Towards Real-Time Speech Segmentation for Glanceable Conversation Visualization

Shanna Hollingworth*
University of Calgary

Wesley Willett†
University of Calgary

Abstract

We explore the use of segmentation and summarization methods for the generation of real-time conversation topic timelines, in the context of glanceable Augmented Reality (AR) visualization. Conversation timelines may serve to summarize and contextualize conversations as they are happening, helping to keep conversations on track. Because dialogue and conversations are broad and unpredictable by nature, and processing must be done in real-time, not all relevant information may be present in the text at the time it is processed. Thus, we present considerations and challenges which may not be as prevalent in traditional implementations of topic classification and dialogue segmentation. Furthermore, we discuss how AR visualization requirements and design practices require an additional layer of decision making, which must be factored directly into the text processing algorithms. We explore three segmentation strategies – using dialogue segmentation based on the text of the entire conversation, segmenting on 1-minute intervals, and segmenting on 10-second intervals – and discuss our results.

Index Terms: Conversation Visualization, Topic Classification, Dialogue Segmentation, Augmented Reality.

1 Introduction

Both AR hardware and Large Language Models (LLMs) have seen recent strides in terms of processing power and capability, resulting in many novel research areas at their intersection. One such area which has taken off in recent years surrounds live conversation support [3]. The immersive visualization nature of AR and the ability of LLMs to analyze a wide range of potentially unstructured information make this pairing uniquely suited to this task. We are currently exploring the creation and visualization of real-time conversation timelines (Fig. 1) in AR. The generation of conversation timelines involve the classification of topics in real time as conversations evolve and change, allowing us to better understand conversation themes. These timelines may help to address common challenges in conversation which tend to disrupt conversation flow, such as losing one's train of thought mid-sentence, or retracing the conversation path to find topic connections.

The way in which we choose to visualize this information will be a primary factor in how we analyze the speech, and the type of information we can present. A visualization encompassing the overarching topics over a full conversation might serve to contextualize and summarize the conversation as a whole. Conversely, conversation visualizations at 10 second intervals may be useful to jog the memory of someone who's lost their train of thought, while 1 minute visualizations may provide a slightly higher level breakdown of subtopic breakdowns within each topic.

We must carefully consider two things in the creation of conversation timelines: (1) the segmentation of the transcribed speech and (2) classification and summarization of the segments to facilitate glanceable visualizations.

We present a set of challenges surrounding effective analysis and visualization of real-time conversation timelines at each of these three levels, and discuss our current work in addressing them.

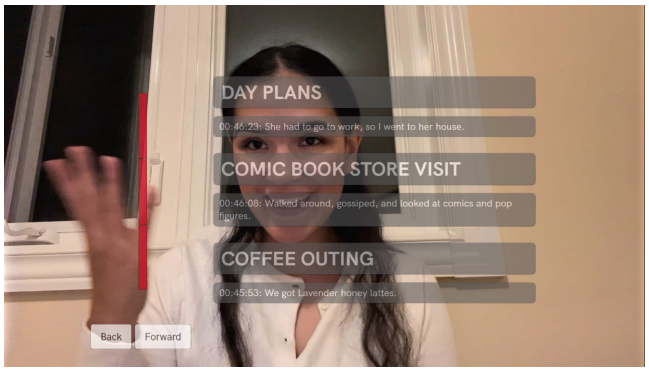

Figure 1: A snapshot of our early system prototype

2 Related Work

Conversation flows are not structurally homogeneous, and may involve countless co-dependencies [9] and variability in the dialogue [2]. Conversation topics constantly change and move away then back to their starting points, and humans are prone to tangential and off-topic thoughts within conversation that all weave into each other in a complex web. This can make it difficult to find a 'natural' way to segment and classify the conversation.

Current work in discourse visualization largely surrounds captioning and referencing [7]. Exploration in captioning has been especially targeted towards those who are deaf or hard of hearing [1, 4], where it has been noted that extraneous context including tone of voice or emotion may be important to accurately represent the conversation. Jain et al. [4] note that the augmentation of this captioning as opposed to traditional displays improve glanceability and maintenance of visual contact with other speakers. Another form of captioning which has risen in popularity is augmenting visual imagery in place of captions [6, 5].

Minimal work has been done around summarizing and itemizing conversation in real time. However, from previous works, we understand that less is more when it comes to AR visualization [8, 7]. Furthermore, through creative use of colour and text placement, we can encode greater amounts of information [1].

3 Approach

Effective information visualization in our AR system plays a major role in how we choose to analyze speech, and what sorts of outputs we consider to be acceptable. The information we collect and present through our analysis will not be useful if it is not quickly and easily understood by the user. Therefore, we must consider the impacts to the final visualization at every step of the process. At a high level, we segment conversation text by some metric (either

*e-mail: shanna.hollingwor1@ucalgary.ca
†e-mail: wesley.willett@ucalgary.ca

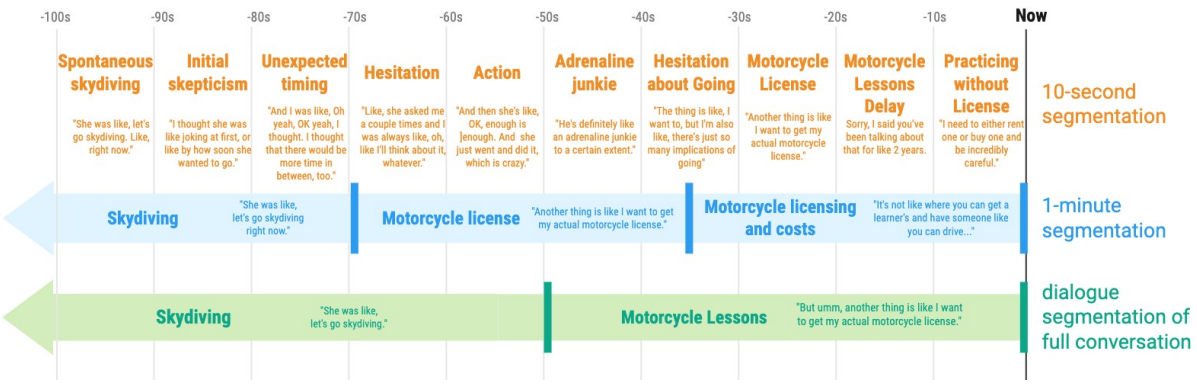

Figure 2: A visualization of the different segmentation points in the three conversation modes over the same 2 minutes of conversation

by sentence semantic similarity, or length of time) in real time, and then classify each segment. We also outline challenges, considerations, and visualization needs that affect our segmentation and summarization approaches.

3.1 Design Considerations

Consistent Visualization. Our first consideration was creating a visualization which is consistent from one minute to another, such that every visualized topic will remain in the visualization until pushed out of the field of view. The goal is to avoid visualizations which are constantly changing and reorganizing as context is gained, as this is likely to be confusing for the user. In terms of our approach, this means that once a portion of the conversation has been segmented and classified, we must commit to the classification and assume that it is a standalone segment. We can no longer include any part of this segment as context for future classifications, as it is assumed not to contribute to the next topic. Furthermore, we should have a way to handle the case where two separate segments were identified to contain the same topic. At the very least, they should not be duplicated side by side.

The trade-off may be that the timelines will not capture topics in the same way that they might if this was done retrospectively, as conversation topics have a tendency to snake around. A topic may momentarily seem to veer off, then tie back into the previous topic after a few minutes. How much this affects the reliability of the system from a user perspective should be more formally addressed as part of a user study.

Topic Contextualization. In order to account for LLM classifications which may be incorrect, or not immediately obvious, in our current iterations we have also included "representative phrases". This are short snippets taken directly from the segment that the LLM suggests best represent the topic. We hypothesize that visualizing a representative phrase alongside the topic will help users to better remember the conversation and more accurately benchmark each topic as they view the visualization.

Topic Specificity. Finally, topics must be short enough to fit neatly within the visualization, but long enough to effectively capture the specific topic. For example, if the discussion was about an upcoming trip to Japan, "Travel" might be a valid topic designation. However, a phrase like "Japan Trip Planning" better captures the specificities of the conversation. This is especially important when we are visualizing short term topics. 10 second segments are not necessarily long enough to contain dramatic topic changes, so more specific topics are likely to be more useful (Table 3).

3.2 Implementation

While topic classification is not a novel technique, the classification of live speech introduces a fresh set of both Natural Language Processing (NLP) and visualization challenges. Traditional topic classification models are not sufficient to properly capture speech, as the potential topics are limited only to what topics the model has been trained on. LLMs are capable of classifying an infinite number of topics at a high level of detail. They also allow us to get very specific with our preferred outputs, allowing us to easily achieve various design requirements, once these requirements have been identified.

However, we found in our experimentation with GPT-4o that LLMs are not yet capable of accurately performing dialogue segmentation on real-time conversation, as the information is constantly updating and changing. Thus, while we can utilize GPT-4o to assist with classification and summarization given the design considerations listed above, we must turn to more traditional approaches in order to segment the conversation.

As segments are created, they are sent one by one to GPT-4o to be classified into a topic. As a final check, if a returned topic classification was the same as the classification for the segment before it, the segments will be 'merged' into one and we will not display a new topic. This is done to ensure that the visualization will not display the same topic twice, as described under the 'Consistent Visualization' design consideration.

To examine the potential of multiple different segmentation and summarization alternatives, we conducted a set of initial design explorations in which we segment and summarize the same short conversation at multiple levels of granularity. As illustrated in Figure 2, we explored using dialogue segmentation based on the entire conversation text, as well as segmenting and summarizing using discrete 1-minute and 10-minute segments. Below, we break down and discuss each of these approaches.

Full Conversation Dialog Segmentation

For the topic segmentation of full conversations, we turned to a more traditional NLP implementation of dialog segmentation, which allows for more structure and mathematical definition. By manually writing a dialog segmentation algorithm, we are guaranteed to output the same result no matter how much of the conversation is provided at once.

We first used NLTK's SentimentIntensityAnalyzer method to assign sentiment scores to each sentence, which outputs a number between -1 and 1 based on the content of the sentence. We then create new segments when the difference between sentiment scores for two sentences passed a threshold of 0.7, using a context window of two sentences. We also defined segments of less than 50 words to

Table 1: Sample segments with topics and key phrases extracted by segmenting a full 26-minute conversation. (These two segments represent roughly the first 2.5 minutes of the conversation.)

Segment #	Topic	Full Conversation Segments	Representative Phrase
1	Skydiving	Time to mentally prepare. I was not ready when she was ready. Oh my gosh. Yeah, no, She was like, let's go skydiving. Like, right now. I know. It's just like, I've never even thought about this before. What do you mean? Yeah, I thought she was like joking at first, or like by how soon she wanted to go. But then yeah, she sends me the pictures of her. And I was like, Oh yeah, OK yeah, I thought. I thought that there would be more time in between, too. Yeah, like, she's she's very much like that. Like, she asked me a couple times and I was always like, oh, like I'll think about it, whatever. And then she's like, OK, enough is enough.	"She was like, let's go skydiving."
2	Motorcycle License	And she just went and did it, which is crazy. I was like, oh, do you think that Emmanuel would go with you? Bro, I don't know, 'cause like he's I, I think he would, 'cause he's definitely like an adrenaline junkie to a certain extent. Like, he really likes to do that type of stuff. I mean, honestly, Yeah, probably. I feel like you could go together then. That'd be cute. Bro, I The thing is like, I want to, but I'm also like, there's just so many implications of going 'cause it's like. Like, yeah, you're with the instructor, but like, things can always go wrong. You know what I mean? But umm, another thing is like I want to get my actual motorcycle license. I want to be 1. Sorry, I said you've been talking about that for like 2 years. I thought you know the lessons. Let me tell you why. OK 'cause I did do the lessons right and like it was fine, whatever. But The thing is like you can't practice without a license and like the way you get the license is with the driven test. So it's like it's not like where you can get a learner's and have someone like like you can drive when someone is there with you, right? That's fully licensed. You can't do that with a motorcycle, right? And then it's also like do I really want to invest like 3-4 K in a bike? And I'm still learning and like, what happens if I drop it? Or like, you know, there's like lots of things that can ruin a bike. Yeah, So you know what I mean? So it's like. I need to either be able to rent one at my lessons place and like be able to get decent, or I have to buy one and be like incredibly careful.	"But umm, another thing is like I want to get my actual motorcycle license."

be too short to encompass a topic, and merged these segments into the ones before them. This approach allows all segmentation to be locally done, limiting calls to the cloud-based LLM. Table 1 shows an example of the resulting text segments and topic classifications for a 26-minute conversation.

1-Minute and 10-Second Segmentation

A multi-level visualization allows for different levels of conversation summarization. Where topic segmentation over the full conversation will provide a high level overview, smaller 1-minute (Table 2) and 10-second (Table 3) segments provide more low-level details regarding conversation specifics. For example small snippets may capture details such as an expression of emotion (Table 3 Segment 4) or a commitment to a task. Ultimately, we chose to separate the chunks by number of words. According to various internet sources, the average native English speaker in the US says around 100-150 words per minute in casual conversation. Thus, we segment the text at the sentence that brings the segment to 140 words or more. While this may not produce exact results, in practice the results appear reasonable (Table 2). We use similar logic for the 10 second segments, segmenting at 24 words.

As seen in Figure 2, we have found that topic alignment across segmentation levels is quite high. For example, topics related to 'Skydiving' and 'Motorcycle License' can be found in all three versions, accompanied by either identical of very similar representative phrases. This may be useful in helping users to visually benchmark the conversation against the other visualizations, improving the ease of comprehension for flipping between visualizations.

4 Future Work

Our approach thus far has been a promising start for generating and visualizing real time conversation. On the implementation side, we will continue to refine and improve our dialogue segmentation algorithm, as there are a multitude of different ways to define a 'turn in the conversation'. Similarly, exploration of a wider range of fixed segment lengths is warranted – as other segment durations (say 30 seconds or 5 minutes) may prove more reflective of the pace of typical conversations. Multi-level visualization impacts the way we prompt the LLM, and can greatly increase the number of calls to it. Therefore we must be intentional in choosing these spans such that each visualization provides a unique and accurate summarization of

Table 2: Segments, topics, and representative phrases from the same 2.5 minute block extracted via 1-minute segmentation.

Segment #	Topic	1 Minute Segments	Representative Phrase
1	Skydiving	Time to mentally prepare. I was not ready when she was ready. Oh my gosh. Yeah, no, She was like, let's go skydiving. Like, right now. I know. It's just like, I've never even thought about this before. What do you mean? Yeah, I thought she was like joking at first, or like by how soon she wanted to go. But then yeah, she sends me the pictures of her. And I was like, Oh yeah, OK yeah, I thought. I thought that there would be more time in between, too. Yeah, like, she's she's very much like that. Like, she asked me a couple times and I was always like, oh, like I'll think about it, whatever. And then she's like, OK, enough is enough. And she just went and did it, which is crazy. I was like, oh, do you think that Emmanuel would go with you?	"She was like, let's go skydiving right now."
2	Motorcycle license	Bro, I don't know, 'cause like he's I, I think he would, 'cause he's definitely like an adrenaline junkie to a certain extent. Like, he really likes to do that type of stuff. But I mean, honestly, Yeah, probably. I feel like you could go together then. That'd be cute. Bro, I The thing is like, I want to, but I'm also like, there's just so many implications of going 'cause it's like. Like, yeah, you're with the instructor, but like, things can always go wrong. You know what I mean? Yeah. But umm, another thing is like I want to get my actual motorcycle license. I want to be 1. Sorry, I said you've been talking about that for like 2 years. I thought you know the lessons. Let me tell you why. OK 'cause I did do the lessons right and like it was fine, whatever.	"Another thing is like I want to get my actual motorcycle license."
3	Motorcycle licensing and costs	But The thing is like you can't practice without a license and like the way you get the license is with the driven test. So it's like it's not like where you can get a learner's and have someone like you can drive when someone is there with you, right? That's fully licensed. You can't do that with a motorcycle, right? And then it's also like do I really want to invest like 3-4 K in a bike? And I'm still learning and like, what happens if I drop it? Or like, you know, there's like lots of things that can ruin a bike. Yeah, So you know what I mean? So it's like. I need to either be able to rent one at my lessons place and like be able to get decent, or I have to buy one and be like incredibly careful.	"It's not like where you can get a learner's and have someone like you can drive when someone who is fully licensed is there with you; you can't do that with a motorcycle."

Table 3: Segments, topics, and representative phrases from the first 90 seconds extracted via 10-second segmentation.

Segment #	Topic	10 Second Segments	Representative Phrase
1	Spontaneous skydiving	Time to mentally prepare. I was not ready when she was ready. Oh my gosh. Yeah, no, She was like, let's go skydiving. Like, right now.	"She was like, let's go skydiving. Like, right now."
2	Initial skepticism	I know. It's just like, I've never even thought about this before. What do you mean? Yeah, I thought she was like joking at first, or like by how soon she wanted to go.	"I thought she was like joking at first, or like by how soon she wanted to go."
3	Unexpected timing	But then yeah, she sends me the pictures of her. And I was like, Oh yeah, OK yeah, I thought. I thought that there would be more time in between, too.	"And I was like, Oh yeah, OK yeah, I thought. I thought that there would be more time in between, too."
4	Hesitation	Yeah, like, she's she's very much like that. Like, she asked me a couple times and I was always like, oh, like I'll think about it, whatever.	"Like, she asked me a couple times and I was always like, oh, like I'll think about it, whatever."
5	Action	And then she's like, OK, enough is enough. And she just went and did it, which is crazy. I was like, oh, do you think that Emmanuel would go with you?	"And then she's like, OK, enough is enough. And she just went and did it, which is crazy."
6	Adrenaline junkie	Bro, I don't know, 'cause like he's I, I think he would, 'cause he's definitely like an adrenaline junkie to a certain extent. Like, he really likes to do that type of stuff.	"He's definitely like an adrenaline junkie to a certain extend."
7	Hesitation about Going	But I mean, honestly, Yeah, probably. I feel like you could go together then. That'd be cute. Bro, I The thing is like, I want to, but I'm also like, there's just so many implications of going 'cause it's like.	"The thing is like, I want to, but I'm also like, there's just so many implications of going"
8	Motorcycle License	Like, yeah, you're with the instructor, but like, things can always go wrong. You know what I mean? Yeah. But umm, another thing is like I want to get my actual motorcycle license.	"Another thing is like I want to get my actual motorcycle license."
9	Motorcycle Lessons Delay	I want to be 1. Sorry, I said you've been talking about that for like 2 years. I thought you know the lessons. Let me tell you why.	"Sorry, I said you've been talking about that for like 2 years."

the conversation. Another potential exploration is how a 'subtopic' visualization would compare, where each segment in the full conversation breakdown (Table 1) would be further broken down into it's subtopics, rather than by timestamp.

Ultimately, we aim to understand which elements of the visualization are the most impactful, and which elements should be removed or reworked. We invite further discussion about how conversation topics might be best visualized in real-time and in which context they might provide the greatest value.

Acknowledgments

We would like to acknowledge Victoria Wong, who helped create mockups for the second iteration of our design and assisted with prototype development and styling. This work was supported in part by Alberta Innovates and the Canada Research Chairs Program.

References

[1] C. a. de Lacerda Pataca, M. Watkins, R. Peiris, S. Lee, and M. Huenerfauth. Visualization of speech prosody and emotion in captions: Accessibility for deaf and hard-of-hearing users. In *Proceedings of the 2023*

CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581511 1

[2] G. Flammia. *Discourse segmentation of spoken dialogue: an empirical approach*. PhD thesis, Massachusetts Institute of Technology, 1998. 1

[3] I. Hirskyj-Douglas, A. Kantosalo, A. Monroy-Hernández, J. Zimmermann, M. Nebeling, and M. Gonzalez-Franco. Social ar: Reimagining and interrogating the role of augmented reality in face to face social interactions. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, p. 457–465. Association for Computing Machinery, 2020. doi: 10.1145/3406865.3418585 1

[4] D. Jain, B. Chinh, L. Findlater, R. Kushalnagar, and J. Froehlich. Exploring augmented reality approaches to real-time captioning: A preliminary autoethnographic study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, DIS '18 Companion, p. 7–11. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3197391.3205404 1

[5] J. Liao, A. Karim, S. S. Jadon, R. H. Kazi, and R. Suzuki. Realitytalk: Real-time speech-driven augmented presentation for ar live storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3526113.3545702 1

[6] X. B. Liu, V. Kirilyuk, X. Yuan, A. Olwal, P. Chi, X. Chen, and R. Du. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1–20, 2023. 1

[7] F. Müller, S. Günther, A. H. Nejad, N. Dezfuli, M. Khalilbeigi, and M. Mühlhäuser. Cloudbits: supporting conversations through augmented zero-query search visualization. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, p. 30–38. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3131277.3132173 1

[8] T. T. Nguyen, D. T. Nguyen, S. T. Iqbal, and E. Ofek. The known stranger: Supporting conversations between strangers with personalized topic suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 555–564. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702411 1

[9] F. Wolf and E. Gibson. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics*, 31(2):249–287, 06 2005. doi: 10.1162/0891201054223977 1