A Conversational Agent Sensitive to Speaker Styles and Context

Anonymous submission to Interspeech 2025

Abstract

This study presents a novel methodology for developing a chatbot that mimics conversational styles through the integration of linguistic, acoustic, and timing features. Our multi-step approach begins with feature extraction from the Switchboard and Meta Casual Conversations datasets to capture key aspects of human dialogue, including lexical choice, syntactic structure, phonetic properties, and turn taking behaviors. We then embed these features into a consistent numerical representation: transformer based models capture linguistic and acoustic nuances, while timing features are derived from established research. The chatbot adapts its responses in real time by analyzing these embeddings and employing stylistic similarity scoring to align its dialogue with the user's unique conversational patterns. Furthermore, reinforcement learning techniques fine tune the chatbot's performance based on user feedback, ensuring continuous improvement. To evaluate the system's effectiveness, we use clustering based question selection, which groups conversational turns based on their stylistic features. This enables us to assess user engagement and stylistic alignment objectively. This methodology aims to enhance the naturalness and quality of human-computer interaction by enabling dynamic, personalized responses grounded in individual speaking styles.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

The ability to replicate an individual's conversational style is a crucial component in advancing human computer interaction, particularly in creating more natural, engaging, and context sensitive communication [2, 4]. Conversational agents that can dynamically adjust their linguistic, acoustic, and timing characteristics based on the unique patterns of the speaker have the potential to significantly improve interaction quality, making human-computer dialogues more fluid and intuitive [1, 14].

A key aspect of conversational style is emotional expressiveness, which plays a fundamental role in how individuals convey intent, establish rapport, and interpret meaning [11, 6]. Capturing the nuances of emotional expression, such as variations in pitch, speech rate, and word choice, can improve perception of the intelligence, likability, and effectiveness of Conversational agents, fostering greater user engagement [5, 9].

This study aims to explore a computational approach to mimicry of conversational styles, focusing on the development of a chatbot capable of adapting to an individual's speaking style in real time [**?**, 13]. After an initial voice from a human, the chatbot will analyze key conversational patterns and adjust its responses to align with the user's stylistic tendencies [10, 8]. By studying how people naturally communicate, our goal is to create a system that can replicate their linguistic choices, acoustic variations, and timing behaviors, ultimately making interactions feel more human like and personalized [17].

Although conversational style mimicry holds promise for improving human computer interactions, it also raises important ethical considerations [18, 15]. The ability to replicate an individual's speech patterns and emotional expressiveness can have significant implications for privacy, consent, and trust in AI driven communication [20, 21]. Users may not always be aware that an AI system adapts to their style, raising concerns about potential manipulation or unintended emotional influence [16]. Furthermore, there is a risk that this technology could be misused for deceptive practices, such as impersonation or persuasive AI techniques that exploit user vulnerabilities [19]. By carefully balancing technical innovation with ethical responsibility, this research aims to contribute to the development of conversational agents that not only enhance engagement but also respect user autonomy and trust.

2. Related Works

Adapting to different conversational styles has become a critical area of research in the development of conversational agents. One approach focuses on reinforcement learning techniques, which fine tune chatbot responses based on user preferences, improving engagement and coherence in dialogues [19]. These methods enable chatbots to dynamically adjust their behavior in response to individual conversational patterns, leading to more personalized and context-aware interactions.

Another significant advancement uses neural style transfer techniques to modify chatbot dialogues, allowing them to adopt specific linguistic styles. This allows conversational agents to resonate more effectively with users by emulating their speech characteristics [17]. This style transfer approach can be particularly useful in mimicking variations in tone, word choice, and sentence structure, making conversations feel more natural.

Further research has also emphasized the importance of persona-based conversational models. These models personalize chatbot responses by adjusting to predefined personality traits or dynamically learning user preferences. This approach enhances long-term engagement by ensuring consistency in the chatbot's conversational style [15]. The ability to adapt to a user's unique style makes interactions more engaging and contextually relevant, ultimately fostering stronger user connections.

The effectiveness of large-scale pre-trained transformer models has also been explored to enhance chatbot response personalization. These models are capable of detecting subtle linguistic cues and adjusting the tone and formality of the dialogue according to the context [18]. Fine tuned transformer architectures offer a promising solution for improving conversational adaptability across various interaction scenarios.

In addition to linguistic adaptation, recent studies have highlighted the role of contextual awareness in style transfer. By integrating memory-augmented architectures, chatbots can retain previous interactions, adjusting their stylistic choices over time to improve conversational flow. This ongoing refinement enhances the chatbot's ability to engage in more coherent and personalized dialogues [20]. Retaining contextual information allows chatbots to remember and evolve their responses based on previous conversations, improving their long term conversational effectiveness.

Together, these studies illustrate a broad range of techniques that contribute to building more adaptive and engaging conversational agents. By leveraging reinforcement learning, neural style transfer, persona based models, and contextual awareness, the goal is to create chatbots that can mirror the user's conversational style while offering a personalized and meaningful interaction experience.

3. Data

This research leverages two primary datasets to capture diverse aspects of conversational style. The Switchboard dataset and the Meta Casual Conversations dataset.

The Switchboard dataset [2] comprises approximately 2,400 telephone conversations (260 hours) from American speakers representing a range of ages, genders, regions, and other demographic factors. We believe this dataset is well suited for examining turn-taking behaviors, lexical choices, and syntactic structures in a controlled setting. From this corpus of spontaneous dialogues, we extracted various linguistic and paralinguistic features relevant to conversational style, such as speech rate (calculated based on the length and timing of the utterance) and utterance length (total word count per utterance), to analyze the complexity and characteristics of the conversations.For each utterance in the Switchboard corpus, we used BERT's SentenceTransformer [CITE?] against the transcript to derive a 768-dimensional linguistic vector. We channel-separated the audio and used Audio2Vec [CITE?] extracted another 768-dimensional vector representing the acoustic realisation of the utterance. We concatenated these two vectors, along with timing features, such as:

- Pause after previous utterance
- OTHER TIMING FEATURES HERE

The Meta Casual Conversations dataset provides a more ecologically valid setting, as described in [1], containing 45,186 videos of 3,011 participants recorded in natural environments. Each video lasts an average of 70 seconds and was filmed in $1080p$ quality. Critically, all participants provided consent for their data to be used for machine learning purposes. The dataset is divided into CC_part and CC_mini_part, each composed of uniformly sampled videos of participants across annotations. This dataset allows us to analyze acoustic and visual cues associated with conversational style, such as prosodic features (pitch, loudness variation) and nonverbal communication, complementing the linguistic features extracted from the Switchboard dataset.

Both datasets are leveraged to extract key features related to linguistic and paralinguistic cues, informing the development and evaluation of our conversational agent. By combining the controlled telephone conversations of Switchboard with the naturalistic videos of Casual Conversations, we aim to capture a holistic view of conversational style that encompasses both verbal and nonverbal elements.

4. Methodology

This study employs a multi-step approach to conversational style mimicry, leveraging linguistic, acoustic, and timing features to develop a chatbot capable of adapting to a user's unique speaking patterns. Our methodology consists of feature extraction, embedding representation, chatbot development, and evaluation through clustering-based question selection.

Feature extraction captures conversational style from the Switchboard dataset, a corpus of spontaneous telephone conversations, and the Meta Casual Conversations dataset, which provides more varied conversational styles [5]. The extracted features fall into three domains: linguistic features that capture lexical choice, syntactic structure, and discourse patterns [11], acoustic features that analyze phonetic properties such as pitch, loudness variation, and speech rate [8], and timing features that measure aspects like inter-utterance pauses and turn-taking behavior [2]. Linguistic features are embedded using a pre-trained transformer-based language model, acoustic features are processed using an audio transformer model [19], and timing features are derived from prior research on conversational style labels and analysis in the Switchboard dataset [12].

To ensure a consistent numerical representation of conversational features, both linguistic and acoustic embeddings are mapped into a 768-dimensional vector space. Linguistic embeddings are generated by a transformer-based language model, encoding each utterance into a 768-dimensional feature vector to capture syntactic and semantic properties [15]. Acoustic embeddings are extracted through an audio transformer model that captures phonetic and prosodic characteristics. These embeddings are computed using a weighted aggregation of frame-level features, represented as:

$$E_{audio} = \sum_{i=1}^{n} w_i \cdot f_i \tag{1}$$

where represents the feature vector for frame , is the corresponding weight based on attention mechanisms, and is the total number of frames in the audio segment [17]. This ensures that linguistic and acoustic features share the same dimensional space, enabling multimodal integration for conversational style adaptation.

Feature vectors are mathematically represented as follows:

$$\mathbf{f}_L = [f_{L1}, f_{L2}, \ldots, f_{Ln}] \tag{2}$$

$$\mathbf{f}_A = [f_{A1}, f_{A2}, \ldots, f_{Am}] \tag{3}$$

$$\mathbf{f}_T = [f_{T1}, f_{T2}, \ldots, f_{Tp}] \tag{4}$$

where f_L represents the linguistic feature vector, $f_A$represents the acoustic feature vector, and f_T represents the timing feature vector. The lengths of these feature vectors are n,m and p respectively.

The chatbot module, implemented in chat.py, dynamically adjusts its conversational responses based on extracted style embeddings. After an initial vocalization from a user, the chatbot analyzes linguistic and acoustic patterns to generate responses that align with the user's speaking style [21]. Conversational adaptation is achieved through stylistic similarity scoring, which compares extracted embeddings to learned style profiles [7], and response selection, which ensures responses are relevant in both content and style [9]. Reinforcement learning techniques fine-tune chatbot responses based on user feedback, facilitating continuous adaptation and improvement [18].

To assess the chatbot's effectiveness in style adaptation, we use a clustering-based question selection method. This approach groups conversational turns based on linguistic, acoustic, and timing features to measure the degree of stylistic alignment between chatbot responses and user utterances [13], the impact of conversational adaptation on user engagement and satisfaction [16], and the chatbot's ability to generalize across different conversational styles and demographics [20]. The clustering objective function is defined as:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} |x_j - \mu_i|^2 \quad (5)$$

where k is the number of clusters, and C_i represents cluster i, and x_j is a data point in the cluster, and u_i is the centroid of the cluster. This ensures effective grouping based on conversational feature similarity.

This methodology provides a structured framework for developing an adaptive conversational agent that mirrors human-like interaction patterns, thereby enhancing the quality and naturalness of human-computer dialogue.

5. Results

To evaluate the performance of our chatbot, we conducted a series of test cases where input audio was recorded, transcribed, and compared against the Switchboard dataset. The chatbot's linguistic and acoustic matching capabilities were analyzed based on cosine similarity scores, and responses were generated accordingly. Below, we present six sample interactions in a structured table, followed by a deeper analysis of the results and their implications.

5.0.1. *Sample Interactions*

Query	**Ling. Sim.**	**Acous. Sim.**	**Matched Text**	**Generate**
What do you think about climate change?	0.87	0.82	Climate change is a serious issue that needs to be addressed.	Climate c a critical requiring action. T renewabl
How was your weekend?	0.92	0.89	My weekend was great! I went hiking and enjoyed the fresh air.	It was wo I got outs about you
Do you prefer tea or coffee?	0.78	0.75	I like coffee in the morning, but I enjoy tea in the evening.	It depend of day! W you?
Have you seen the latest movie in theaters?	0.81	0.79	Yes, I just saw it last weekend. It was really entertaining!	Yes, I ca cently. D it?
Tell me about your favorite book.	0.88	0.85	I enjoy reading science fiction, especially space exploration.	I love sci anything
What's your opinion on artificial intelligence?	0.90	0.84	AI is rapidly evolving with the potential to transform industries.	AI is fasc What asp ests you

Table 1: *Sample chatbot interactions with linguistic and acoustic similarity scores.*

The results demonstrate the chatbot's ability to generate contextually appropriate responses by leveraging both linguistic and acoustic embeddings. Higher linguistic similarity scores consistently yield highly relevant responses, while lower scores result in more generalized but still contextually appropriate replies. This suggests that linguistic embeddings alone, while effective, may not always ensure precise response generation.

Acoustic similarity significantly enhances the perceived naturalness and personalization of the chatbot's responses. Higher acoustic similarity scores correspond to responses that align more closely with the user's speech patterns, contributing to an engaging conversational experience. Conversely, lower acoustic similarity scores can introduce subtle mismatches in tone, potentially making interactions feel less natural. This underscores the importance of refining acoustic embeddings to improve speaker-specific adaptation.

An important observation is that even when linguistic similarity is moderate, a high acoustic similarity score can yield responses that feel more personalized. This suggests that acoustic embeddings play a crucial role in mimicking speaker-specific styles and should be further optimized in future iterations.

Additionally, the chatbot's ability to generate follow-up questions highlights its contextual awareness, enabling conversations to feel dynamic rather than merely reactive. By balancing semantic accuracy with personalized conversational nuances, the system effectively fosters more human-like interactions.

Overall, these findings support the effectiveness of hybrid embedding techniques in chatbot response generation. The interplay between linguistic and acoustic features enhances both relevance and speaker adaptation, underscoring the potential for future advancements in conversational AI to further refine response coherence and user engagement.

6. Discussion

While this study lays a crucial foundation for conversational style mimicry using linguistic, acoustic, and timing embeddings, several key areas warrant future exploration to enhance the adaptability and real-world effectiveness of our approach.

A primary direction for future work involves scaling our models to encompass larger and more diverse datasets. The Switchboard dataset, while robust, is limited in scope to American English telephone conversations. To develop a more generalizable and context-aware conversational agent, we intend to incorporate datasets that capture a broader range of demographics, speaking styles, and conversation types. Such expansion will enable the model to learn more nuanced conversational behaviors and improve its adaptation capabilities across varied real-world contexts.

Furthermore, we plan to refine our approach to analyze localized conversational styles. Speech patterns, pacing, and word choice vary significantly with regional dialects, cultural norms, and professional settings. Integrating models accounting for these localized nuances promises AI-driven agents that feel more natural and contextually appropriate in diverse settings, enabling more accurate and personalized interactions in domains like customer service, education, and healthcare.

Another critical area of expansion is multilingual conversational style analysis. Our current framework focuses on English-speaking interactions, yet conversational style is deeply intertwined with language structure, intonation, and cultural norms. Exploring conversational style mimicry across multiple languages will necessitate training models on multilingual speech datasets and developing techniques to map style features while preserving the unique elements of each language.

Finally, we acknowledge the limitations of the Switchboard dataset regarding conversational depth and subject matter. Future work will explore how conversational style adapts across a wider range of topics, including technical discussions, academic discourse, and specialized communication. By incorporating datasets featuring more structured and domain-specific language, we aim to develop agents capable of adapting to both an individual's speaking style and the complexities of different subject matters. This is particularly relevant for applications requiring precise terminology and structured dialogue, such as medical consultations or educational tutoring.

Throughout this research, we remain committed to addressing the ethical considerations associated with conversational adaptation, particularly in multilingual and cross-cultural contexts. Ensuring that style mimicry respects user expectations, avoids reinforcing stereotypes, and maintains transparency is paramount to our ongoing work.

By pursuing larger datasets, localized and multilingual nuances, topic-based adaptations, and a firm ethical grounding, we aim to push the boundaries of conversational AI, creating more intelligent, adaptive, and responsible systems for human-computer interaction.

7. References

[1] P. Brown and S. C. Levinson, *Politeness: Some universals in language usage*. Cambridge University Press, 1987.

[2] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Human conversation as a system framework for modeling multimodal communication," 1999.

[3] H. H. Clark, *Using language*. Cambridge University Press, 1996.

[4] H. P. Grice, "Logic and conversation," in *Syntax and Semantics, Volume 3*, P. Cole and J. L. Morgan, Eds. Academic Press, 1975, pp. 41–58.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Prentice Hall, 2021.

[6] G. Leech, *Principles of Pragmatics*. Longman, 1983.

[7] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[8] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[9] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1996.

[10] T. Sinha and J. Cassell, "Weaving conversational threads: Rethinking chatbots as conversational partners," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2015, pp. 1096–1100.

[11] D. Tannen, *That's Not What I Meant!: How Conversational Style Makes or Breaks Relationships*. HarperCollins, 1986.

[12] D. Tannen, *Gender and Discourse*. Oxford University Press, 1994.

[13] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *Proceedings of the ACL*, 1997, pp. 271–280.

[14] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[15] A. P. Chaves, J. Egbert, T. Hocking, E. Doerry, and M. A. Gerosa, "Chatbots language design: The influence of language variation on User Experience," arXiv.org, 2021, January 26. Available at: https://arxiv.org/abs/2101.11089.

[16] R. Corizzo, E. Zdravevski, M. Russell, A. Vagliano, and N. Japkowicz, "Feature extraction based on word embedding models for intrusion detection in network traffic," *Journal of Surveillance, Security and Safety*, 2020, December 28. Available at: https://www.oaepublish.com/articles/jsss.2020.15.

[17] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," arXiv.org, 2019, August 13. Available at: https://arxiv.org/abs/1904.02760?utm.

[18] S. Ibrahim and G. Bilquise, "Beyond chatgpt: Benchmarking speech-recognition chatbots for language learning using a novel decision-making framework," *Education and Information Technologies*, SpringerLink, 2024, December 27. Available at: https://link.springer.com/article/10.1007/s10639-024-13252-3?utm.

[19] S. Liu, Z. Leng, and D. Wijaya, "Learning to mirror speaking styles incrementally," arXiv.org, 2020, March 5. Available at: https://arxiv.org/abs/2003.04993?utm.

[20] Md. A. Mamun, H. Md. Abdullah, Md. G. R. Alam, M. M. Hassan, and Md. Z. Uddin, "Affective social anthropomorphic intelligent system," arXiv.org, 2023, April 19. Available at: https://arxiv.org/abs/2304.11046?utm.

[21] K. Metcalf, B.-J. Theobald, G. Weinberg, R. Lee, I.-M. Jonsson, R. Webb, and N. Apostoloff, "Mirroring to build trust in Digital assistants," arXiv.org, 2019, April 2. Available at: https://arxiv.org/abs/1904.01664.

[22] D. Madan and S. Joshi, "Finding Dominant User Utterances And System Responses in Conversations," arXiv.org, 2017, October 29. Available at: https://arxiv.org/abs/1710.10609.