Not Annotated* (the voice note did not correspond to this text passage). We labelled the text passages highlighted by the participants as *Annotated* and all other text passages as *Not Annotated*, which served as the ground truth for our model.

To train a machine learning model we experimented with Logistic Regression, SVM and Random Forest models by training them on the dataset of 32 participants described in Section 3.1. As the Logistic regression model gave the highest performance score across the testing folds and does not require extensive hyperparameter tuning, it was selected as the final model for the anchoring task.

As note-taking behaviours differ from user to user, we evaluated our model in two ways. First, we performed a *user-dependent* evaluation using leave-one-note-out cross-validation. For this purpose, we conducted separate evaluations for each participant by building classifiers on the data from the same participant on which it was trained but from a different voice note. For example, if a participant made ten voice notes, we trained the model 10 times, each time training on the nine voice notes and evaluating it on the remaining last voice note. Second, to avoid the risk of overfitting the model to the note-taking behaviour of any particular participant, we performed a *user-independent* evaluation. For this, we trained the classifier 32 times using leave-one-participant-out cross-validation. Each time, we trained the model on the voice notes of 31 participants and then evaluated it on the voice notes of the remaining last participant. The results reported in the next sections are averaged by the total number of participants.

We observed that for each voice note, the number of text passages fed in the classifier labelled as *Annotated* was much smaller than the ones labelled as *Not Annotated* (on average, 10% of text passage were labelled as *Annotated*). This made the employed dataset highly imbalanced. Therefore, we evaluate the classifiers using two metrics commonly used in the literature for reporting the performance of models on imbalanced datasets [13]. First, we used the F_1-Score, which is the harmonic mean of the precision and recall score. Second, we employed the Average Precision (AP) score, which summarises the model's Precision-Recall plot as the weighted average of precision scores achieved at each threshold. For both evaluation metrics, the score ranges between 0 to 1, with 0 depicting the worst and 1 depicting the best value that the model could achieve in terms of its performance.

4.5 Stage 5: Anchoring text passages to voice notes

In the final stage of the pipeline, the results of the classifier were used to recommend text passages for anchoring voice notes. For a particular voice note, the text passages that the model predicted as *Annotated* are mapped to the recorded voice note.

5 RESULTS

In this section, we first report the classification performance of the user-dependent and independent model trained on gaze and speech data. Then, we report that for which type of voice notes our proposed user-independent model exhibits the highest performance. Lastly, we investigate the importance of the features on the performance of the user-independent model trained on the gaze and speech data for the task of anchoring voice notes.

5.1 Model performance on gaze and speech features

We computed the F_1-Score and Average Precision (AP) to evaluate the classification performance of the models shown in Table 1. Further, we report the normalized confusion matrix for each of the classifiers in Figure 4. Our results show that the proposed approach of leveraging features from both gaze and speech modality achieves a high classification performance (F_1-Score = .90) for predicting the relevant text passages regarding which a voice note was recorded. To demonstrate the effectiveness of leveraging multimodal data for anchoring voice notes, we compared the performance of our proposed approach with Khan et al.'s [19] gaze-only approach. The results reported in Table 1 show that for both user-dependent and independent models, our proposed approach of employing gaze and speech data outperformed the gaze-only approach by an average AP score of 12%. This increase in classification performance can be explained by the fact that our pipeline leverages data from both gaze and speech modalities, which captures a different dimension of the user's input. Speech modality captures what the user is speaking,

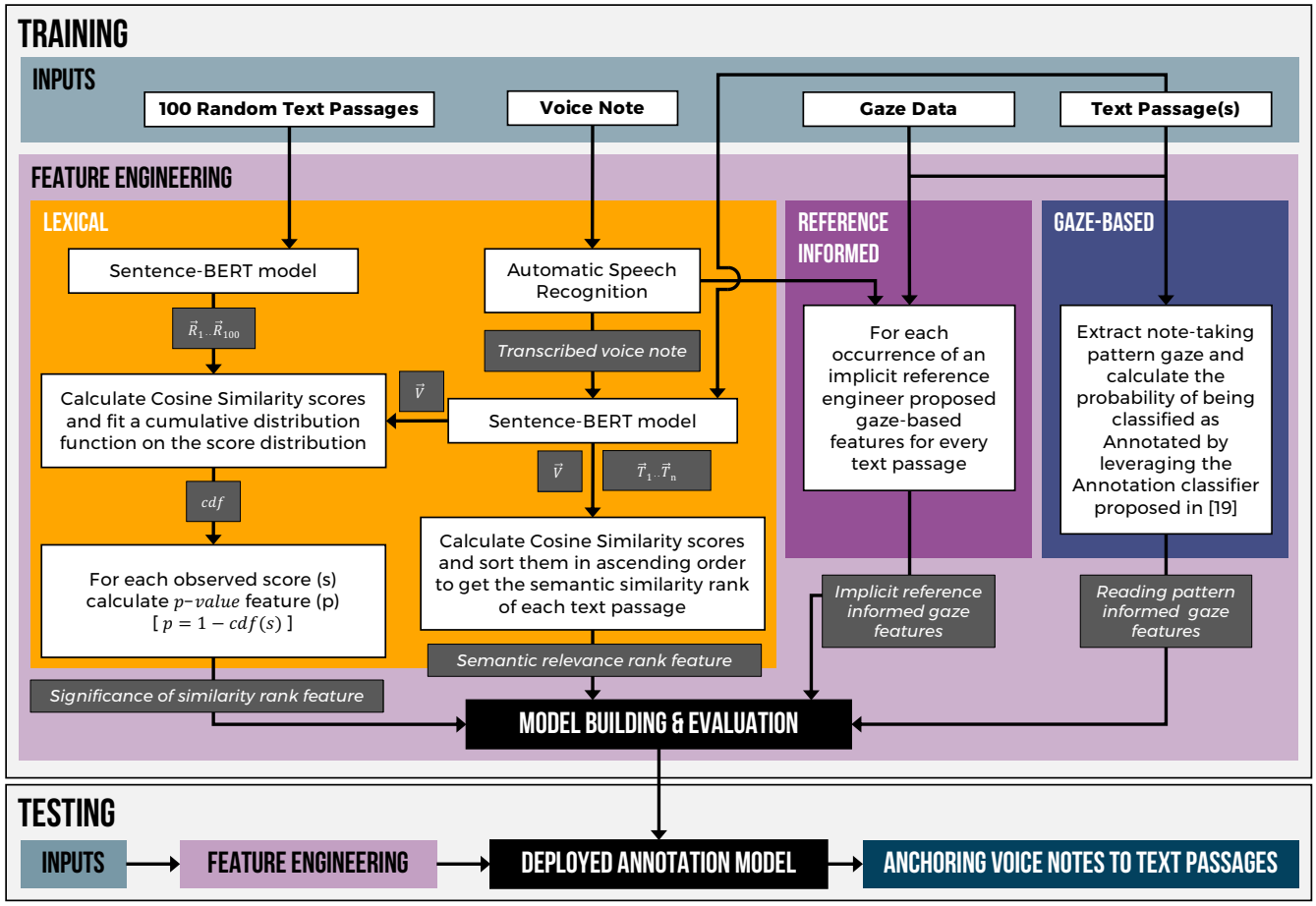

Figure 3: Proposed five-stage machine learning-based pipeline for combining gaze and speech data to implicitly anchor voice notes to text passages. In this figure, $\vec{V}$, $\vec{T}$ and $\vec{R}$ represents the BERT transformed embedding vectors for the transcribed user recorded voice note, candidate text passage lying in the defined ROA of the voice note and arbitrary random text passage extracted for calculating sampling distribution, respectively.

Table 1: Performance of Khan et al. [19] and our proposed approach leveraging ASR and manually transcribed participants' notes for anchoring voice notes.

Classifiers	Evaluation Measures	Khan et al.	Proposed Approach	
			Manual	ASR
User-Dependent	F_1-Score	.88 ± .08	.90 ± .18	.88 ± .16
	Average Precision	.79 ± .16	.88 ± .08	.87 ± .17
User-Independent	F_1-Score	.79 ± .11	.89 ± .02	.87 ±.06
	Average Precision	.68 ± .14	.83 ± .09	.79 ± .10

and gaze depicts the global reading pattern of the user while making voice notes. Additionally, the leveraged synchronicity between gaze and speech input captures the user's visual attention when implicit references are uttered. Thus, when data from both modalities are combined, they complement each other and are more informative for predicting the relevant text passage for voice annotation than leveraging features from just the gaze modality.

Because of the lack of control of natural speech, automatic speech recognition (ASR) algorithms might struggle to perfectly transcribe users' utterances. As such, we evaluated our approach both on the output of ASR and on manually transcribed ground truth data. Table 1 shows a drop in classification performance when the proposed user-dependent and independent models leverage speech features from the ASR transcriptions rather than a manual transcription

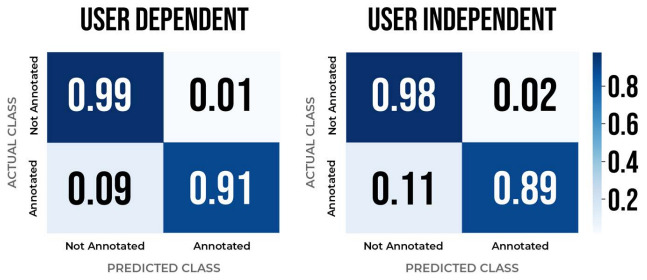

Figure 4: Normalized confusion matrix for user-dependent classifier (Left) and user-independent classifier (Right) trained on the combination of gaze and speech features for the task of anchoring voice notes to reference text passages

Figure 5: SHAP values for top 8 feature indicating the influence of the features on gaze and speech combined model for predicting the labels for text passage as *Annotated*

of the voice notes. The decrease in the performance of the models is because of the use of the ASR transcribed user notes whose correctness is highly dependent on the accent of the speaker and the word vocabulary employed for transcription. To evaluate the performance of the ASR on the recorded users' notes, we used the Word Error Rate (WER), which computes the minimum edit distance between the ground-truth sentence and the ASR transcribed sentence. The WER obtained for the transcribed notes was 64%. The observed high WER explains the decrease in model performance as lexical features were less meaningful when they were extracted from erroneous ASR transcription of notes. Interestingly, even with erroneous ASR transcription, the performance of the proposed approach built using the combination of gaze and speech features remains better than Khan et al. [19] gaze-only approach by an average AP score of 10%.

Moreover, the employed dataset contained 55 voice notes that were manually annotated to more than one text passage by the participants during ground truth collection. We observed that our proposed multimodal approach employing gaze and speech modality was robust enough to correctly map voice notes to multiple text passages. Out of the 55 voice notes, the user-dependent model was able to map 48 (87%) voice notes and the user-independent model was able to map 45 (81%) voice notes to correct multiple text messages.

5.2 Model performance for different note types

To understand why the multimodal approach outperforms unimodal ones, we tested the user-independent models split by the three different kinds of voice notes observed (see Section 3.3). We report the classification performance of the models trained using gaze-only (gaze-based), speech-based (lexical and implicit reference-based) and the combination of gaze and speech features for anchoring different types of voice notes in Table 2. The results indicate that for all types of voice notes, the prediction accuracy of the model is boosted when trained on the combination of gaze and speech features compared to using gaze or speech features alone. Further, we observed that the model trained on both feature sets had the highest performance for anchoring content-related voice notes. These results were expected as participants often tended to summarise the read content while recording this type of note, i.e., by using words written in the relevant text passage and fixated more frequently on the passages that elicited the note. Thus, for content-related notes, both gaze and speech features contributed significantly to predict the relevant text passage and boosted the model performance to a relatively higher F_1-Score of 0.93.

Lastly, we observed that our proposed model trained on the combination of gaze and speech features had a relatively lower F_1-Score of 0.87 for reflective voice notes as compared to the other two types of notes. These results could be explained by the observations made in our exploratory data analysis. While recording reflective notes, users often vocalised their own opinion of the read content that had no obvious connection with the text of the referent passage. Consequently, the extracted speech features are not sufficient to predict the relevant text passage, as depicted by the F_1-Score of 0.36 for the model trained on just the speech features. However, for this note type, the gaze features were a good predictors (F_1-Score: 0.79) as they tend to capture the broader reading pattern of the user that can help in identifying the relevant text passage which informed the user's verbalised opinion.

5.3 Feature importance

We employed the SHAP algorithm [25] to investigate the importance of each feature on the performance of the model trained on the combination of gaze and speech features. We use the Shapley values of the top 8 features obtained from this algorithm to plot the distribution of the importance of each feature on the model's prediction, as shown in Figure 5. The features in the plot are ordered in decreasing importance from top to bottom. Each row corresponds to a feature, and a point in each row represents a text passage. The x coordinate of the point is determined by the Shapley value and represents the impact of that feature on the model's prediction for the specific text passage. The colour of each point depicts the value of that feature for the text passage under consideration, with red depicting high and blue depicting low value.

As shown in Figure 5, the gaze feature capturing the reading patterns of the user was most influential in predicting whether a voice note was made regarding a text passage. The red points on the right side of the corresponding row depict the text passages that received a higher probability score by the Annotation classifier [19] are more likely to be predicted as the candidate passage for

Table 2: Performance of the user-independent classifiers trained on gaze-based, speech-based (lexical and reference informed) and the combination of gaze & speech features for implicitly anchoring different kind of voice notes to text passages.

Type of Note	Evaluation Measures	Gaze-only	Speech-based	Gaze & Speech
Content Related	F_1-Score	.84	.55	.93
	Average Precision	.73	.48	.88
Reflective	F_1-Score	.79	.36	.87
	Average Precision	.65	.29	.78
Implicitly Referenced	F_1-Score	.74	.71	.89
	Average Precision	.63	.61	.83

voice annotation. We also observe that some text passages are being classified as *Annotated* although they obtained a low probability value (shown through the blue dots on the right side of the row). This reveals that for these text passages, other features contributed to the model's prediction.

Similarly, we observed that for a subset of text passages, the semantic similarity rank feature (shown through Similarity_Rank label) and p-value of the cosine score (shown through Significance_Similarity_Rank label) show a negative association for classifying a text passage as *Annotated*. This pattern is reflected by the blue dots on the right side of the respective feature row. A possible explanation of this behaviour is that users often make voice notes to summarise their understanding of the read content by using words written in the candidate text passage. As a result, the semantic similarity rank between such users' note and the text passage would be closer to 1, and the p-value of the observed cosine score would be low, which depicts a higher significance of the calculated score.

Lastly, we observed that the text passages upon which the user fixated more often with a prolonged duration (depicted through the fixation count and duration feature for *this*, *these*) while making an implicit reference are more likely to be classified as a referenced text passage. This behaviour could be explained by the observation of our exploratory data analysis that users often directly looked at the object about which they are talking while making an implicit reference in their speech.

6 DISCUSSION

In this work, we propose a multimodal pipeline that combines natural gaze and speech to facilitate implicit anchoring of voice notes. Our results showed that our proposed pipeline could efficiently combine gaze and speech to anchor voice to referent text passages accurately. Below, we discuss the most important observations from our results.

6.1 Overall performance of the multimodal pipeline

Using the our machine learning-based pipeline, we constructed two gaze and speech trained classifiers—user-dependent and user-independent—for the task of implicitly anchoring voice notes to relevant text passages. Our results show that both our models were able to anchor the majority of users' voice notes correctly (see Table 1). These results demonstrate our pipeline's effectiveness for combining natural gaze and speech patterns to facilitate implicit anchoring of voice notes.

Moreover, our results demonstrate that combining users' gaze and speech results in a more robust anchoring of all types of voice notes than using just their gaze behaviour (see Table 2). This boost in performance is especially notable when predictions are made for *implicitly referenced notes* (increase in F_1-Score = 15%, Average Precision = 20%). This is because while recording such voice notes, participants made ambiguous references to the content displayed on the screen. Hence, for such type of note, when natural speech is combined with gaze, the proposed pipeline captures three distinct dimensions of interaction (1) the lexical similarity between vocal utterances and read text passages; (2) users' visual attention, while making an implicit reference to content; and (3) the overall global gaze pattern depicted by the user while uttering a note. This information, when integrated, enabled the classification model to achieve a more thorough understanding of the user's context and consequently predict accurate text passage for anchoring implicitly referenced voice notes.

Motivated by the high performance of our proposed gaze and speech combined models, we believe that they could be deployed in different settings. For instance, the user-dependent model could be used to build a personalized voice annotation system for a specific user (e.g., an academic researcher editing their own papers), as the model can improve its performance by learning from the user's note-taking behaviours. On the other hand, the proposed user-independent model could be employed in a real-time voice annotation system used by a general audience for voice annotation, such as markers at a university grading students' papers.

6.2 Comparison with existing approaches

We compared our multimodal approach against the current state of the art gaze-only approach proposed by Khan et al. [19] to demonstrate the effectiveness of combining gaze and speech for implicit anchoring of voice notes. The results indicate that our proposed multimodal approach outperforms the gaze-only approach by an average of 12% for the Average Precision score. This increase in performance can be explained by giving two examples. First, as reported by Khan et al., whenever similar reading patterns are observed in two text passages, the gaze-only approach might anchor voice notes to one correct text passage and an additional incorrect text passage [19]. This error is resolved in the multimodal approach, where the speech-based features break the tie by giving

lower weights to text passages that have a lower semantic similarity with the voice notes and a lower feature value for reference informed gaze features. As a result, our proposed model would correctly classify the passage as *Not Annotated* even though the gaze-based feature value will be high for the misclassified text passages.

Second, the findings of [19] also suggest that the gaze-only approach occasionally failed to anchor voice notes when users exhibit a skimming pattern over the relevant text passage while uttering the voice note. We observed that this behaviour was often exhibited when users recorded reflective voice notes. For such cases, our multimodal approach is more effective (see Table 2) as it performs a more fine-grained analysis of users' gaze with speech to resolve implicit references and leverages the semantic similarity of the content annotated notes. Thus, our proposed model would correctly classify the passage as *Annotated* even though the gaze-based feature value will be low for the misclassified text passages.

These observations show that the combination of both input modalities captures different aspects of the user's behaviour that are important for voice annotation. The gaze-based features capture the user's global reading patterns during the recording of a voice note, while the speech-based features capture the semantic relevance between the recorded notes and the relevant text passages and help resolve implicit referrals in the uttered voice notes. Hence, the interaction offered by our multimodal approach may break for fewer annotation scenarios and, consequently, lead to a better overall reading experience for users.

6.3 Towards robust anchoring of digital content

Our results demonstrate the effectiveness of the proposed multimodal pipeline for implicitly anchoring users' voice notes during reading. However, we believe that our proposed pipeline can be generalised to implicitly map users' utterances with any digital content that can be represented as text or word embeddings. Below we provide examples of three such scenarios where our proposed pipeline can be used to implicitly map users' vocal utterances with relevant regions of the digital content.

6.3.1 Automating usability testing. User experience researchers commonly use the 'think-aloud' protocol in usability testing sessions. In these sessions, participants are asked to verbalise their thoughts while they perform tasks on the system interface. Current usability software tools require researchers to manually determine which UI elements are being referred by participants as they voice their thoughts. Our proposed pipeline can be used to build an automated usability testing tool that leverages participants' natural gaze and speech patterns to implicitly map their comments to the areas of interest on the interface. However, to use our pipeline for building such a tool, it is necessary to convert UI elements into vector embeddings so that semantic relevance between participants comments and visual content can be measured. Prior work in the domain of *screen understanding* has put forward various deep models that convert UI elements (e.g., textual content, visual design and layout patterns of the screen, images) into word embeddings [24, 44]. By leveraging such an embedding model, during usability testing, each UI element can be converted into word vectors that capture the semantic information embedded in interfaces. This implicit interaction would enable researchers to more efficiently identify regions of the interface that needs improvement.

6.3.2 Automating web annotations. People often create both private and public annotations on web pages to express and share their opinions regarding the web resources [12]. However, this process requires them to manually mark web items and then append a voice or text note to them [1, 2]. Our pipeline can be used to build a web extension that can automatically map users' vocal utterances to the relevant web content. In order to build such a web extension, items on a web page should first be converted to word embeddings. Prior work in the field of natural language processing has put forward approaches to convert web elements (e.g., buttons, form inputs, bookmarks, text content) into vector embedding that captures their semantic and spatial context [30]. Our pipeline can leverage these context-rich web embeddings and combine them with users' gaze and speech patterns to identify web regions that are to be anchored with a vocal utterance. This implicit interaction would enable users to conveniently express and share their thoughts while browsing through the Web.

6.3.3 Facilitating annotations of data visualisations. Data analysts often vocalize their thoughts when comparing multiple visualisations (e.g., charts and graphs) displayed together. However, to save their insights for later referral or to share them with peers, they manually select the visualisation and then voice their insights about it [20]. Our multimodal pipeline can make this interaction natural and convenient for analysts by implicitly mapping their vocal utterances insights with the referent visualisation. Though, to use our pipeline for facilitating such an interaction, it is necessary to first convert visualisation into a text description. For this purpose, image-based deep models can be used that captures the semantics of visual elements (e.g., charts and images) in textual descriptions [5, 43]. By leveraging such image-based deep models, our pipeline can obtain the textual descriptions of visualisations and then combine them with the analyst' gaze and speech patterns to identify the visualisation that is to be anchored with their vocal utterance. This implicit anchoring of visualisation with voice comments would allow data analysts to express their insights without breaking their flow.

6.4 Limitations and future Work

We acknowledge four important limitations of our work. First, to identify an occurrence of implicit reference to textual content, our pipeline relies on the presence of the three demonstrative pronouns in the user's speech, i.e. *these*, *this* and *those*. However, for voice notes such as *"The results in the previous paragraph are important"*, the user is making an implicit reference to the last read text passage without explicitly uttering a demonstrative pronoun. Hence for such notes, the employed technique which relies on the explicit usage of pronouns would fail to identify a voice note in which an implicit reference is made. In order to make our pipeline more robust, future research could leverage NLP techniques employing the knowledge of a spoken language understanding [9] to identify a broader range of vocal utterances where an implicit reference to content is made.

Second, we demonstrate the effectiveness of the multimodal pipeline for combining natural gaze and speech on a specific anchoring task, i.e. (anchoring voice notes to text during digital reading). Although this approach is suitable for showing the technical feasibility of our pipeline, future research is required to demonstrate its effectiveness for anchoring voice utterances to other various digital content, such as videos and images.

Third, our work demonstrates that users' natural gaze and speech behaviour can enable implicit anchoring of voice notes to relevant textual paragraphs with high accuracy. However, the anchoring of voice notes can be made more precise by restricting the scope of the textual region to the line component of the page. This will result in voice notes being mapped to the exact sentences referenced by the user in the recorded note. Therefore, in future research, we plan to demonstrate the feasibility of our proposed gaze and speech pipeline for enabling implicit anchoring of voice notes to smaller text regions such as sentences or words.

Lastly, we employed sentence-BERT model to effectively encode text passages and voice notes to vector embeddings. This was because the text passage and voice notes in the employed dataset were within 756 characters limits employed by the sentence-BERT model to effectively represent text. However, to encode voice notes and anchor them to relevant text passages that are relatively longer, future research can leverage our pipeline and experiment with document or paragraph embeddings (e.g., Doc2Vec [23] and paragraph vector [11]) models which capture the semantics of documents in vector embeddings.

7 CONCLUSION

Our work contributes a multimodal pipeline that combines natural gaze and speech behaviour to facilitate implicit anchoring of users' vocal utterances to digital content. For this purpose, we leverage and combine users' natural gaze activity, the semantic knowledge from their vocal utterances and the synchronicity between gaze and speech input to build a machine learning model that could predict text passages for implicitly anchoring of voice notes. Through such a pipeline, our work reveals insights into how two modalities that are key in enabling implicit interactions helps address recognition errors of one another to result in a robust anchoring of vocal utterances to relevant content. We evaluated our approach on the dataset collected by Khan et al. [19] and obtained a high classification performance (F_1-Score = 0.90) for predicting text passages that are to be mapped with voice notes. Overall, the findings of our work reveal the combined feasibility of using gaze and speech modality to anchor voice notes implicitly. Further, it enables the design of multimodal systems that combines natural gaze and speech patterns to enable novel natural user experiences.

ACKNOWLEDGMENTS

We wish to thank Bayu Trisedya for the useful discussion on this work. Eduardo Velloso is the recipient of an Australian Research Council Discovery Early Career Researcher Award (Project Number: DE180100315) funded by the Australian Government. Anam Ahmad Khan is supported under the Melbourne Graduate Research Scholarship.

REFERENCES

[1] 2021. Hypothesis. https://web.hypothes.is/
[2] 2021. reflect-in-seesaw. https://chrome.google.com/webstore/detail/reflect-in-seesaw-extensi/lhgiigkiddoalobhmmcpdhddlccindjj
[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing* 9, 4 (2016), 478–490.
[4] Thomas Bader, Matthias Vogelgesang, and Edmund Klaus. 2009. Multimodal Integration of Natural Gaze Behavior for Intention Recognition during Object Manipulation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) *(ICMI-MLMI '09)*. Association for Computing Machinery, New York, NY, USA, 199–206. https://doi.org/10.1145/1647314.1647350
[5] Abhijit Balaji, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. 2018. Chart-text: A fully automated chart image descriptor. *arXiv preprint arXiv:1812.10636* (2018).
[6] TR Beelders and PJ Blignaut. 2011. The usability of speech and eye gaze as a multimodal interface for a word processor. *Speech Technologies* (2011), 386–404.
[7] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (Seattle, Washington, USA) *(SIGGRAPH '80)*. Association for Computing Machinery, New York, NY, USA, 262–270. https://doi.org/10.1145/800250.807503
[8] Matteo Casarini, Marco Porta, and Piercarlo Dondi. 2020. A Gaze-Based Web Browser with Multiple Methods for Link Selection. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 17, 8 pages. https://doi.org/10.1145/3379157.3388929
[9] Asli Celikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2014. Resolving referring expressions in conversational dialogs for natural user interfaces. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2094–2104.
[10] Siyuan Chen, Julien Epps, and Fang Chen. 2013. Automatic and Continuous User Task Analysis via Eye Activity. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 57–66. https://doi.org/10.1145/2449396.2449406
[11] Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998* (2015).
[12] Xin Fu, Tom Ciszek, Gary Marchionini, and Paul Solomon. 2005. Annotating the Web: An exploratory study of Web users' needs for personal annotation tools. *Proceedings of the American Society for Information Science and Technology* 42, 1 (2005).
[13] Qiong Gu, Li Zhu, and Zhihua Cai. 2009. Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications*. Springer, 461–471.
[14] Andrea Guazzini, Eiko Yoneki, and Giorgio Gronchi. 2015. Cognitive dissonance and social influence effects on preference judgments: An eye tracking based system for their automatic assessment. *International Journal of Human-Computer Studies* 73 (2015), 12–18. https://doi.org/10.1016/j.ijhcs.2014.08.003
[15] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 263–266.
[16] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and Turn-Taking Behavior in Casual Conversational Interactions. *ACM Trans. Interact. Intell. Syst.* 3, 2, Article 12 (Aug. 2013), 30 pages. https://doi.org/10.1145/2499474.2499481
[17] Casey Kennington, Spyridon Kousidis, and David Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. *Proceedings of SigDial 2013* (2013).
[18] Anam Ahmad Khan, Sadia Nawaz, Joshua Newn, Jason M. Lodge, James Bailey, and Eduardo Velloso. 2020. Using voice note-taking to promote learners' conceptual understanding. arXiv:2012.02927
[19] Anam Ahmad Khan, Joshua Newn, Ryan M. Kelly, Namrata Srivastava, James Bailey, and Eduardo Velloso. 2021. GAVIN: Gaze-Assisted Voice-Based Implicit Note-Taking. *ACM Trans. Comput.-Hum. Interact.* 28, 4, Article 26 (Aug. 2021), 32 pages. https://doi.org/10.1145/3453988
[20] Yea-Seul Kim, Mira Dontcheva, Eytan Adar, and Jessica Hullman. 2019. Vocal Shortcuts for Creative Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300562
[21] Dimosthenis Kontogiorgos, Elena Sibirtseva, Andre Pereira, Gabriel Skantze, and Joakim Gustafson. 2018. Multimodal reference resolution in collaborative assembly tasks. In *Proceedings of the 4th International Workshop on Multimodal*

Analyses Enabling Artificial Agents in Human-Machine Interaction. 38–42.
[22] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*. IEEE, 1–6.
[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.
[24] Toby Jia-Jun Li, Lindsay Popowski, Tom Mitchell, and Brad A Myers. 2021. Screen2Vec: Semantic Embedding of GUI Screens and GUI Components. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[25] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.
[26] Chandra Sekhar Mantravadi. 2009. *Adaptive multimodal integration of speech and gaze*. Ph.D. Dissertation. Rutgers University-Graduate School-New Brunswick. https://doi.org/10.7282/T3QC03PM
[27] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.
[28] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. *Patterns for How Users Overcome Obstacles in Voice User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173580
[29] Cuong Nguyen and Feng Liu. 2016. Gaze-Based Notetaking for Learning from Lecture Videos. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2093–2097.
[30] Panupong Pasupat, Tian-Shun Jiang, Evan Zheran Liu, Kelvin Guu, and Percy Liang. 2018. Mapping natural language commands to web elements. *arXiv preprint arXiv:1808.09132* (2018).
[31] Annie Piolat, Thierry Olive, and Ronald T Kellogg. 2005. Cognitive effort during note taking. *Applied cognitive psychology* 19, 3 (2005), 291–312.
[32] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214
[33] Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 20–29.
[34] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[35] Dario Salvucci and Joseph Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*, 71–78. https://doi.org/10.1145/355017.355028
[36] Korok Sengupta, Min Ke, Raphael Menges, Chandan Kumar, and Steffen Staab. 2018. Hands-free web browsing: Enriching the user experience with gaze and voice modality. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–3.
[37] Hao Shen and Jaideep Sengupta. 2018. Word of mouth versus word of mouse: Speaking about a brand connects you to it more than writing does. *Journal of Consumer Research* 45, 3 (2018), 595–614. https://doi.org/10.1093/jcr/ucy011
[38] Malcolm Slaney, Andreas Stolcke, and Dilek Hakkani-Tür. 2014. The relation of eye gaze and face pose: Potential impact on speech recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 144–147.
[39] Cagkan Uludagli and Cengiz Acarturk. 2018. User interaction in hands-free gaming: A comparative study of gaze-voice and touchscreen interface control. *Turkish Journal of Electrical Engineering and Computer Sciences* 26 (07 2018). https://doi.org/10.3906/elk-1710-128
[40] Jan Van der Kamp and Veronica Sundstedt. 2011. Gaze and voice controlled drawing. In *Proceedings of the 1st conference on novel gaze-controlled applications*. 1–8.
[41] Jan van der Kamp and Veronica Sundstedt. 2011. Gaze and Voice Controlled Drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (Karlskrona, Sweden) *(NGCA '11)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages.
[42] Diogo Vieira, João Dinis Freitas, Cengiz Acartürk, António Teixeira, Luís Sousa, Samuel Silva, Sara Candeias, and Miguel Sales Dias. 2015. " Read That Article" Exploring Synergies between Gaze and Speech Interaction. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 341–342.
[43] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
[44] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. *arXiv preprint arXiv:2108.03353* (2021).
[45] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* 13, 3 (2018), 55–75.

A INDIVIDUAL CONTRIBUTION OF GAZE AND SPEECH FOR ANCHORING VOICE NOTES

In this appendix, we highlight the individual contribution of natural gaze and speech for anchoring voice notes. For this purpose, we report the performance of the user-independent models trained individually using the gaze-only (gaze-based) and speech-based (lexical and reference informed) features in Table 3. We observed that the classifier trained on features extracted from analysing just the speech data of users' notes had a lower average precision score of .47 (see Table 3). The lower performance of the speech-based models was because the text passages lying in the ROA of the voice notes are in close proximity with each other. Hence, often these passages have similar textual content, which results in an approximate same cosine similarity score for them. Consequently, although the speech features could reduce the candidate text passages for anchoring voice notes, it may be challenging to disambiguate between related passages using just these features. nevertheless, when combined with the gaze features, these speech-based features can boost the classification performance to a relatively high F_1-Score of .89.

Table 3: User-independent classifier performance trained on gaze-only and speech-based engineered feature.

Evaluation Measures	Gaze-only	Speech-based	
		Manual	ASR
F_1-Score	.80 ± .08	.53 ± .06	.43 ± .18
Average Precision	.70 ± .12	.47 ± .20	.35 ± .18

B GAZE AND SPEECH FEATURES FOR ANCHORING VOICE NOTES

In this appendix, we provide the complete list of features used to train the machine learning model for anchoring voice notes to reference text passages. We further provide explanation of calculating some references informed and lexical features.

The feature vector used to train the Logistic regression classifier consisted of lexical, reference-informed and gaze-based features. The complete list of these features is given in Table 4.

In Table 4, the minimum distance from the closest gaze point feature is the smallest distance ($d[.]$) between the text passage (T) and any gaze fixation point (P) that occurred in the 4 seconds window of analysis. The computation of this feature is shown through equation 1, where n is the total number of gaze fixation points in the defined window of analysis.

$$distance = min_{x \in [0,n]}(d[T, P_x]) \quad (1)$$

Further, in Table 4 the p-value feature of each text passage is calculated using Equation 2, where cdf is the cumulative distribution function fitted to the score distribution of random text passages

Table 4: Gaze and speech features extracted for each text passage for the task of anchoring voice notes.

Category	Features
Lexical	Semantic similarity rank with transcribed voice note p-value of the cosine similarity score
Reference Informed	Presence of *this*, *these*, *those* pronouns Fixation count for *this*, *these*, *those* pronouns Sum of fixation duration for *this*, *these*, *those* pronouns Minimum distance from the closest gaze fixation during the utterance of *this*, *these*, *those* pronouns
Gaze-based	Probability of a text passage being predicted as *Annotated* based on the classifier proposed by Khan et al. [19]

and x is the observed cosine similarity score of the candidate text passage.

$$p - value = 1 - cdf(x) \quad (2)$$