ConFides: A Visual Analytics Solution for Automated Speech Recognition Analysis and Exploration

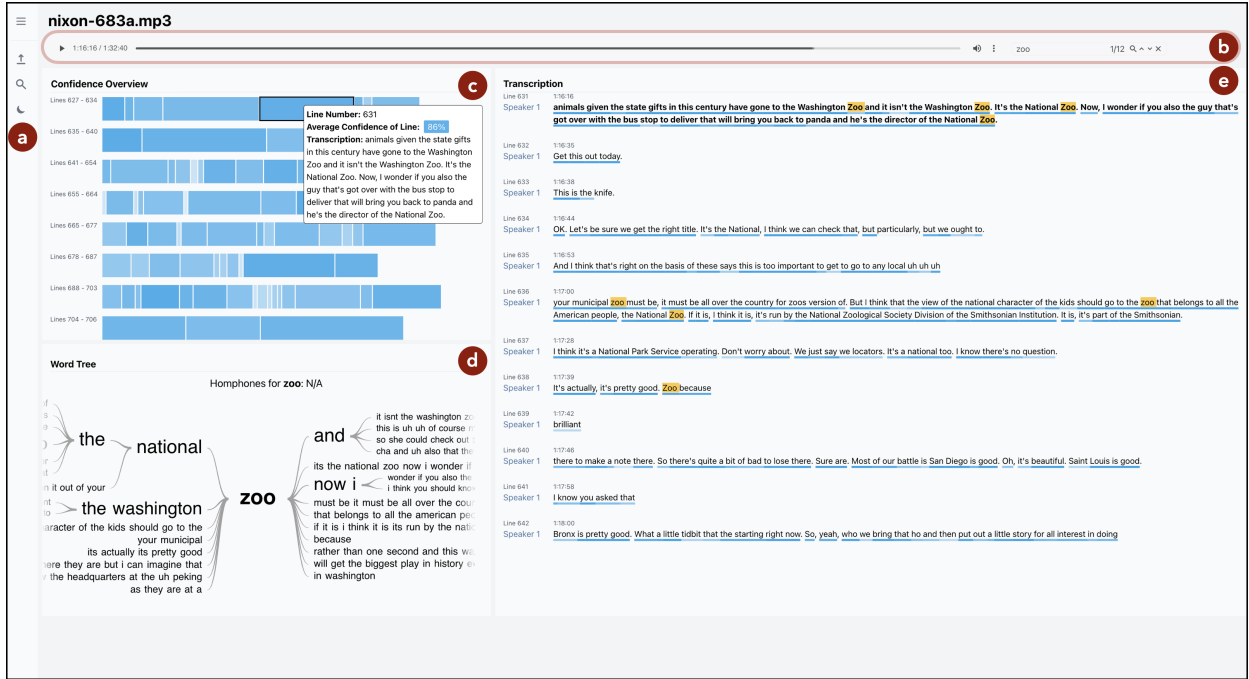Sunwoo Ha, Chaehun Lim, R. Jordan Crouser, and Alvitta Ottley

Fig. 1: An overview of **ConFides**: (a) The collapsible side menu contains controls for login, selecting, uploading, and transcribing audio files via AWS Transcribe. (b) At the top of the dashboard are the audio player and search bar. (c) The confidence overview displays the length and average confidence value of each line segment in the transcription (encoded by the width and opacity of each rectangle, respectively). (d) The word tree provides context to a specific search term and shows which words most often follow or precede it. (e) The user can view and edit the transcription; each word is underlined where its opacity indicates the confidence score.

Abstract— Confidence scores of automatic speech recognition (ASR) outputs are often inadequately communicated, preventing its seamless integration into analytical workflows. In this paper, we introduce **ConFides**, a visual analytic system developed in collaboration with intelligence analysts to address this issue. **ConFides** aims to aid exploration and post-AI-transcription editing by visually representing the confidence associated with the transcription. We demonstrate how our tool can assist intelligence analysts who use ASR outputs in their analytical and exploratory tasks and how it can help mitigate misinterpretation of crucial information. We also discuss opportunities for improving textual data cleaning and model transparency for human-machine collaboration.

Index Terms—Visual analytics, confidence visualization, automatic speech recognition

✦

1 Introduction

The volume of audio data grows exponentially with each passing hour. However, turning raw audio data into actionable intelligence can be complex, as it relies heavily on the accuracy of the algorithms, transcription services employed, and

- S. Ha, C. Lim, and A. Ottley are with Washington University. E-mails: sha@wustl.edu, chaelim@wustl.edu, alvitta@wustl.edu.
- R. J. Crouser is with Smith College. E-mail: jcrouser@smith.edu.

the quality of the original audio files. Despite significant advances in AI and Automatic Speech Recognition (ASR), the analysis of audio data remains fraught with inaccuracies and is notably time-consuming [4]. As a result, fields such as intelligence analysis urgently require solutions that enable them to efficiently and accurately harness this vast and rapidly expanding resource.

Existing solutions that provide automatic speech-to-text and editing capabilities of the transcription (i.e., Rev, Happy Scribe) fall short as they often inadequately or rarely highlight and convey metrics that may bring forward potential inaccuracies to the user. The lack of transparency between the machine and the analyst within these existing services hinders the complete integration and effective utilization of speech-to-text technology within their workflows and analytical processes [24]. Relying on AI, especially in human-machine collaborations, without awareness of these uncertainties can be detrimental, as the

quality of an analyst's work is a direct result of the trust [8] in and accuracy of the information presented to them.

To address this challenge, we present a new visual analytics system called **ConFides**, which we developed in collaboration with intelligence analysts. Our system enhances understanding of speech-to-text results by showing how confident the model is in its transcription. It uses the speech-to-text service from Amazon Web Services (AWS) and aims to make exploring and editing post-AI transcription easier by providing visual representations of confidence levels. By using interactive visualizations and multiple views, we demonstrate how our tool could improve the analysis of speech-to-text output and foster trust in human-machine collaborations. We also discuss how our system could support common analytical and exploratory tasks for intelligence analysts working with audio transcriptions. Finally, we explore opportunities to improve model transparency and textual data cleaning to encourage more effective human-machine collaboration. Through this work, our main contributions include:

- We outline specific design goals derived from tasks that intelligence analysts perform within their workflow when inspecting audio transcriptions.
- We develop a visual analytics system, **ConFides**, that allows analysts to easily transcribe their audio files and explore the transcription while being aware of uncertainties within the data.
- To demonstrate the applicability and usefulness of the system, we present a realistic use case scenario that forages for information from the Nixon White House tapes.

2 Related Work

We briefly overview the prior work on transparency in human-machine collaborations in visual analytics and AI-assisted decision-making settings and visualizing confidence in speech-to-text outputs.

2.1 Communicating Model Uncertainty

From the point of view of mixed-initiative systems, Sacha et al. [24] discussed the role of uncertainty, awareness, and trust in visual analytics and argued that users' confidence in the machine teammate's results depends on their degree of awareness of the different types of uncertainty that are present or generated in the system. Additionally, research has shown that visually communicating uncertainty can support users' interpretation of data and support decision-making [10, 15, 16]. Therefore, communicating uncertainty is vital, especially when an AI agent assists in a decision-making process as it can potentially mitigate unwanted behaviors from the user such as underutilization [20, 9, 17, 7, 12] or overreliance [6, 14, 12] on AI suggestions. Wang et al. [29] explored various explanation methods for recidivism prediction and forest cover prediction tasks. Their study found supportive evidence suggesting that providing information about feature contribution allowed participants to be aware of the uncertainties within the model and appropriately calibrate their trust.

2.2 Visualizing Confidence in Speech-To-Text Outputs

Various methods change the appearance of text to convey speech-to-text confidence (e.g., alternating the font size [22], font color [5], font opacity [4], and underlining [28]). Vertanen et al. [28] utilized underlining, color, and opacity to visually embed confidence in speech recognition outputs, where the opacity of the colored underline of the text is based on the degree of confidence that the predicted word is accurate. In addition to embedding confidence in the transcription text, Wu et al. [31] developed a prototype visualization system that provides an overview of the confidence of speech-to-text outputs with a bar chart. In this visualization, each bar is mapped to a segment in the transcription, and all the bars of one transcribed text are sequentially visualized. We take inspiration from these prior works to communicate confidence to analysts and design the interactive visualizations used in our system.

3 Design Goals

Numerous applications exist for text-to-speech data, including legal environments where depositions are recorded [23], healthcare settings where doctors record audio notes for medical records [18], automatic captioning for online videos, and academic research involving interview studies. Each use case scenario has different and unique tasks. Therefore, to create an ecologically valid solution, we ground our work in government intelligence analysis settings where analysts often analyze transcribed audio data for matters related to national security.

We partnered with the Laboratory of Analytical Sciences, which facilitates collaborations between the Department of Defense's (DoD) intelligence community and academia. Our tasks and design goals were derived from an initial one-hour, semi-structured interview with a language specialist from the DoD. We inquired about current tools, tasks, and goals that are supported, typical analysis workflows, collaborative activities, pain points, and features that are missing. This initial interview provided context and grounded the design process. We then adopted an interactive design methodology with bi-weekly feedback meetings from January 2023 to December 2023. Attendees included academic partners, DoD-affiliated analysts, and the projects' program manager. Below we summarize the design goals (**DG**) for the visual analytics system:

- **DG1: Interactively visualize transcription and assess the quality of the transcription output**. Current transcription services lack transparency and rarely communicate the uncertainties and confidence values of automatic transcriptions. In many tasks, especially ones with consequential decisions, the analyst must determine when to trust the automatic transcription versus taking the time to listen to the source audio. Analysts must also determine whether additional data cleaning is necessary.
- **DG2: Correct and perform basic textual data cleaning**. The system should allow the analyst to easily explore and playback audio segments, and make corrections to the transcription (e.g., adding, deleting, viewing alternative words, and replacing the text if desired).
- **DG3: Discern patterns and extract actionable intelligence within the transcription**. The system should facilitate the extraction of insights and allow the analyst to identify patterns and trends across the audio transcription data. It should enable the analyst to explore the context and frequency of spoken content, enhancing the ability to track discourse over time.

To support the analyst in performing these tasks, ConFides offers three main interactive views, as detailed in Section 4.2. Additionally, the backend of the system is connected to AWS Transcribe, enabling effortless transcription of audio files and exploration of the outputs in a single interface.

4 Visual Analytics System Design

Through several iterative designs and feedback from our collaborators, we developed **ConFides**. The following sections describe the framework of the system and how each view will support the analyst with the design goals outlined in Section 3.

4.1 Framework

The backend of our system utilizes the service offered by AWS Transcribe to automatically transcribe audio files uploaded by the analyst. Despite the vast amount of cloud-based speech-to-text services available, we chose the service provided by Amazon as the transcription output contains valuable insights, such as confidence scores, speaker labels, and alternative texts. Once the analyst uploads the audio file with our tool and the

transcription of the file is received from AWS Transcribe, the transcription is available for exploration and analysis by the analyst. Figure 2 shows the framework of the system.
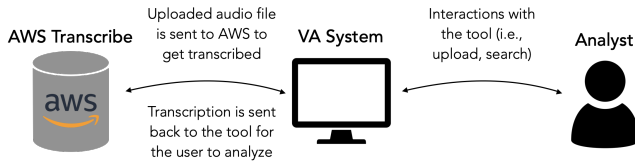

Fig. 2: The framework of **ConFides**. The audio files are uploaded and sent to AWS for automatic transcription. Users can select which transcriptions to explore and analyze.

4.1.1 AWS Transcribe Output

AWS Transcribe offers a line segmentation feature in their transcription output that will divide the transcript if it detects a pause or change in speaker. The service offers up to a maximum of 10 different speakers and will use its speaker diarization algorithm to determine the speaker at the current time in the audio. AWS Transcribe also supports word-level confidence, which is defined as a value between 0 and 1. Based on their documentation, "a larger [confidence] value indicates a higher probability that the identified item correctly matches the item spoken" [2]. Due to confidentiality, AWS does not release any information on how this confidence score is calculated. Therefore, we stress that the confidence score should not be treated as absolute truth but rather as a metric that offers the analyst a way to gauge the potential flaws within the transcription.

4.2 Views

To promote transparency and encourage calibration of trust in the machine's output, each view of the system is embedded with the machine's confidence in the transcription, whether that is term or segment-based.

4.2.1 Confidence Overview

To provide analysts with a visual overview of the transcription data, we designed the following view shown in Figure 1(c). In this view, each rectangle element corresponds to a segment in the transcript, with the width representing the audio length of the segment. Similar to the bar chart in the prototype proposed by Wu et al. [31], the rectangle segments are sequentially ordered from left to right. Each segment's average confidence determines its corresponding rectangle's opacity, meaning the lower the confidence, the more transparent the rectangle will be. This opacity value is calculated by averaging the confidence score associated with every word within the segment.

Within the confidence overview, analysts can rapidly gauge the confidence of segments and playback certain portions of the audio by clicking on the rectangle element associated with the segment (**DG1**). Clicking on the rectangle element prompts the audio player to move and play the corresponding segment instantly. When the analyst hovers over a rectangle element within the view, a tooltip dynamically updates to provide the analyst with specific details on demand about the corresponding segment. This includes information such as the segment's line number, the rolling average of confidence values within the segment, and the segment's text.

4.2.2 Transcription Editor

The transcription view, as seen in Figure 1(f), is designed to assist the analyst in all the tasks listed in Section 3. This view displays the automatic transcription of a singular audio file from AWS. Each line represents a segment of the audio spoken by a specific speaker. We utilize color coding to distinguish speakers. To promote transparency, we provided both a visual representation (shown by the opacity of the underline

– inspired by Vertanen et al. [28]) and a textual representation (the tooltip of each word shows the confidence score) of the confidence for each word within the text (**DG1**). As the audio plays, the transcription will follow along and automatically scroll to the current segment of the audio being played. The segment currently being played will also be indicated by the boldness of the text within the transcription. By utilizing the search bar shown in Figure 1(b), the analyst can query for specific keywords and traverse through all instances of the word within the transcription (**DG3**). The search word is highlighted in yellow within the transcription for easy viewing. The last actionable item the analyst can perform within the transcription view is editing the output. If and when the analyst finds errors within the transcription, the analyst can add or delete text, or even replace a word within the transcription with an alternative suggestion from AWS (**DG2**).

4.2.3 Context Word Tree

Unlike the other two views above, the word tree view will only be populated when the analyst provides a specific keyword they are interested in exploring with the search bar. This view, as seen in Figure 1(d), depicts multiple parallel sequences of words. Based on visualizations such as the Word Tree [30] and Sententree [13], we visualize a node-link diagram where nodes are words and links indicate word co-occurrence within the same segment. The size of the words in this visualization is proportional to the number of occurrences of the word observed. The average confidence of the keyword is also communicated to the analyst within this view.

In addition to the visualization, this view provides the analyst with a list of phonetically similar words – also known as homophones – to the current search keyword. The goal of presenting homophones in this view is to bring awareness of the nuances of the English language, allow the analyst to discern between similar-sounding words, and reduce misunderstandings in spoken and written discourse. The list of homophones shown to the analyst is gathered from an online platform that has curated an extensive collection of phonetically similar words [3].

With the word tree visualization, we aim to show which words most often follow or precede the specific keyword indicated by the analyst (**DG3**). This view will allow the analyst to explore and understand the context in which the specified word is being said. The analyst can also click on any neighboring word to navigate the tree visualization.

5 Case Study: Panda Diplomacy and the Nixon Tapes

During the Nixon Administration, China gifted two pandas – Ling-Ling and Hsing-Hsing – to the US. This case study will use the *Nixon White House Tapes* [1] to learn about the *Panda Diplomacy*. We demonstrate how analysts can use **ConFides** to find relevant data and enable them to answer their key intelligence questions more efficiently. We focus on two questions: "*When were the pandas expected to arrive in the US?*" and "*Which locations were considered for housing the pandas?*" We provide a video [1] walking through the case study. The reader can also click on 🔗 for each question in subsections 5.3.

5.1 About the Nixon White House Tapes

U.S. President Richard Nixon's administration secretly recorded conversations held in the White House from 1971 to 1973. These recordings, infamously as the Nixon White House Tapes surfaced during the Watergate Scandal and ultimately led to Nixon's resignation. The administration installed recording devices in the Oval Office and other locations in the White House, intending to document meetings and conversations for historical purposes and to aid in decision-making. As

---

[1] `https://youtu.be/hbeDn5D-GCg`

these tapes were captured with concealed microphones, the audio quality is often poor, making the task of transcribing and extracting information after transcription strenuous and overwhelming for analysts. This is the ideal scenario for leveraging the strengths of **ConFides** and visual analytics. The analysis of this data can reveal pivotal information, but ASR outputs are fraught with inaccuracies due to the poor recording qualities. Since manually transcribing and analyzing the data is tedious, the analyst can offload the computationally heavy work to the machine and utilize their knowledge and perception to generate insights from the transcription output.
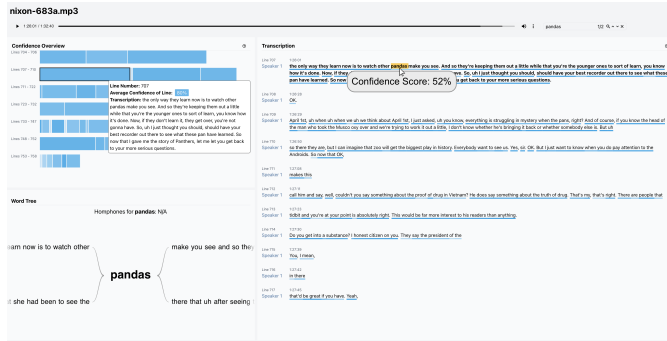

Fig. 3: Searching for "pandas" in the current transcription revealed two instances. We can observe that the first instance of this search term has a confidence score of 52%.

5.2 Reconnaissance, Quality Assessment, and Editing

The analyst begins by searching for instances of "pandas" within the transcription (See Figure 3). The query results indicate that "pandas" is believed to be said only twice within this hour-and-a-half-long audio file. Puzzled, the analyst reads the first segment (line 707) that contains the term "pandas" and observes that the individual confidence in the term "pandas" (52%) is quite low compared to the overall average confidence of this line segment (80%). However, the terms "pan" and "panther" also appear in the segment. Considering that "pan" is a subword/substring of the desired search term and "panther" is phonetically similar to "pandas," the analyst plays the source audio and confirms that "pandas" is indeed inaccurately labeled as "pan" or "panther." The analyst edits the transcription to reflect the audio. We note that there is no search-and-replace feature as a precaution, so the analysts should listen to the audio before making changes to the transcription.
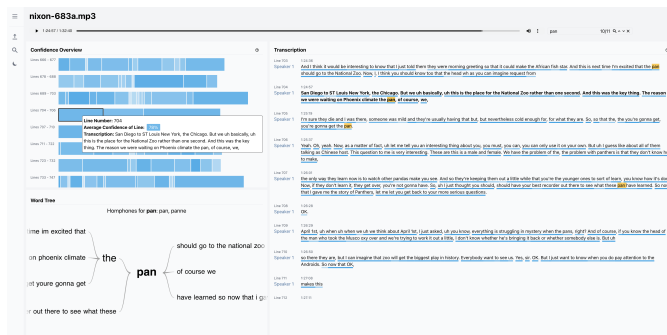

Fig. 4: After searching for "pan," we observe "zoo" in the word tree. This indicates that "panda" was misclassified as "pan" and hints that a zoo may be where the pandas will be kept.

5.3 Extracting Intelligence and Deciding When to Rely on the AI's Output

When were the pandas expected to arrive in the US? 🔗
While updating the transcription, the analyst discovered that

line 709 refers to "pans" and "April 1st," the latter with two instances and a confidence score of 100% for both. The analyst decided the source audio was unnecessary as this line segment showed the pandas arriving in the United States with high confidence on April 1st.

Which locations were considered for housing the pandas? 🔗
As the analyst searches for more instances of "pan," (See Figure 4), they observe that word tree mentions "national zoo." Additionally, line 704 lists several cities (San Diego, St. Louis, New York, and Chicago) known for their zoos. The transcription shows high confidence for the listed cities (ranging from 93% to 100%). Even though the confidence level for this particular segment is relatively high, the analyst decides to cross-check the source audio due to grammatical errors in the transcription. They discovered that despite considering the four listed cities, the pandas will be kept at the National Zoo in Washington, D.C., as one speaker reasoned that it is a tradition that all animals gifted to the United States are homed to this zoo. This fact can be uncovered by searching for "zoo" and observing the first instance of the term, as seen in line 631 in Figure 1.

6 Discussion and Future Work

Through this work, we aim to underscore the importance of AI transparency in fostering and calibrating appropriate trust in visual analytics, especially in scenarios involving domain experts making high-risk decisions. Research has shown that domain experts tend to perform tasks on their own [8, 12] despite the presence of AI assistance. To promote effective usage of AI tools, designers should carefully iterate through intuitive visual representations of uncertainty that do not add cognitive load to the analyst when performing a task [32]. However, even with the communication of model confidence, analysts are likely to engage with AI suggestions differently [25], influenced by their own definition/interpretation of confidence and their criteria for reliability [21]. Further research is needed to explore and develop design guidelines for visual presentations of confidence that are responsive to individual analysts' needs within decision-making tools such as **ConFides**.

In the transcription view of our system, we utilized underlining and color opacity to highlight uncertainty and possible errors within the output. As prior works have found inconclusive findings on the effectiveness of highlighting potential errors to aid in the correction of automatic speech-to-text data [28, 27, 26], future work involves conducting user studies to validate the system's ability to assist analysts in data cleaning and also uncovering relevant data more efficiently. In the future, we would also like to expand upon this work by providing additional assistance to analysts through guided suggestions. Through the observation of the analyst's interactions and their analytical provenance, we can leverage user modeling algorithms [11, 19] to offer the analyst relevant data points of interest in real-time for enhanced exploration and analysis.

7 Conclusion

We developed **ConFides** alongside intelligence analysts to enhance the analyst's understanding of speech-to-text results by providing visual representations of uncertainty in several views of our system. In this paper, we identify clear design goals that guided the development of this system and explore a realistic use case scenario investigating the Nixon White House tapes to demonstrate the system's applicability. Finally, we discuss lessons learned from the development of this system and opportunities for improving textual data cleaning, promoting transparency, and fostering trust. Moving forward, continued efforts to improve the communication of model uncertainty and build appropriate reliance will further enhance the efficacy of human-machine collaborations within visual analytics.

Acknowledgments

References

[1] National Archives, White House Tapes — Richard Nixon Museum and Library, 1999.

[2] Amazon Web Services Inc., AWS Transcribe Documentation, 2023.

[3] A. Aloisi. Homophone Dictionary, 2008.

[4] L. Berke, C. Caulfield, and M. Huenerfauth. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 155–164. ACM, 2017.

[5] K. Brodlie, R. Allendes Osorio, and A. Lopes. A Review of Uncertainty in Data Visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pp. 81–109. Springer London, London, 2012. doi: 10.1007/978-1-4471-2804-5_6

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287

[7] F. Dabek and J. J. Caban. A Grammar-based Approach for Modeling User Interactions and Generating Suggestions During the Data Exploration Process. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):41–50, 2017.

[8] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne. Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):271–280, 2017.

[9] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi: 10.1037/xge0000033

[10] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, Montreal QC Canada, 2018.

[11] S. Ha, S. Monadjemi, R. Garnett, and A. Ottley. A Unified Comparison of User Modeling Techniques for Predicting Data Interaction and Detecting Exploration Bias. *IEEE Transactions on Visualization and Computer Graphics*, 29:483–492, 2023.

[12] S. Ha, S. Monadjemi, and A. Ottley. Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics. *Computer Graphics Forum*, 2024.

[13] M. Hu, K. Wongsuphasawat, and J. Stasko. Visualizing Social Media Content with SentenTree. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):621–630, 2017. doi: 10.1109/TVCG.2016.2598590

[14] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1):108, 2021. doi: 10.1038/s41398-021-01224-x

[15] M. F. Jung, D. Sirkin, T. M. Gür, and M. Steinert. Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2201–2210. ACM, Seoul Republic of Korea, 2015. doi: 10.1145/2702123.2702479

[16] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103. ACM, San Jose California USA, 2016. doi: 10.1145/2858036.2858558

[17] A. Kim, M. Yang, and J. Zhang. *When Algorithms Err* : Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *ACM Transactions on Computer-Human Interaction*, 30(1):1–36, 2023. doi: 10.1145/3557889

[18] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis. Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2021. doi: 10.1109/RBME.2020.3006860

[19] S. Monadjemi, R. Garnett, and A. Ottley. Competing models: Inferring exploration patterns and information relevance via bayesian model selection. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):412–421, 2020.

[20] S. Monadjemi, S. Ha, Q. Nguyen, H. Chai, R. Garnett, and A. Ottley. Guided data discovery in interactive visualizations via active search. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 70–74, 2022.

[21] A. Ottley. *Adaptive and personalized visualization*. Springer Nature, 2022.

[22] A. Piquard-Kipffer, O. Mella, J. Miranda, D. Jouvet, and L. Orosanu. Qualitative investigation of the display of speech recognition results for communication with deaf people. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pp. 36–41. Association for Computational Linguistics, Dresden, Germany, 2015. doi: 10.18653/v1/W15-5107

[23] R. Prasad, L. Nguyen, R. Schwartz, and J. Makhoul. Automatic transcription of courtroom speech. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pp. 1745–1748. ISCA, 2002. doi: 10.21437/ICSLP.2002-520

[24] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016.

[25] T. B. Sheridan. Individual Differences in Attributes of Trust in Automation: Measurement and Application to System Design. *Frontiers in Psychology*, 10, 2019. doi: 10.3389/fpsyg.2019.01117

[26] T. H. Soe, F. Guribye, and M. Slavkovik. Evaluating AI assisted subtitling. In *ACM International Conference on Interactive Media Experiences*, pp. 96–107. ACM, Virtual Event USA, 2021. doi: 10.1145/3452918.3458792

[27] B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8(1):60–98, 2001. doi: 10.1145/371127.371166

[28] K. Vertanen and P. O. Kristensson. On the benefits of confidence visualization in speech recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1497–1500. ACM, Florence Italy, 2008. doi: 10.1145/1357054.1357288

[29] X. Wang and M. Yin. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, pp. 318–328. ACM, 2021.

[30] M. Wattenberg and F. Viegas. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008. doi: 10.1109/TVCG.2008.172

[31] T. H. Wu, Y. Zhao, and M. Amiruzzaman. Interactive Visualization of AI-based Speech Recognition Texts. *EuroVis Workshop on Visual Analytics (EuroVA)*, 2020. doi: 10.2312/EUROVA.20201091

[32] J. Zhou, S. Z. Arshad, S. Luo, and F. Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP conference on human-computer interaction*, pp. 23–39. Springer, 2017.