FAME: Exploring Expressive Facial Avatars for Lyrical and Non-Lyrical Music Visualization for d/Deaf Individuals

Suhyeon Yoo*
Computer Science
University of Toronto
Toronto, Ontario, Canada
suhyeon.yoo@mail.utoronto.ca

Yifang Pan*
Dynamic Graphics Project
University of Toronto
Toronto, Ontario, Canada
evan.pan@mail.utoronto.ca

Ashish Ajin Thomas
Computer Science
University of Toronto
Toronto, Ontario, Canada
ashish.ajinthomas@mail.utoronto.ca

Karan Singh
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
karansher.singh@utoronto.ca

Khai N. Truong
Computer Science
University of Toronto
Toronto, Ontario, Canada
khai.truong@mail.utoronto.ca

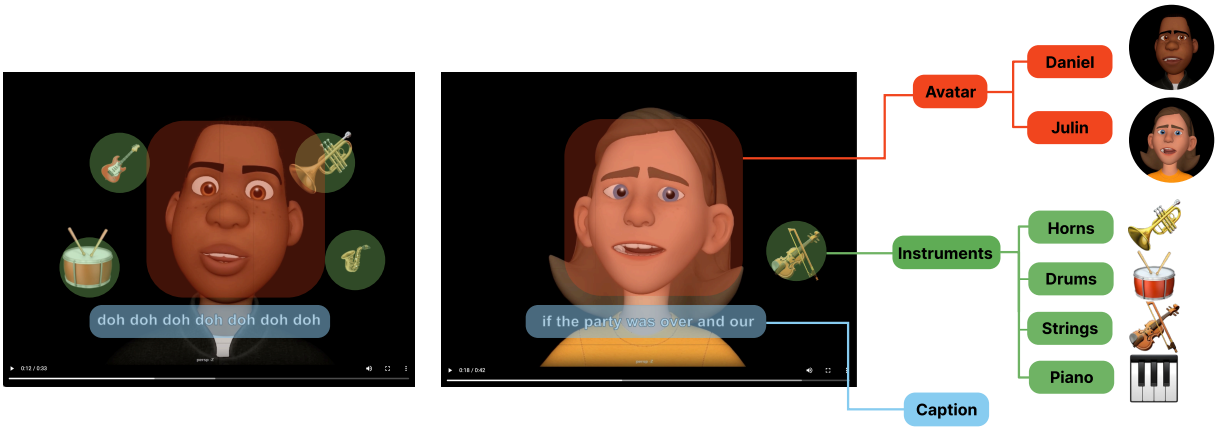

Figure 1: The FAME design probe provides multimodal visualizations of music for DHH users through an expressive facial avatar. Avatars (left: Daniel visualizing "Uptown funk - Bruno Mars", right: Julin visualizing "Die with a smile - Lady Gaga") convey lyrics, rhythm, and emotion using facial expressions, lip-sync, and body movements. Additional layers include instrument highlights (e.g., horns, drums) and captions, which can be selectively combined to support different listening preferences.

Abstract

d/Deaf and Hard of Hearing (DHH) individuals often engage with music through a multimodal approach, where visual modalities are also used rather than relying on sound alone. While tools like captions and visualizers offer partial support, they often fail to capture the emotional depth and structural nuances of music. To explore new possibilities, we adopted an iterative, probe-based approach. Through a formative study with 9 DHH participants, we identified key design requirements for visualizing rhythm, emotion, and lyrics. We developed FAME (Facial Avatar for Musical Expression), a design probe that conveys music through expressive facial animation, instrument highlights, and synchronized captions, lip-syncing to lyrics or scat-singing to melodies. Through a two-phase exploratory study with 12 DHH users, we examined FAME's efficacy, applicability, and requirements for representing musical elements. Our findings refine design requirements for avatar-based systems and highlight the potential of avatars as expressive and socially meaningful tools for music accessibility.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools**; **Empirical studies in accessibility**.

Keywords

Deaf Music, Facial Avatar, Music Visualization, Scat Singing

ACM Reference Format:
Suhyeon Yoo, Yifang Pan, Ashish Ajin Thomas, Karan Singh, and Khai N. Truong. 2026. FAME: Exploring Expressive Facial Avatars for Lyrical and Non-Lyrical Music Visualization for d/Deaf Individuals. In *Proceedings of*

*Both authors contributed equally to this research.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790402

the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3790402

1 Introduction

Music is a multimodal art form that combines rhythm, melody, and emotional expression. d/Deaf and Hard of Hearing (DHH) individuals often engage with music not only through sound but also via visual and tactile modalities, such as vibrations from speakers, captioned lyrics, and sign language interpretation [26, 34]. In social settings, DHH individuals often draw on cues from how others around them interact with music, such as expressive head movements, dancing, and singing along, to perceive rhythm, energy, and mood [19, 99]. They often attend to a performer's facial expressions, lip movements, and gestures, which offer additional cues for understanding emotional and structural qualities of music [4].

Despite the diverse ways DHH individuals experience music, most research and commercial accessibility efforts have primarily focused on mapping quantifiable musical elements (e.g., pitch, rhythm, volume) to visual or tactile feedback [35, 65, 99]. However, these approaches often fall short in conveying key aspects of music, such as emotional and structural information [118]. Vibrotactile systems are frequently limited in expressive range or require specialized hardware, while visualizers tend to rely on abstract representations that lack embodied context [77, 84, 89].

Recent advances in real-time avatar animation [50, 94] and generative systems [21, 44] have opened new possibilities for visualizing music through animated characters that lip-sync [8], express emotions [70], and move in sync with rhythm [5]. Prior research has explored animated singing heads to convey music and lyrics, emphasizing facial expression and emotional tone [86, 107, 117]. However, several gaps remain. First, although lip-sync is often emphasized, prior work shows that lipreading can be cognitively demanding and is rarely sufficient for comprehension without supporting context [13, 102]. Second, very few of these avatar systems have been co-designed or evaluated with DHH users, leaving questions about their usability, expressiveness, or alignment with everyday musical practices. Third, much of the existing work focuses narrowly on pop music with lyrics [106], while excluding non-lyrical genres that can be significant to many DHH individuals [118].

To address these gaps, we adopt an iterative, probe-based research-through-design approach. Our goal is to understand how avatar-based visualizations can amplify the strategies DHH individuals already use to engage with both lyrical and non-lyrical music. We structured our work in two stages. In the first iteration, a study with 9 DHH individuals examined emotional responses and reactions to an early avatar prototype, yielding high-level requirements for avatar appearance, animation style, and multimodal cues such as captions and instruments. In the second iteration, we developed a refined probe, **FAME (Facial Avatar for Musical Expression)**, which operationalized these requirements. FAME presents avatar-based music videos with additional features such as instrument highlights, captioned lyrics, and scat-style vocalizations, nonsensical syllables paired with expressive singing, to convey melodies.

We then conducted a two-phase user study with 12 DHH participants to explore how well these design criteria worked in practice and to elicit feedback on gaps and desired improvements. In the comparison phase, participants provided insights about how FAME improved recognition accuracy and comprehension of lyrical and emotional elements over a visualizer baseline. In the application phase, participants envisioned avatars as performers, interpreters, and companions, and suggested new requirements for avatar appearance, body movement, and background visualization. In doing so, we ground our work in a cultural model of disability [11, 12] and perspectives from aural diversity [37, 49], expanding visual musical expressivity in ways that align with Deaf cultural performance.

Taken together, our work makes the following contributions:

- An understanding of DHH users' musical preferences, accessibility strategies, and attitude towards avatar-based visualization for both lyrical and non-lyrical music.
- A demonstration of the effectiveness of avatar-based representations for enhancing comprehension of lyrics, emotion and enjoyment through expressive cues.
- Design requirements for an avatar-based system that visualizes musical elements (emotion, lyrics, rhythm) through facial and bodily gestures, with lip-syncing lyrics and non-lyrical scat-style vocalization.

Positionality Statement. Our team comprises researchers with expertise in accessibility and human–computer interaction, as well as expressive avatars and computer graphics. All members of the team identify as hearing, and we approached this work with an awareness of how this positionality shapes our interpretation of Deaf and hard-of-hearing (DHH) experiences. We drew on critical disability perspectives [95] throughout the research process.

2 Related Work

2.1 Music Experience of DHH individuals

d/Deafness exists on a spectrum [1] and as such, the musical experiences and preferences of Deaf and Hard of Hearing (DHH) individuals vary widely [34, 118]. Many DHH individuals describe music as a multi-sensory, spatio-temporal experience that integrates visual, tactile dimensions alongside auditory input [16, 17]. Kolb's reflections on experiencing music with cochlear implant, *Sensations of Sound* [61], similarly emphasize perception through vibration, spatial cues, and visual movement. Prior studies have found that DHH listeners tend to prioritize rhythm, timing, emotional resonance, and especially lyrical understanding, particularly when the song is already familiar to them [34, 109].

One expressive art form is song signing, in which performers visually interpret lyrics using sign language, along with facial expressions and spatial gestures to convey musical elements [32, 73]. While this approach is primarily designed for culturally Deaf audiences, it can also be applied with lyrics captions for people who are deaf or hard of hearing who are unfamiliar with ASL [109]. These performances allow DHH viewers to engage with the musical narrative by visually conveying mood, beat, and meaning [74, 109].

[1] In research, the acronym DHH (Deaf and Hard of Hearing) is commonly used to describe individuals with a range of hearing levels and communication preferences [45]. The term "deaf" (lowercase "d") typically refers to individuals with a clinical diagnosis of hearing loss, who may use assistive technologies such as hearing aids or cochlear implants, and may rely on spoken language or lip reading. In contrast, "Deaf" (capital "D") refers to individuals who identify culturally as members of the Deaf community, use sign languages and embrace Deaf culture and identity [93].

In social contexts, DHH individuals frequently observe friends, family members, and others engaging with music (e.g., singing, dancing), as a way of understanding the rhythm and emotional tone of a piece [16]. These embodied, interpersonal cues serve as valuable scaffolding for music perception, especially in group settings [19, 99]. Additionally, individuals often develop personalized heuristics for engaging with music, through songs that are instrumental or rhythm-heavy, and relying on visual and haptic feedback to help them perceive the music [118]. While some commercial haptic technologies, such as the Music Not Impossible vest [52] and the Cutecircuit Soundshirt [33], have been developed to enrich musical experiences for DHH users, their high cost can limit accessibility.

2.2 Music Visualization and Accessibility

For lyrical music, captioning is a well-studied visualization strategy. Many captioning works from the HCI community focus on enhancing captions using expressive typography, modulating font size, weight, motion, and layout to reflect pitch, timbre, and volume [18, 46, 58]. Other systems, like EnACT [104] and its follow-up work [80] also experimented with conveying the emotional tone. Works from DHH artists and scholars [98] offer a crucial perspective on how access is deeply integrated into their holistic experience. Olivia Ting [98], for example, explores the multi-modal and fragmented nature of deaf musicality in *Song Without Words* using "twitchy," "smudged" captions alongside haptic radios. Similarly, Louise Hickman's film *Captioning on Captioning* [47] and Christine Sun Kim's film *Closer Captions* [59] critique hearing-centered notions of access, reframing captioning not as an automated utility but as an intimate, care-based, and embodied form of human labor. While some HCI work examines how captions can be consumed along with other media, such as background visualizations [58] or music videos [72], these forms of integration often lack the deep, lived-experience-driven perspective demonstrated by DHH creators. This gap highlights a clear opportunity for in-context studies that feature co-design with members of the DHH community. Further, the visualization of non-lyrical music remains underexplored, limiting the applicability of caption-based approaches.

An alternative approach to music visualization leverages graphics that directly map audio features to visual elements. These systems use algorithmic analysis—such as frequency, amplitude, and temporal data—to generate visual representations like time plots, tonal landscapes, or spectrograms [81]. One example is the work by Pouris and Malekian, who visualized each pitched note as a moving 3D cylinder with varying height [87], while Szucs and Kozek developed a multi-instrument visualizer that displayed virtual instrument models alongside dynamic bar graphs representing pitch ranges in 3D [97]. Building on the established associations between pitch and color [53], many systems also use color to encode musical notes or octaves. Prior work has visualized music using colored particles or morphing 2D geometries to melody and rhythm [30, 42, 64].

Despite a growing interest in music visualization, there is limited research assessing whether these systems provide sufficient informational and emotional value for DHH users. Few literature tends to focus on developing new visualization prototypes [18, 35, 46]. For instance, ViTune [35] maps different pitch ranges to a 2D virtual piano keyboard, helping DHH users understand musical structure. Similarly, systems like [87] use 3D visuals to engage users through dynamic representations of notes and instruments. However, prior studies suggest that while visually appealing, many music visualizers overlook emotional and narrative dimensions, making them less effective for appreciating lyrical or expressive pieces [43, 87]. Nanayakkara *et al.* demonstrated that beat-synchronized lip movements (e.g., repeating "ba") of a performer can support DHH users' perception of musical timing, whereas conductor-like full-body human gestures communicate phrasing and emotional nuance more effectively than abstract displays. [4]. These findings show that embodied cues can enhance music perception for DHH users, yet they have not been studied in avatar-based visualizations.

2.3 Facial Avatar-based Music Performance

The emergence of audio-driven generative models for lip-sync animation [39, 96, 117] has opened new possibilities for designing music visualizers in accessibility contexts. Lip-sync animations, in particular, have shown potential as effective "visual hearing aids" for DHH individuals by visually conveying the timing and articulation of lyrics [28]. Compared to traditional abstract visualizers, Wang *et al.* demonstrated that avatar-based visualizations are significantly more effective at communicating the emotional tone and mood of music for DHH participants [106].

Recent developments in learning-based lip-sync generation models in academia [39, 96, 117] show promising potential as speech visualizers. Models such as Media2face [117], primarily trained on speech corpora, have demonstrated the ability to generalize to singing. However, due to the limited representation of singing examples in the training data, these models can generate realistic animations but lack the fine-grained controllability needed for precise emotional expression. SingingHead [107] addresses part of this gap by constructing an extensive dataset of singing and training an auto-regressive generative model. Unfortunately, as this dataset contains only neutral singing, the emotion dimension cannot be effectively authored or controlled.

In contrast, procedural lip-sync models based on visemes offer superior flexibility for constructing visualizations that communicate the rich information embedded in songs. The commercial system JALI [39] decouples the control of visemes (i.e., mouth shapes for different sounds) from emotional motion, enabling precise editing of emotional intensity and quality. Similarly, Pan *et al.*'s adaptive framework VOCAL uses a rule-based system that automatically generates expressive viseme-based lower-face animations, capturing the fine-grained technical details unique to singing, such as pitch variation, vowel modification, enunciation, and vibrato [86]. These singing-specific nuances are precisely what make procedural approaches more suitable for our context, where communicating the emotional richness of musical performance is paramount.

Complementary to these efforts, a growing body of academic work has explored systems that support music performance in virtual environments. Many of these systems focus on enhancing the social experience of music by creating a sense of "social presence" [51, 56, 103]. Virtual avatars in these environments have been shown to improve audience immersion, emotional engagement, and perceived co-presence during performances [29, 54]. Some of this work focuses on avatar-driven virtual concerts [20, 66], where the

Table 1: Comparison of avatar-based music visualization systems in terms of conveyed musical elements. FAME expands the expressiveness beyond emotion to include lyrical, rhythmic, and instrumental elements.

System	Emotion	Lyrics	Beats	Pitch
VOCAL (Siggraph '22)[86]	–	lip-sync	–	mouth opening size
Music-to-Facial (AAAI '23)[106]	facial expressions	–	–	–
SingingHead (Arxiv '23)[107]	–	lip-sync	head motion	–
Media2face (Siggraph '24)[117]	facial expressions	lip-sync	head motion	–
Initial Probe	facial expressions	lip-sync	head motion	mouth opening size
Refined Probe (FAME)	facial expressions	lip-sync + caption	head + body motion	mouth opening size

goal is to enhance audience participation and emotional resonance [100, 108]. Other systems integrate singing avatars into collaborative or karaoke-style settings [14, 25], employing MIDI to generate expressive avatar performances in real-time.

Our work builds on this body of research through a two-stage, probe-based approach (See Table 1). We designed our probe to incorporate concepts explored in prior avatar and music visualization work, such as lip-sync [86], expressive facial movement [39], and rhythmic head motion [6], as a means of providing DHH participants with a rich, integrated experience of what avatars can offer. By grounding our probe in the existing literature, we were able to examine how these strategies collectively support music accessibility for DHH users, while also surfacing new design requirements for representations that effectively communicate musical meaning.

Figure 2: Early probe: Avatar visualizations for contrasting songs. (Left) A joyful rendition of 'Happy' by P. Williams. (Right) A somber rendition of 'Someone Like You' by Adele.

3 Study 1 (Formative)

The design of probes was guided by a disability centered and interdependence oriented approach [12, 48], emphasizing early and sustained engagement with d/Deaf and Hard of Hearing (DHH) individuals throughout the design and evaluation stages. This approach responds to broader critiques of the systemic exclusion of DHH people from digital accessibility research.

As the first step in our iterative, probe-based approach, we sought to surface preliminary design requirements for an avatar-based music visualization system. We conducted a formative study with DHH individuals to better understand how they experience music, and to gather their reactions on an early exploratory probe: a simple avatar singing along to different music (See Figure 2). Rather than aiming to evaluate a finished system, this probe served as a catalyst for reflection and discussion. Through semi-structured interviews, we investigated participants' preferences, challenges, and expectations related to how avatars might represent musical elements to identify requirements to guide the design of a more refined probe.

3.1 Methods

3.1.1 Participants. We recruited nine DHH individuals (U1 – U9[2]), including five who identified as Deaf and four as Hard of Hearing (See Table 2). Participants ranged in age from 27 to 46 years. All participants reported using assistive technologies such as hearing aids or cochlear implants and were fluent in both American Sign Language (ASL) and English. Inclusion criteria required participants to (1) self-identify as Deaf or Hard of Hearing, (2) be 18 years of age or older, (3) have prior experience using music visualization tools (e.g., captions, visualizers), and (4) be familiar with basic musical elements such as rhythm, pitch, or melody. Recruitment was conducted via email outreach, social media platforms (e.g., Facebook groups and Reddit channels), and snowball sampling.

3.1.2 Procedure. Each participant was invited to a Zoom session lasting between 50 and 70 minutes. We ensured accessibility by using Zoom's chat feature, closed captioning, and sign language interpreters for DHH participants who preferred to sign. The interview consisted of two parts. In the first part, participants were asked about their typical music practices, tools they use, and challenges they face. In the second part, participants watched eight avatar-generated music videos and provided feedback on each. They were asked to reflect on the effectiveness of the avatars in conveying musical elements, as well as to suggest potential improvements.

3.1.3 First Probe: Avatar Videos. We generated avatar-based videos for 8 selected songs (See Table 5). The selection criteria included songs with high arousal levels, balancing 4 lyrical and 4 non-lyrical songs, as well as 4 with positive valence and 4 with negative valence, to ensure diversity in both emotional and lyrical content.

The avatar videos were generated and rendered with Autodesk Maya, using the ValleyVillage Suite of characters from JALI [39], which offers photo-realistic human models from the neck up. The suite includes six characters (three male, three female) with African, Asian, and Caucasian features. We ensured demographic diversity by utilizing all six characters across our stimuli (See Table 6).

We generate the motion of the avatars procedurally. For lyrical videos, lip-sync and facial expressions were automatically generated using the JALI plugin [39], which converts emotionally-tagged

[2]We denote participants in this formative study as U1, U2, … to distinguish them from participants in the main study, who are denoted as P1, P2, …

Table 2: Formative study participant information ($N = 9$). Values represent total counts unless otherwise noted.

Description	Details
Group	Deaf (5), Hard of Hearing (4)
d/Deaf Onset	Born d/Deaf (7), Late-deafened (2)
Age	Mean = 36.6, Median = 35, Min = 27, Max = 46
Gender	Female (8), Male (1)
Music Visualization Usage	Daily (6), Weekly (1), Monthly (2)
Assistive Technology	Hearing aids (8), Cochlear implants (1)
Communication	Sign or talk (8), Sign and talk (1)

audio into emotional facial expressions, and lip movements reflecting the "visemes" (lip shapes during sound production) of the different "phonemes" (sounds). For videos with positive valence, we tagged the generation with the emotion "joy" at 100%, while for low valence, we used the tag "grief" at 100% (See Figure 2). For non-lyrical videos, we omitted the lip-sync motion as it would not correspond with the audio. We further generated head bobbing motion by using MediaPipe [71] to obtain head rotation from an actor instructed to sing or hum along to the music.

3.2 Findings and Design Requirements

We identified three major themes from our formative interviews with DHH participants: their diverse music preferences, the essential but limited role of visualization tools, and perceptions of avatar-based music visualization. In this section, we present these findings along with the corresponding design requirements.

3.2.1 Captioned Lyrics Can Contribute to Understanding of Meaning and Timing. Many participants reported using captioning features on platforms like Spotify and YouTube. They found real-time captions and karaoke-style displays especially helpful for understanding lyrics and tracking musical timing. However, they also noted inconsistencies in caption availability, synchronization, and accuracy across platforms. As U9 described, *"There were different options for every song, like the video, the video with captions, and then just the lyrics. It would be interesting to have it all in one place."*

Although captions were seen as essential for comprehension, they are insufficient on their own. Participants expressed the need to complement the avatar with cues about which lyrics are sung, which instrument is playing, and even to have sign language overlays. As U8 described, *"I need to know which part is sung. . . captions don't always help. Refer to 'Andy Mineo - Hear My Heart'* [3]*. I really like how it visualizes the instruments!"*

DR1: Combine Captions and Instrument Cues with Avatar Performance. Avatars should integrate captioned lyrics and instrument highlights to complement their expressions. These complementary cues provide richer context and better support both comprehension and engagement.

3.2.2 Avatars Show Promise but Need Greater Expressiveness and Clarity. Participants saw potential in avatar-based music visualizations to convey emotion, rhythm, and lyrical content. Facial expressions, head movements, and lip-sync were recognized as valuable cues for interpreting music. As U3 noted, *"When I can see the lips moving. It's like, you're putting on the performance with the lyrics. So I can relate to this avatar."* However, current avatars often lacked sufficient emotional depth and motion variation. As U4 noted: *"The head should be moving more! Like the way Bruno Mars did. It felt like stale… but this song is rocking pop hard."* Despite their realism, the avatars felt artificial to some users, landing in the "uncanny valley." As U9 described, *"I feel indifferent to it… It's very clear that they are an avatar. It's not lifelike."* Participants also noted that the avatars lacked sufficient emotional range and intensity. *"They didn't seem that happy…that's a very upbeat song."* (U9)

DR2: Avoid Photorealism and Emphasize Expressive Clarity. Rather than striving for photorealism, a more stylized or cartoon-like appearance might feel less uncanny and more emotionally legible. Furthermore, avatars should emphasize clear, exaggerated facial and bodily expressions to communicate musical mood.

3.2.3 Different Musical Elements are Prioritized Depending on the Music. We found that DHH participants prioritized different musical elements, such as beat, lyrics, and instrumentals, depending on whether music was lyrical or non-lyrical. For lyrical music, lyrics were often seen as the most important, supported by captions or familiarity. As U4 shared, *"Beat and lyrics. Because I can hear the beat, and I can follow the lyrics to get the meaning of songs."* In contrast, non-lyrical music was often preferred for its calming effect and lower cognitive demand, with melody emerging as the key element for following and appreciating the song. As U3 described: *"I tend to gravitate towards non-lyrical, because then I don't have to stress myself so much about 'listening'... to listen to the lyrics carefully."*

This distinction highlights a challenge for avatar-based visualizations. While captions and lip-syncing can support lyrical music, non-lyrical music lacks an equivalent visual cue for melody. When the avatar's mouth remained static during non-lyrical segments, participants described the experience as *"underwhelming"* (U1) or disconnected from the song's structure. U1 further noted that "the different mouth positions can signal different things," explaining that even subtle movement would "make it easier to focus when the avatar's mouth moved." Others echoed that even simple mouth movements could provide valuable cues, as U4 emphasized: *"Slow, fast or whatever…it would be good cues if lips are moving."*

DR3: Convey Melody in Non-Lyrical Music Through Scat Singing. To address this gap, avatars should provide visual cues

[3] https://www.youtube.com/watch?v=ZD0fhCZFjjI
A song by hip-hop artist Andy Mineo. The music video integrates visual effects designed to help viewers "see" musical rhythms using instrument visualization.

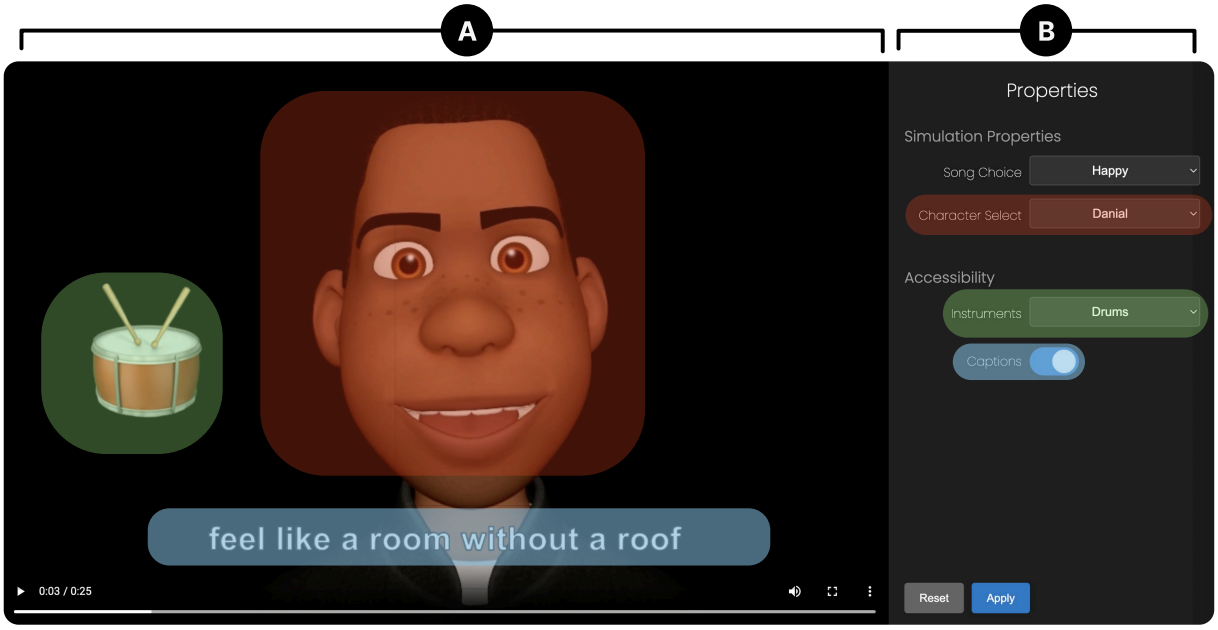

Figure 3: Study interface used with the FAME design probe. (A) Playback view showing the avatar music video with synchronized captions and instrument highlights. (B) Properties panel for selections: song, character, instrument highlighting, and a caption.

for melody during non-lyrical segments so that the music experience remains expressive and engaging. Avatars should scat sing (i.e., vocalize the tune using wordless syllables) to convey melodic contours and maintain continuity in musical expressiveness even when no lyrics are present.

4 Study 2 (Exploratory)

As the second step in our iterative, probe-based approach, we developed **FAME** (Facial Avatar for Musical Expression), an avatar-based music visualization that supports the design requirements identified in Study 1 (See Figure 3). We then used FAME as a refined probe in a two-phase study. In phase 1 (Comparison), participants interacted with both FAME and a visualizer-based probe (ViTune [35]) while music was playing, allowing them to compare whether avatars offer benefits over existing visualizer approaches. In phase 2 (Application), participants used FAME with 4 different types of music, enabling them to experience avatar-based visualization applied to diverse genres and reflect on its effectiveness. This mixed-methods design gave participants first-hand experience with FAME and provided us with grounded feedback, allowing us to validate and refine the preliminary requirements from Study 1 while also identifying challenges and opportunities for future design.

4.1 Second Probe: FAME

Guided by insights from Study 1, FAME embodies three design requirements: integrating captions and instrument cues with avatar performance (DR1), emphasizing expressive clarity over photorealism (DR2), and prioritizing melody in non-lyrical music (DR3).

The avatar is presented as a bust-up cartoon character with exaggerated features, chosen to enhance emotional expressiveness and reduce the uncanny valley effect (See Figure 3A). Depending on lyrical presence, FAME adapts its vocal animation strategy. For lyrical music, the avatar lip-syncs to lyrics with phoneme- and stress-aligned articulation, reflecting participant feedback and DR3. For non-lyrical music, the main melody is extracted with Melodia [91] and converted into a synthetic scatting vocal track ("da") using ACE Studios [1], allowing the avatar to "sing" the melody and visually convey its contours. Lip-sync is generated using the VOCAL system [86], which accounts for pitch modulation, vowel shifts, and vibrato, and was configured with default settings (See Figure 4C).

Emotional tone is encoded through expressive facial actions such as eyebrow shape, mouth curvature, and eye openness, reflecting Ekman's universal emotions [40]. For example, sad segments elicit drooping eyes and furrowed brows, while joyful segments produce smiles and raised eyelids (See Figure 4A). These expressions are generated with JALI [39] at an exaggerated setting (180% of standard blend strength) and blended with the lip sync. The high emotional intensity is intended to highlight emotion more clearly than in the formative study, consistent with DR2 (See Figure 12).

Pitch is also rendered visually: higher notes widen the mouth and raise brows, lower notes narrow articulation, and vibrato appears as subtle oscillations of the jaw (See Figure 4B).

Rhythmic structure is conveyed through head nods and upper-body movements aligned to beats (See Figure 4D). We generate motion by controlling joint rotations of the head and chest, applying Laplacian smoothing and a slight delay to produce anatomically plausible propagation from head to torso. This choice prioritizes visibility of the head over full-body dance movements [101], making facial expression and lip sync easier to perceive.

Complementary musical cues provide additional context, following DR1. Instrumental layers, such as drums and piano, are

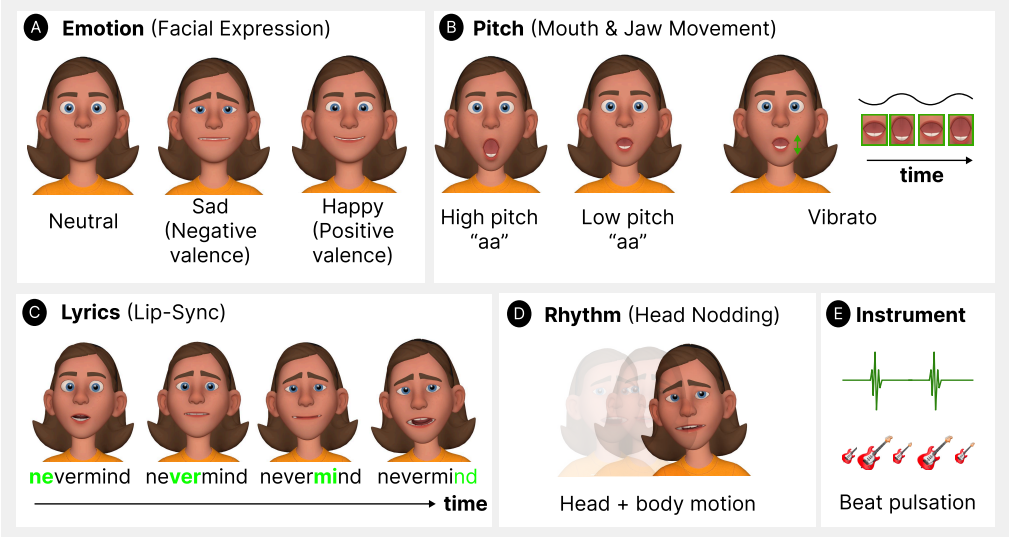

Figure 4: Visual mapping of musical elements to avatar animation components. (A) Emotional tone is represented through facial expressions such as brow movement and lip curvature. (B) Pitch is expressed through mouth openness and jaw vibration. (C) Lyrics are conveyed via lip-sync for lyrical songs and scat-singing for instrumental music. (D) Rhythmic patterns are encoded in head nods and upper-body motion aligned to beats. (E) Pulsating Instruments to showcase instrument-specific beats.

visualized with pulsing icons synced to beats (See Figure 4E). The pulse expands and contracts in proportion to the audio intensity of each instrument, reflecting the beats of the instrumental tracks.

In addition, toggleable real-time captions are aligned with lyrics and positioned below the avatar's face for readability without distracting from its performance. For non-lyrical music, captions were omitted, as they would have only repeated the scat syllable ("da").

4.2 Participants

We recruited 12 DHH participants through various venues, including social media platforms such as Facebook groups and Reddit channels, email list, word of mouth, and snowball sampling. Six participants self-identified as deaf, and the other half of the participants identified themselves as hard of hearing. Our inclusion criteria required participants to (1) identify oneself as deaf or hard of hearing, (2) have used visualization tools to experience music, and (3) understand musical concepts. Table 3 summarizes participants' demographic information. Ages ranged from 24 to 38 years, with a gender breakdown of 5 male, 5 female, and 2 non-binary. Four participants reported prior exposure to singing avatars through platforms such as virtual concerts, TV shows, and social media.

4.3 Study Procedure

Each study session lasted approximately 60–75 minutes and included four sequential parts: a pre-study survey, a comparison phase, an application phase, and a post-study interview and survey. All sessions were conducted individually over Zoom with video and audio recording, including auto transcription for accessibility and data analysis. Depending on their preference, the session was conducted using spoken language with live captioning, Zoom chat, or with an ASL interpreter. The study protocol was approved by the university's Institutional Review Board (IRB). We offered participants 35 USD for their participation.

4.3.1 Pre-study survey. Before the main session, participants completed a screening survey that collected their demographics, deafness history, communication preferences, and musical experiences. This ensured eligibility and helped personalize aspects of the study.

4.3.2 Comparison Phase. The first phase was designed to help participants compare FAME with an existing visualizer-based approach, ViTune[4] [35]. To structure this comparison, we selected four songs that varied in both valence (positive/negative) and lyrical presence (lyrical/non-lyrical) (See Table 7). For each song, we generated short video clips with either FAME or ViTune. Each clip was paired with the audio in three ways: (1) a correct match, (2) a mismatch with a different segment of the same song, and (3) a mismatch with a different song altogether. Participants were asked to choose the video that best matched the audio.

This setup was not intended as a performance test, but rather as a way of prompting participants to reflect on the benefits and limitations of each visualization. By encountering both correct and mismatched pairings, participants could more readily perceive how well each system conveyed timing, lyrics, and melody, and where alignment broke down.

4.3.3 Application Phase. In the second phase, participants experienced FAME applied to four different types of music, again spanning lyrical and non-lyrical as well as positive and negative valence (See Figure 3 for the probe interface). This phase broadened their exposure to avatar-based visualization and enabled them to reflect on

[4]We chose ViTune as a baseline since it has been evaluated with DHH participants in prior work [35]. Similarly, *Music-to-Facial Expressions* also compared its avatar-based visualization against ViTune [106].

Table 3: Main study participant information (N = 12). Values represent total counts unless otherwise noted.

Description	Details
Group	Deaf (6), Hard of Hearing (6)
d/Deaf Onset	Born d/Deaf (9), Late-deafened (3)
Age	Mean = 32.82, Median = 33, Min = 24, Max = 38
Gender	Female (5), Male (5), Non-binary (2)
Music Visualization Usage	Daily (9), Rarely (3)
Assistive Technology	Hearing aids (5), Cochlear implants (2), Both (2), None (3)
Communication	Sign and talk (6), Talk only (6)

how well the system supported different musical contexts. After viewing each song with FAME, participants shared feedback on its effectiveness, challenges, and potential points for improvement.

4.3.4 Post-study debriefing. Participants completed a 15-minute semi-structured interview followed by a 5-minute Likert-scale survey. The interview explored their perceptions of the FAME probe, its usability, and suggestions for improvement. The final survey collected quantitative ratings on musical element representation and FAME features.

4.4 Data Analysis

4.4.1 Comparison Phase Data. We analyzed responses from the comparison phase. Key metrics included (1) accuracy of user selections (i.e., frequency of incorrect answers), (2) response time (measured from the start of the page load to user selection), and (3) Likert scale ratings. Quantitative data were analyzed using Python's `pandas` and `scipy` libraries.

4.4.2 Application Phase Data. We examined interaction logs collected during the application phase. For each song, the probe recorded detailed usage data, including timestamped events for each configuration change (i.e., selecting characters, toggling captions, and modifying instrument highlights), and the final settings selected when the participant clicked "Apply."

4.4.3 Interviews and Survey Data. To understand participants' perspectives and expectations of FAME, we analyzed data from post-study debriefing interviews, surveys, and Zoom recordings (video, audio, and chat). We conducted a thematic analysis [15, 31], where three researchers independently coded initial transcripts, followed by 2 inter-rater discussions to resolve discrepancies and refine the codebook. We applied this codebook across data types, and triangulated patterns across sources. Recurring codes were then organized into overarching themes in comparison and application phases.

5 Study 2 Results

In this section, we present how participants compared FAME's avatar-based approach to visualizer baseline (ViTune) [35] and the ways they engaged with FAME's individual features. We also describe how participants felt these features conveyed musical qualities, such as rhythm, lyrics, and emotion, and conclude with their reflections on avatar roles, potential use cases, and opportunities for enhanced expressiveness and customization.

5.1 Comparison Phase Results

This phase was designed to give participants opportunities to compare FAME with ViTune and reflect on the benefits and challenges of each approach. Below, we report outcomes from the matching task and participants' subjective impressions. These findings highlight where avatars and visualizers provided advantages, and what tradeoffs participants perceived.

5.1.1 Alignment with Music. Across all 48 total instances where participants were shown a visualization (12 participants × 4 conditions), they correctly identified the matching video in 39 cases (81.25%). Accuracy was notably higher for FAME, with participants making the correct selection in nearly all cases (95.8%, with only one error), compared to ViTune, where the correct selection was made two-thirds of the time (66.7%, with eight errors). A paired samples t-test confirmed that participants made significantly more correct identifications with FAME (M = 1.92, SD = 0.29) than with ViTune (M = 1.42, SD = 0.67), $t(11) = 2.57$, $p = .026$.

Errors with ViTune occurred most often for lyrical songs such as Happy and Someone Like You, where abstract visuals made it harder to follow the lyrics. Five participants misidentified the Happy video, and two misidentified Someone Like You. Of these, five errors came from selecting a different part of the same song, while three involved choosing a completely different song. By contrast, the

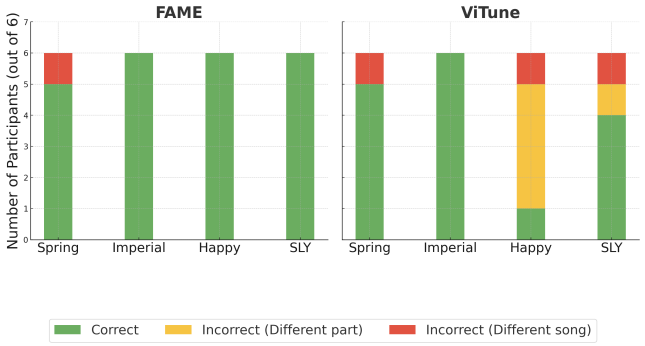

Figure 5: Participant responses in the comparison phase of Study 2 across four songs—Spring, Imperial, Happy, and Someone Like You (SLY)—for two probes: FAME (left) and ViTune (right). Bars represent how many participants (out of 6) selected each version as best matching the music: green for correct video–audio pairings, yellow for incorrect part from the same song, and red for incorrect song.

only FAME error was selecting a video from a different song. These patterns suggest that participants generally found the avatar-based cues in FAME more intuitive for aligning with music, especially when lyrics were present.

5.1.2 Response Time. We also measured how long participants took to select the video that best matched the music audio, as a rough indicator of how easily the visualizations supported alignment. On average, participants took slightly less time with FAME ($M = 138.69s$, $SD = 62.30$) compared to ViTune ($M = 142.52s$, $SD = 55.49$), though this difference was not statistically significant.

Looking more closely at lyrical and non-lyrical songs, different trends emerged. For lyrical music, participants were quicker with FAME ($M = 118.41s$, $SD = 76.2$), than with ViTune ($M = 146.35s$, $SD = 70.5$), with a difference of 27.94 seconds. For Non-Lyrical Music, the patterns reversed: ViTune led to somewhat faster responses ($M = 138.70s$, $SD = 76.4$) compared to FAME ($M = 158.98s$, $SD = 74.3$), with a difference of 20.28 seconds.

While neither of these contrasts reached significance, they suggest a tradeoff between speed and accuracy occurred when non-lyrical music was involved. ViTune's abstract visualizations sometimes enabled faster judgments, but these judgments also led to more errors. By contrast, FAME may have required slightly more time in some cases, but generally resulted in more accurate alignment, particularly when lyrics were present.

5.1.3 Perceived Usefulness and Enjoyment. Participants also reflected on how each system helped them understand and enjoy the music. When comparing the two approaches in terms of supporting understanding, more participants preferred FAME ($n = 6$) over ViTune ($n = 3$), with three considering them equally helpful. However, in terms of enjoyment, an equal number of participants preferred FAME ($N = 5$) to those who preferred ViTune ($N = 5$). These preferences illustrate how the two systems offered different strengths: FAME was valued for making lyrics and emotions clearer, while ViTune was appreciated for its ease and efficiency.

Participants described the effort required to lipread with FAME as a challenge when songs were unfamiliar. For example, P6 shared that *"lip reading is hard. It's not straightforward, so it's a lot of work looking at every part of the face to make it work"*. However, P9 found that lipreading can add to enjoyment if they already know the lyrics and captions were available: *"if you know the song already, yeah, and if you have the caption, it's even better to be reading the lips."*

Participants adopted different strategies with each system. When using FAME, participants primarily relied on the avatar's facial expressions and lip-sync accuracy to interpret the emotional tone and lyrical content of the music, often describing the avatar as conveying the overall "vibe" of a performance. In contrast, with ViTune, they focused on color changes and geometric shapes to perceive rhythmic patterns and pitch variations, which some found clearer for structural elements of the music.

Overall, participants saw FAME as more emotionally engaging and performative, while ViTune offered clearer representations of instrumental structure. These complementary qualities highlight the tradeoffs between avatar-based and abstract visualizations.

5.2 Application Phase Results

In this phase, participants explored how FAME's features (avatar, instrument highlights, and captions) shaped their engagement with music (See Figure 6). They also evaluated how well the system conveyed musical elements, suggested design improvements, and imagined potential roles and contexts for avatar-based music experiences. Across four songs, participants tried 176 unique feature configurations ($M = 44.0$, $SD = 6.52$) (See Table 4), covering different combinations of avatars, instrument highlights, and captions.

Table 4: Summary of feature selection per song. Values represent total selections across all participants.

Song	Feature Choices
Super Mario Bros. ($N = 37$)	Character: boy (22), girl (15) Instrument: All (Mixer) (25), None (12)
Moonlight Sonata ($N = 38$)	Character: boy (20), girl (18) Instrument: All (Piano) (20), None (12)
Die With a Smile ($N = 53$)	Character: boy (30), girl (23) Instrument: All (30), None (10), Drums (9), Piano (4) Captions: on (38), off (15)
Uptown Funk ($N = 48$)	Character: boy (32), girl (16) Instrument: All (36), None (5), Drums (2), Strings (3), Horns (2) Captions: on (44), off (4)

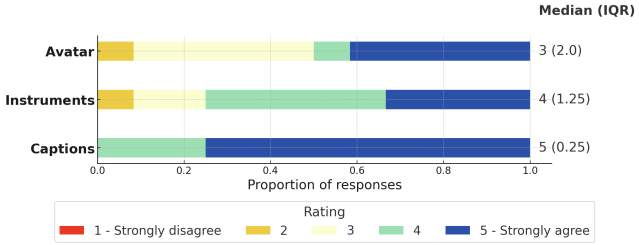

Figure 6: Agreement ratings (N = 12) on the perceived usefulness of individual FAME features, using a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). Rated features include avatar, instrument highlighting and captioning.

5.2.1 User Feedback on FAME's Interface Features.

Avatar ($M = 3$, $IQR = 2$). Reactions to the avatar were mixed. Many participants ($N = 5$) generally valued the avatar's role in communicating lyrics. For example, P9 shared: *"When you look at the avatar, you could kind of follow along… especially when the singing mumbles"*. Some participants ($N = 3$) also found that the avatar enhances the emotional delivery of the song. P1 remarked that *"[the avatar] helps me a lot to understand the mood behind the music, because they have facial expression."*

At the same time, several participants ($N = 3$) voiced concerns about the avatar's uncanniness and mismatches between what was seen and heard. P5 noted the lack of natural eye movements,

describing the avatar as having *"lifeless eyes… staring at me, with a song like this where it's supposed to be super emotional, right? I don't think Adele's ever sung this song with her eyes open."*

Generally, participants found value in the mouth movements, noting that it helps them to understand the lyrics and stay engaged with the non-lyrical music. For example, P9 viewed FAME's scatting as "kind of like a miracle thing" for non-lyrical music, explaining that "with the lyrical music, I can rely on the caption, but with non-lyrical, you can't really hear clearly. But we can focus on the mouth moving." Some critiqued the synthesized scatting lip-sync used for non-lyrical music because the repeated "da" syllable made it difficult to imagine the corresponding sounds and map the mouth shapes to the melody. For example, P5 noted, *"I don't like the mouth flaps… what I hear and what I see are not matching."* Similarly, P8 shared *"I found it very hard to match the lips"* to the music. This challenge was amplified because the same syllable was used throughout the whole song; as P12 noted, *"It would come to be a boring situation. People wouldn't get this really...It's very repetitive."*

Lastly, the avatar's appearance also influenced its perceived effectiveness. The *Daniel (male)* avatar was preferred across all four songs, especially for *Die With A Smile* ($N = 30$) and *Uptown Funk* ($N = 32$). Participants noted that Daniel's facial features made lip-reading easier. As P8 explained, *"I liked the male avatar better just because the mouth was bigger, so you could read the lips better."*

Instrument ($M = 4$, $IQR = 1.25$). Instrument highlights helped participants perceive the musical structure, adding a layer of richness to the visualization. For example, P7 noted, *"I like that you can see the instrument popping, you can see the layer of the music."* High-energy songs, such as *Uptown Funk* ($N = 36$) and *Die With A Smile* ($N = 30$), elicited frequent use of the "All instruments" option, while more contemplative pieces like *Moonlight Sonata* prompted simpler configurations (e.g., "Piano" or "None").

While participants generally felt instrument highlights are useful for understanding what instruments are in the song, participants felt that the way instrument highlights were shown could be improved. Some felt the highlights were visually overwhelming. For example, P5 explained, *"I know the pianos are playing, but I feel like it's a little bit distracting."* Participants also felt the instrument icons' movements should better reflect how each instrument is physically played. *"I think that's cool. I get told what the instruments are, right? But maybe they need different movements."* (P10)

Captions ($M = 5$, $IQR = 0.25$). Captions were consistently described as essential, particularly for lyrical songs such as Die With A Smile ($N = 48$) and Uptown Funk ($N = 44$). Participants emphasized how captions complemented the avatar. For example, P1 said, *"...with the lyrics and the facial expressions, I could better understand the overall mood and rhythm. When the captions matched the avatar's mouth movements, it doubled the effect, helping me follow the song."*

At the same time, participants suggested design improvements, such as moving captions lower to avoid covering facial expressions and adopting karaoke-style *"word by word highlighting"* (P9). Thus, while captions were seen as essential, participants discussed the need for more dynamic and integrated presentation.

5.2.2 Musical Elements Representation. Participants expressed interest in using FAME in the future and moderate trust in its accuracy for representing musical elements, with an average rating of 3.58 out of 5 ($SD = 0.69$) (See Figure 7). FAME was reported to be particularly effective in conveying the lyrics and emotion, somewhat effective for rhythm and pitch, and least effective for melody.

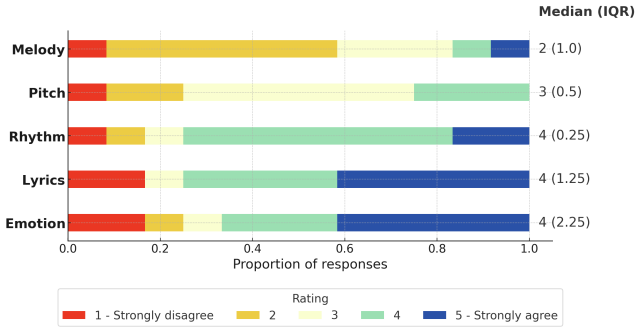

Figure 7: Agreement ratings (N = 12) on FAME's effectiveness in visually conveying musical elements, using a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). Statements addressed how well FAME expressed rhythm, lyrics, emotion, and pitch through visual animation.

- **Lyrics** ($M = 4$, $IQR = 1.25$): Understanding lyrics received the highest rating, suggesting that the avatar's lip-sync was especially helpful in communicating lyrics. As P8 shared, *"The avatar is great because you can read their lips."*
- **Emotion** ($M = 4$, $IQR = 2.25$): Participants reported that the avatar's facial expressions effectively communicated the mood of the music. They frequently referenced the eyebrows and lips as cues for interpreting emotion. As P1 shared, *"If I play the song 'Happy,' then their face is smiling. I can tell."*
- **Rhythm** ($M = 4$, $IQR = 0.25$): Participants found the avatar helpful in conveying rhythm through lip and head movements. For instance, P2 mentioned, *"I can catch the rhythm by watching their mouth."* P6 added, *"The avatar keeping the beat with their head movements helped when [I'm] trying to figure out the strong beat."* Others noted that scat aligned well with instrumental timing, with P5 observing that the avatar's mouthing matched the moment of a violin pluck.
- **Pitch** ($M = 3$, $IQR = 0.5$): Perceptions of pitch received moderate ratings. Some participants identified subtle visual cues, such as vibrato. For instance, P5 noted, *"There's some strain to it, certain ways the lips tremble, like you can feel it."* However, others, especially Deaf participants, found pitch harder to distinguish. As P7 shared, *"I'm not sure what pitch really means. I didn't feel a big difference."*
- **Melody** ($M = 2$, $IQR = 1.0$): Melody received the lowest average rating among the musical elements. While participants were often able to align the avatar's scat singing with the melody, some found the use of nonsensical syllables unnatural or unfamiliar. As P9 explained, *"I don't really like scats because I can't understand what kind of sound that's trying to make… unless you know the music."* This suggests that the challenge was not melody recognition itself, but

the perceived unnaturalness of scat-style vocalization as a representational strategy.

5.2.3 Desired Improvements. The design probe also highlighted additional areas where participants envisioned avatars becoming more expressive and immersive. Suggestions centered on expanding expressive channels, increasing diversity and customization, and combining avatar performance with other forms of visualization.

Hands and Body Movements. Participants repeatedly stressed the importance of incorporating full-body and hand movements, especially given the centrality of gesture in sign language and DHH culture. P6 explained: *"Full body, including the hands and lower body! Some people who use sign language are used to having bodily movements."* Others also discussed that having a body would make it easier to emotionally connect with the performance (P12) and make the viewing experience more enjoyable (P1). Moreover, participants suggested incorporating hands and body movement to shift the experience of using an avatar-based visualizer by feeling "less like a tool, but more like entertainment" (P6).

Avatar Appearance and Customization Participants expressed a strong desire for avatars to reflect diverse identities and, in some cases, to visually resemble the original artist. P6 emphasized that resemblance to the performer would enhance immersion *"because [he looks] just like the artist and makes you feel like you're at the concert yourself."* Similarly, P12 noted that mismatches between the avatar and the singer disrupted the experience: *"When I look at somebody, and it's not the representation of the person who sang the song, it just kind of doesn't fit."* Beyond resemblance, participants also wanted more diverse options in gender, ethnicity, and attire.

Combination with Other Visualizations Many participants ($N = 6$) noted the absence of background imagery in our design probe and suggested using the background as an additional visualization layer provide context and complement the avatar. For example, P1 commented, "I'll choose the background with the concert hall. For Uptown Funk, we can capture one of the scenes from the music video." Others highlighted a desire for pitch-related cues, such as dynamically changing background colors (P2, P9) or integrating a spectrum analyzer (P3) [81]. Overall, participants suggest holistic integration between avatar and background visualization. As P3 noted: "Some sort of combination of visualization bars plus the avatar would create a fuller experience."

5.2.4 Potential Roles of Avatars in Music Experiences. Participants described avatars not only as visual aids but also as aids for socially meaningful ways of experiencing music. Their reflections highlight three main roles: **performers, interpreters, and companions**.

Performer. Avatars were often framed as entertainers embodying the expressive energy of music. For example, P5 likened the avatar to Japan's Vocaloid phenomenon: *"So it's a Vocaloid, a Japanese invented pop star. They're avatars that sing. It kept me engaged enough to visualize her. So I'd say 'yes', it did help me enjoy the music."* Others, such as P1, compared avatars to orchestral conductors whose physicality conveys emotion and structure: *"It reminded me a lot of orchestra. The conductors, their facial expression is like, as soon as they hit the higher notes, they go up with their eyebrows."*

Interpreter. Some participants saw avatars as visual translators, paralleling the role of sign language interpreters at live events. P11 explained: *"ASL interpreters on stage… they hype up the crowd, show tiny beats, and adjust to genre (the Carpenter = romantic creamy, Doja Cat = sharp, hype)."* Similarly, P3 drew parallels to interpreters shown on-screen: *"It looked like the interpreters that you see in like a little bubble on the side of a screen."*

Companion. Some participants imagined avatars as friendly presences who co-experience music with them. Rather than just showing music, avatars could create a sense of shared engagement. P9 reflected: *"It would be a way for me to experience music by watching someone else."* Similarly, P12 commented: *"I can picture myself walking with them to the beat in that setting. This is an expression for that music. Those would be my thoughts I wouldn't have got otherwise."*

5.2.5 Contexts for Using Avatars in Music Experiences. Participants envisioned avatars being integrated into both individual and social music experiences. Overall, they expressed general interest in incorporating avatar-based visualization into everyday contexts.

Individual Use. Some participants saw value in using avatars for practicing singing and better understanding lyrics. For example, P9 remarked: *"I wouldn't mind generating avatar videos so that I can learn how to understand the lyrics. When you look at the avatar, you could kind of follow along."* Others wanted to personalize avatars to reflect their identity. P1 shared: *"If I customize the avatar to look like me, then I can create videos of myself singing and enjoying music."*

Social Use. Avatars were also imagined in shared contexts, such as concerts, karaoke, and parties. At concerts, they were seen as a way to enhance focus and reduce visual overwhelm. P7 commented that they would *"Use an avatar when you're in a concert, so that you can concentrate on one avatar rather than trying to follow all the different things happening on stage and screens."* In karaoke, their playful qualities were described by P6 as a good fit for the atmosphere: *"The avatars look a little bit goofy, and karaoke is a goofy time anyway."* At parties, avatars were imagined as making background music more accessible. P7 described: *"If you have an avatar right there on the screen, people at the party may better understand the sound. We sometimes have parties where music is just played in the background, but with FAME, people could watch the video instead."*

5.3 Summary of Findings

Our findings reveal both the strengths and limitations of avatar-based visualization, as well as its potential applications. FAME significantly improved musical comprehension compared to a visualizer baseline (95.8% vs. 66.7% accuracy, $p = .026$), particularly for lyrical music. Participants highlighted the avatar's effectiveness in conveying lyrics, emotion, and rhythm, while noting ViTune offered clear presentations for pitch and melody. Captions were consistently valued as essential for lyrical music, with requests for enhancements such as karaoke-style highlighting.

Participants emphasized areas for improvement, including clearer support for non-lyrical music, expanded expressive modalities (e.g., dancing and signing), and greater control over avatar appearance and backgrounds. Beyond these technical considerations, participants envisioned avatars as performers, interpreters, and companions, positioning them as cultural and social participants in music experiences. They further imagined avatars playing

roles in both individual engagement (e.g., practicing singing, personalization) and social settings (e.g., concerts, karaoke, parties).

6 Discussion: Design Implications for Expressive Music Avatars for DHH individuals

Through two probe studies, we explored the potential of avatars to support music engagement for DHH audiences by conveying lyric timing, rhythm, and emotional tone. Participants especially valued the performative qualities of the avatar—its facial expressions, lip-sync, and rhythmic motion—which helped them connect emotionally and imagine the music being embodied. These strengths highlight avatars not only as accessibility aids but also as expressive companions that add enjoyment and depth to music appreciation. At the same time, the probe exposed several limitations and tradeoffs: lip-sync was useful for lyrics but struggled in non-lyrical passages, facial-focused animation felt incomplete without hands or body movement, and realism heightened uncanny effects.

To situate these findings, it is important to contextualize FAME alongside existing Deaf cultural practices such as song signing. Song signing offers rich emotional and rhythmic artistic translation [73] through skilled human performance, but is resource-intensive and difficult to produce at scale. FAME explores how expressive avatars might algorithmically render some aspects of musical performance in visually meaningful ways. These approaches are not substitutes for song signing, but rather early design explorations into how automated approaches might complement existing practices by generating accessible information more efficiently and at scale. In doing so, our findings highlight both the potential and the cultural tensions of translating musical expression into avatar form.

We note that perceptions of these strengths and limitations varied widely across the spectrum of how DHH audiences experience music [45, 118]. This diversity underscores the need for selective visualization, where users can decide which layers of information to prioritize based on their preferences. Building on this framing, the following sections present five design implications: (1) enhancing scat singing for non-lyrical music, (2) integrating hand movements and sign language, (3) extending to full-body performance and dance, (4) balancing realism and stylization in avatar appearance, and (5) layering avatars with captions and visual backgrounds. We then conclude with a refined list of design requirements.

6.1 Refining Lip-Sync and Synthesizing Scatting

Participants recognize lip-sync as a distinctive strength of avatar-based visualization: it lets DHH users follow lyrics while keeping their gaze on the face to pick up emotional cues. Similarly, scat singing gave the avatar a way to remain active in non-lyrical segments, allowing participants to "see" the melody.

Despite these strengths, participants found limitations in both features. Lip-sync felt flat when not paired with expressive eyes or brows, and several described the animation as "uncanny". Moreover, participants emphasized that lip-sync was most helpful when paired with captions or when the lyrics were already familiar, underscoring the importance of contextual anchors. For scat, the repeated use of "da" syllables was perceived as monotonous and made it difficult to understand the instrumental contours. This echoes Nanayakkara *et al.*'s findings that beat-synchronized "ba" mouth shapes of a human performer supported timing perception for DHH listeners but offered only coarse cues about expressive contour [4].

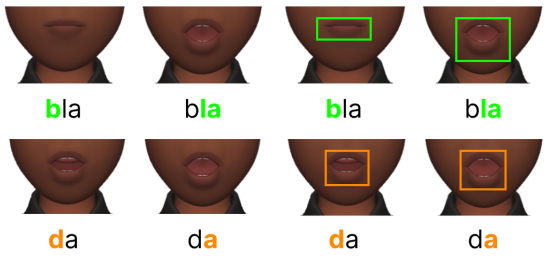

Figure 8: Comparison of mouth shapes produced by different scat syllables (Top: "bla" v.s. Bottom: "da". Bottom: "da"). Colored boxes highlight the areas of the mouth size.

To address these gaps, future systems could augment lip-sync with exertion cues, such as jaw trembling during vibrato, brief eye closure on high notes, or visible cheek tension to convey vocal effort [3, 86]. For scat, instead of a single syllable, designers could draw from jazz scat literature [9, 92, 113] to employ a set of visually distinct syllables (e.g., bla) chosen according to local melodic context [10], with articulation modulated to reflect phrasing and contour (See Figure 8). Technical advances in expressive speech animation [22] and syllable-conditioned gesture generation [7] could support this by varying articulation in sync with rhythm and timbre. Anchoring mechanisms, such as adding synthetic scat audio or displaying syllable-aligned captions or melodic traces, could further clarify the connection between what is seen and what is heard.

Enhancing lip-sync and scat introduces tensions between legibility and naturalness. Exaggerated articulation improves visual clarity but can appear artificial if not balanced with expressive motion elsewhere. Similarly, diversifying scat syllables prevents monotony but risks overloading viewers with too much variation. Designers must balance these extremes, providing just enough exaggeration and variety to support comprehension without distraction.

6.2 Integrating Hand Movements and Sign Language

While participants appreciated that FAME's avatar could already convey lyrics and emotion through facial expressions and lip-sync, they pointed out that the absence of hands and body limited additional ways of meaningfully expressing important musical elements. Participants requested the addition of hand movements and sign language, framing the avatar as a potential interpreter. As P11 explained, *"ASL interpreters on stage… they hype up the crowd, show tiny beats, and adjust to genre."* This reflects how DHH individuals experience music through signed music (or "song signing"), a valued practice in Deaf culture that goes beyond translating lyrics into ASL to embody rhythm, spatiality, and emotion [109]. Similarly, in classical music, DHH participants reported stronger engagement when conductor gestures were shown in sync with the music [4].

Future systems could integrate generative sign language avatars [60, 112] or gloss-based translation methods [110] to render lyrics

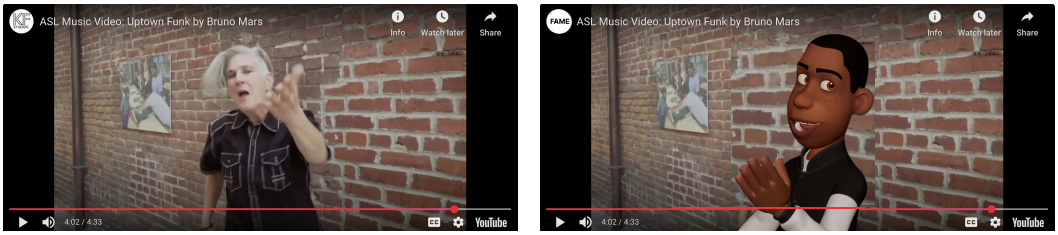

Figure 9: FAME avatar with hand gestures. (Left) Example sign of "up" used in song signing "Uptown Funk". (Right) Daniel's upper body, including "up" hand gestures. Song signing video source: https://youglish.com/pronounce/uptown/signlanguage/asl.

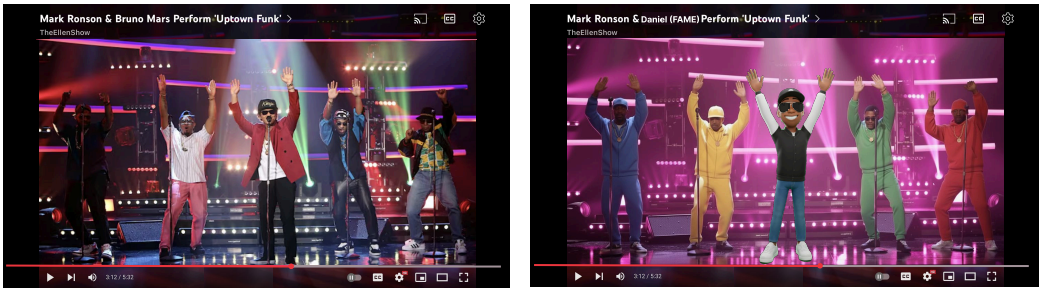

Figure 10: FAME avatar with full body. (Left) Bruno Mars performing "Uptown Funk" on a talk show (https://www.youtube.com/watch?v=P-WdrMLLpPg). The image captures the energy of the live performance, with all dancers, including Bruno Mars, striking a synchronized pose with their hands raised. (Right) The FAME avatar mimics the exact pose from the performance.

in a structured and culturally meaningful way [27]. Tools such as Sign Dance Maker demonstrate how lyric-aligned signing can respect both rhythm and emotional nuance in musical performance [79]. Beyond full signing, advances in co-speech gesture generation [7, 22, 23] show how models can synthesize rhythmic and semantic arm and hand movements aligned with speech prosody, while offering fine-grained control over expressiveness. Applied in a musical context, these methods could enable avatars to clap, snap, or sway in time with the music, providing additional rhythmic and affective cues that enhance accessibility and engagement (See Figure 9).

Incorporating hand and arm movements introduces both opportunities and risks. While gestures and signs can greatly enrich the performance and increase acceptance within the DHH community, they also raise challenges of accuracy, cultural sensitivity, and visual complexity. These concerns echo critiques of signing avatars as "disability dongles" when developed without sustained Deaf involvement [4], underscoring the need for Deaf-led evaluation to ensure cultural and linguistic meaningfulness. Designers must decide when to prioritize faithful linguistic representation versus simplified gestural cues, and how to balance expressiveness with the risk of cognitive overload when multiple visual signals (face, hands, captions, instruments) compete for attention.

6.3 Extending to Full-Body Performance and Dance

Participants trusted FAME's accuracy in representing musical content and rated it highly for conveying core elements such as emotion and rhythm. These impressions aligned with our design focus on expressiveness and clarity through facial expressions and rhythmic head motion (DR1, DR2). As U4 observed, *"The head should be moving more! Like the way Bruno Mars did."* However, head nodding alone was seen as insufficient. Participants wanted to see richer upper-body animation and dancing to better capture musical rhythm and emotion. As P6 remarked, adding dance would make the avatar feel "less like a tool, but more like entertainment." This resonates with prior work showing that DHH individuals often rely on full-body movement to experience the energy and affective qualities of music [49, 99]. For example, DHH users showed stronger engagement and clearer perception of phrasing when viewing synchronized conductor gestures in classical music. [4]

Future systems could incorporate skeleton-based animation [88] and gesture-generation methods from dance recognition [69, 83] to enrich full-body movement in avatars. More recently, music-conditioned motion generation approaches using transformers and diffusion models [24, 67, 69] demonstrate how rhythm-aligned and stylistically varied dance can be synthesized directly from audio. Prior CHI work shows how avatars and embodied interaction support learning, self-expression, and social participation in dance [41, 115]. In our context, such techniques could allow avatars to improvise body movements during instrumental interludes, sustaining engagement throughout the song (See Figure 10).

However, emphasizing full-body performance introduces trade-offs. While dancing enhances emotional connection and enjoyment, it can also draw attention away from subtle but important cues like lip movements, facial expressions, or caption alignment. Designers

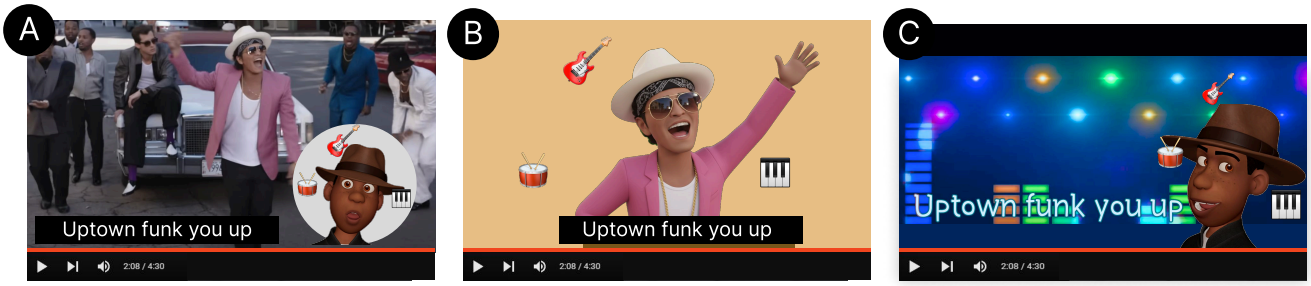

Figure 11: Three potential roles of FAME avatars in music experiences: (A) Interpreter: The avatar appears in the corner of a music video; (B) Performer: The avatar takes center stage; (C) Companion: The avatar joins a karaoke scene.

must therefore balance the richness of body motion with the clarity of other modalities to avoid overwhelming users.

6.4 Balancing Realism and Stylization in Avatar Appearance

Our probes featured photorealistic (Study 1) and stylized avatars (Study 2). The photorealistic avatar got praise for accurate lip sync (U3, U4, U9), but many found it uncanny and low on emotion, which matches prior reports of the uncanny valley [62, 82]. The stylized avatar used larger, more exaggerated facial features and helped people read the song's mood (P1, P6, P12) and read lips (P5, P7, P8), in line with work showing that clearer, amplified features improve emotion reading and articulation visibility [2, 114, 116].

At the same time, stylization introduced new challenges. Errors in lip sync appeared more noticeable, and during non-lyrical music, repeated "da" syllables without a clear source made the avatar feel puppet-like. Several participants also described the avatar's eyes as "lifeless," noting missing cues such as gaze shifts and natural blink timing. These limitations suggest that stylization enhances expressivity, but without grounding in realistic motion, it risks looking artificial or distracting.

Future work should keep in mind the expressive benefits of stylized geometry while grounding motion in realism. Concretely, use human-like gaze strategies [85], realistic blink timing, and eye-closure patterns that reflect vocal effort; anchor lip sync to the true source of sound production, including phonetic timing, emphasis, and visible micro-articulations; reveal musical phenomena such as vibrato with subtle jaw or lip micro-oscillation [86]. For non-lyrical music, pair any mouth motion with an auditory or visual anchor so the source is clear, as outlined in Section 6.1. When the avatar plays the role of the performer, we also recommend aligning the avatar's identity with the perceived gender and race of the singer to improve recognition and acceptance (see Figure 11).

Future work could also consider tuning avatar proportions to match communicative goals. When emotion legibility is the priority, enlarging the eyes and brows can enhance expressivity [105]. When supporting lip reading, enlarging the lips offers clearer articulation cues [2]. In contrast, avatars with realistic proportions are generally preferred in immersive contexts [36]. Regardless of the proportions, these adjustments should always be paired with motion constraints to avoid the failure mode in which amplified features also amplify animation errors.

6.5 Layering Avatars with Captions and Visual Backgrounds

Our refined probe shows that layered avatar, captions, and instrument highlights were generally useful, with participants using the different layers to understand different elements of music. The avatar helped people follow mood, lyrics, and rhythm, and captions reinforced this effect when synced with the mouth: as P1 explained, "*When the captions matched the avatar's mouth movements, it doubled the effect, helping me follow the song.*" Instrument highlights provide structure by showing which parts were playing. Beyond these components, participants also requested a background layer to add musical context (e.g., a concert hall scene) or pitch cues like color and spectrum visualizers.

However, some participants also voice concerns that too much motion in the visualization (such as the instruments) is a bit distracting (P5), and for the mellow songs like the *Moonlight Sonata*, many users show a preference for the minimal set of instrumental highlights ($N = 12$). While more layers of information can be helpful, it can also push cognitive load [68] and cause post-task fatigue for DHH users [90]. Studies have also shown that while more visual information is more engaging to DHH users, it does not always translate to understanding [43]. Further, lip-reading is a challenging task and requires full attention; splitting the user's attention with additional visualization may reduce the communicative effectiveness of lip reading [63, 75].

When designing avatar-based systems, it is important to align visualization choices with the communicative goal. Our results point to the synergistic effect of avatars and static captions for lyric comprehension, with users suggesting avatars can be used as a companion during karaoke (see Figure 11C). We point to future work to study different combinations of avatar and additional visualizer. Such as the effect on lyric comprehension of showing timed-dynamic captions with lip-sync animation [46, 72, 80], understanding pitch comprehension by combining avatar motion with pitch visualizer [35, 53, 57], and how background influences enjoyment and presence [55, 68, 111]. Given the diversity among DHH users, practical visualizers must be personalized to accommodate individual needs [76].

6.6 Refined and Extended Design Requirements

Our evaluation largely supports the formative design requirements (DRs) derived from Study 1, while also suggesting refinements and extensions. Below, we revisit each DR and summarize the new requirements that emerged.

DR1 (*Combine captions and instrument cues with avatar performance*) is supported, but requires refinement. Captions were consistently valued, receiving a Median usefulness rating of 5. Participants emphasized that captions as essential for lyric comprehension across unfamiliar and familiar songs. However, participants requested lower-placed captions so the text would not obstruct facial expression. Instrument highlights were also widely used, especially in high-energy songs such as Uptown Funk and Die with a Smile, helping participants to see musical structure. At the same time, participants felt highlights can be visually overwhelming or distracting when instrument icons moved too aggressively.

- **Refined DR1: Support flexible layering of captions and visualizations.** Captions should be positioned to avoid occluding facial expressions and ideally adopt dynamic or karaoke-style timing, which reinforces lip-sync and improves lyric comprehension. Instrument highlights should remain simplified and precise, providing structural cues without excessive motion. Additional layers, such as contextual backgrounds, can enrich musical interpretation, but must be balanced to prevent cognitive overload.

DR2 (*Prioritize emotionally expressive facial and bodily movements*) is strongly supported. Participants relied on facial expressions and rhythmic head motion to interpret mood and beat. Participants noted that stylized features improved lip readability. However, they critiqued incomplete or unnatural facial behaviour, such as lifeless eyes, minimal blinking, and mismatched articulation in some segments. They also reported that facial-focused animation felt incomplete without body motion when that current expressiveness *"stops at the neck."* Participants wanted fuller body engagement, particularly during instrumental sections, noting that head nodding alone was insufficient for conveying musical energy, while requesting hand and arm movement to complete the expressive performance. They also described dance movement as helpful for emotional engagement and for matching the vibe of upbeat songs.

- **Refined DR2: Balance stylization and realism in avatar appearance.** Stylized geometry (e.g., enlarged eyes, brows, or lips) can enhance expressivity, while realistic gaze, blinks, and articulation patterns ground motion in naturalness. Tuning appearance to communicative goals (e.g., lip-reading vs. immersion) can increase both clarity and acceptance.
- **New DR3: Extend to upper-body and** full-body performance. Avatars should use coordinated body and arm movement and dance to capture rhythm, mood, and energy, particularly during instrumental sections. These motions can enhance emotional engagement while being balanced so as not to obscure facial cues or captions.

DR3 (*Convey melody in non-lyrical music through scatting*) is supported. Participants felt that scatting successfully kept the avatar visually active and synced during non-lyrical songs. However, some participants disliked the repetitive use of the syllable "da," which they felt was monotonous and made it hard to understand the instrumental contour.

- **Refined DR4: Convey melody in non-lyrical music through varied and anchored scatting.** Scat syllables should be visually distinct and temporally aligned with melodic contour, with articulation varied to reflect phrasing and emphasis. Anchoring mechanisms, such as synthetic scat audio, syllable-aligned captions, or simple melodic traces, can further strengthen the connection between visual and audio.

7 Limitations

A first limitation is that our study primarily involved DHH participants who were already actively engaged with music, often using captions to support their appreciation. While this aligns with prior work on DHH music accessibility [109, 118] showing that captions are the most frequently used visualization tools, it may not capture the full spectrum of Deaf community perspectives, including those less musically engaged or who experience music primarily through tactile channels. As such, these findings should be interpreted as an early design exploration rather than a generalizable evaluation. Future research should include participants with a broader range of musical exposure and communication modalities to better understand how avatar-based visualizations are received.

Second, our study primarily examined songs with relatively high valence and arousal, where strong emotions and pronounced rhythms made avatar-based visualizations more salient. While this provided a useful testbed for expressive features, it limits our understanding of how avatars might function with calmer, lower-arousal, or ambient genres. This limitation also extends to our use of scat-style vocalization for melody visualization, which originates in Black Jazz traditions [10, 113] and may not generalize well to genres with different stylistic conventions. But other vocalizations, such as solfège and ragga, might be applicable. Future work should broaden the scope to include styles such as lo-fi [38], meditation soundtracks [78], to assess how genre-specific properties and the suitability of scat-based melody cues shape the effectiveness.

Third, all evaluation sessions were conducted online via Zoom, where participants used either a laptop or a desktop with an external monitor in a quiet, interruption-free environment. Although this ensured consistency, it does not capture how FAME might operate in socially dynamic contexts, such as concerts or crowded public spaces, where maintaining sightlines and visual orientation might be more difficult for DHH users. Future work should investigate FAME's usability and effectiveness in such real-world settings.

Finally, our web-based probe relied on a multi-step pipeline involving audio separation, preprocessing, and manual adjustments, which constrained scalability and availability. This design is consistent with a probe-based approach, where feasibility and exploration take precedence over full automation, and it demonstrated the feasibility of avatar-based visualization. Future work should streamline this workflow by automating preprocessing and enabling real-time operation of the system. Such automation would also make it feasible to support long-term, in-the-wild deployments.

8 Conclusion

This work explores how expressive facial avatars can support music accessibility for d/Deaf and Hard of Hearing (DHH) individuals through an iterative, probe-based research approach. Across a formative study and a two-phase exploratory study, we investigated how participants engaged with lyrical and non-lyrical music through avatars, in order to understand how to design avatar-based systems that better support music accessibility. By treating our studies as evolving design probes rather than finished solutions, this work revealed opportunities, challenges, and directions for making music more inclusive. Our findings highlight the importance of emotional expressiveness, multimodal visual cues, and flexible layering of representations in shaping accessible music experiences. We contribute a refined set of design requirements grounded in DHH users' practices and needs by capturing participants' visions of how avatars can serve as effective interpreters, performers, and companions for enhancing rhythm, emotion, and lyric understanding.

Acknowledgments

This work was supported by an NSERC grant RGPIN-2021-04268.

References

[1] ACE Studio. 2025. AI Singing Voice Generator. https://acestudio.ai/ Accessed 2025-04-03.
[2] Simon Alexanderson and Jonas Beskow. 2014. Animated Lombard speech: motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech & Language* 28, 2 (2014), 607–618.
[3] Deepali Aneja, Daniel McDuff, and Shital Shah. 2019. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International conference on multimodal interaction*. 69–73.
[4] Robin Angelini. 2023. Contrasting technologists' and activists' positions on signing avatars. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.
[5] Taravat Anvari, Kyoungju Park, and Ganghyun Kim. 2023. Upper body pose estimation using deep learning for a virtual reality avatar. *Applied Sciences* 13, 4 (2023), 2460.
[6] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–19.
[7] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.
[8] Bastien Arcelin and Nicolas Chaverou. 2024. Audio2Rig: Artist-oriented deep learning tool for facial and lip sync animation. In *ACM SIGGRAPH 2024 Talks*. ACM, 1–2.
[9] William Bauer. 2007. Louis Armstrong's Skid Dat De Dat: Timbral Organization in an Early Scat Solo. *Jazz Perspectives* 1, 2 (2007), 133–165.
[10] William R Bauer. 2002. Scat singing: a timbral and phonemic analysis. *Columbia University* (2002).
[11] H-Dirksen L Bauman and Joseph J Murray. 2014. *Deaf gain: Raising the stakes for human diversity*. U of Minnesota Press.
[12] Cynthia L Bennett, Erin Brady, and Stacy M Branham. 2018. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th international acm sigaccess conference on computers and accessibility*. 161–173.
[13] Lynne E Bernstein, Nicole Jordan, Edward T Auer, and Silvio P Eberhardt. 2022. Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training. *American Journal of Audiology* 31, 2 (2022), 453–469.
[14] Alexandre Berthault, Takuma Kato, and Akihiko Shirai. 2023. Avatar Fusion Karaoke: Research and development on multi-user music play VR experience in the metaverse. In *2023 IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom)*. IEEE, 281–289.
[15] Virginia Braun and Victoria Clarke. 2024. Thematic analysis. In *Encyclopedia of quality of life and well-being research*. Springer, 7187–7193.
[16] Sylvain Brétéché. 2019. Visual music? The Deaf experience. In *14th International Symposium on Computer Music Multidisciplinary Research*.
[17] Sylvain Brétéché and Christine Esclapez. 2018. Music (s), Musicology and Science: Towards an Interscience Network: The Example of the Deaf Musical Experience. In *Music Technology with Swing: 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers 13*. Springer, 637–657.
[18] João Couceiro e Castro, Pedro Martins, Ana Boavida, and Penousal Machado. 2019. Máquina de Ouver-from sound to type: finding the visual representation of speech by mapping sound features to typographic variables. In *Proceedings of the 9th International Conference on Digital and Interactive Arts*. 1–8.
[19] Doga Cavdir. 2024. Development of embodied listening studies with multimodal and wearable haptic interfaces for hearing accessibility in music. *Frontiers in Computer Science* 5 (2024), 1162758.
[20] Gizem Çelik. 2023. A new field in music production: metaverse concerts. *Ege Üniversitesi İletişim Fakültesi Medya ve İletişim Araştırmaları Hakemli E-Dergisi* 12 (2023), 4–24.
[21] Vinay Chamola, Gaurang Bansal, Tridib Kumar Das, Vikas Hassija, Siva Sai, Jiacheng Wang, Sherali Zeadally, Amir Hussain, Fei Richard Yu, Mohsen Guizani, et al. 2024. Beyond reality: The pivotal role of generative ai in the metaverse. *IEEE Internet of Things Magazine* 7, 4 (2024), 126–135.
[22] Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. 2025. Motion-example-controlled Co-speech Gesture Generation Leveraging Large Language Models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.
[23] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7352–7361.
[24] Zeyuan Chen, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xin Chen, Chao Wang, Di Chang, and Linjie Luo. 2025. X-dancer: Expressive music to human dance video generation. *arXiv preprint arXiv:2502.17414* (2025).
[25] JaeHyeok Choi, Jonggwun Chong, Woojin Lee, and WonHyong Lee. 2021. VR Karaoke Using Expressive 3D Avatars. In *International Conference on Robot Intelligence Technology and Applications*. Springer, 543–552.
[26] Youjin Choi, Junryeol Jeon, ChungHa Lee, Yeo-Gyeong Noh, and Jin-Hyuk Hong. 2024. A Way for Deaf and Hard of Hearing People to Enjoy Music by Exploring and Customizing Cross-modal Music Concepts. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[27] YouJin Choi, JaeYoung Moon, Kyung-Joong Kim, and Jin-Hyuk Hong. 2024. Exploring the Potential of Generative AI in Song-Signing. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 816–820.
[28] Zubin Choudhary, Gerd Bruder, and Greg Welch. 2023. Visual Hearing Aids: Artificial Visual Speech Stimuli for Audiovisual Speech Perception in Noise. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*. 1–10.
[29] Nicole Christoff, Nikolay N Neshov, Krasimir Tonchev, and Agata Manolova. 2023. Application of a 3D talking head as part of telecommunication AR, VR, MR system: systematic review. *Electronics* 12, 23 (2023), 4788.
[30] Peter Ciuha, Bojan Klemenc, and Franc Solina. 2010. Visualization of concurrent tones in music with colours. In *Proceedings of the 18th ACM international conference on Multimedia*. 1677–1680.
[31] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.
[32] Jody Cripps. 2018. Ethnomusicology & signed music: A breakthrough. *Journal of American Sign Languages & Literatures* 6 (2018).
[33] CuteCircuit. 2025. SoundShirt. https://cutecircuit.com/soundshirt/ Accessed November 11, 2025.
[34] Alice-Ann Darrow. 1993. The role of music in deaf culture: Implications for music educators. *Journal of Research in Music Education* 41, 2 (1993), 93–110.
[35] Jordan Aiko Deja, Alexczar Dela Torre, Hans Joshua Lee, Jose Florencio Ciriaco IV, and Carlo Miguel Eroles. 2020. Vitune: A visualizer tool to allow the deaf and hard of hearing to see music with their eyes. In *Extended Abstracts of the 2020 CHI conference on human factors in computing systems*. 1–8.
[36] Georgiana Cristina Dobre, Marta Wilczkowiak, Marco Gillies, Xueni Pan, and Sean Rintel. 2022. Nice is different than good: Longitudinal communicative effects of realistic and cartoon avatars in real mixed reality work meetings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
[37] John L Drever and Andrew Hugill. 2022. Aural diversity: General introduction. In *Aural Diversity*. Routledge, 1–12.
[38] Melanie Pius Dsouza, Ankitha Shetty, Sara Ellen Dsouza, Elisha Buthello, and Nachiket Gudi. 2025. Vibing the Young Consumer to Wellness: Exploring Lo-Fi Music Consumption Through the Positive Design Lens. *Sage Open* 15, 1 (2025), 21582440251318806.
[39] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)* 35, 4 (2016), 1–11.

[40] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[41] Isabel Sophie Fitton, Jeremy Dalton, Michael J Proulx, and Christof Lutteroth. 2022. Dancing with the avatars: Feedforward learning from self-avatars. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[42] Joyce Horn Fonteles, Maria Andréia Formico Rodrigues, and Victor Emanuel Dias Basso. 2013. Creating and evaluating a particle system for music visualization. *Journal of Visual Languages & Computing* 24, 6 (2013), 472–482.

[43] David W Fourney and Deborah I Fels. 2009. Creating access to music through visualization. In *2009 ieee toronto international conference science and technology for humanity (tic-sth)*. IEEE, 939–944.

[44] Yuan Gan, Ruijie Quan, and Yawei Luo. 2024. Expavatar: High-fidelity avatar generation of unseen expressions with 3d face priors. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[45] Caroline Guardino and Joanna E Cannon. 2016. Deafness and diversity: Reflections and directions. *American Annals of the Deaf* 161, 1 (2016), 104–112.

[46] Kaixin Han, Weitao You, Shuhui Shi, and Lingyun Sun. 2024. Hearing with the eyes: modulating lyrics typography for music visualization. *The Visual Computer* 40, 11 (2024), 8345–8361.

[47] Louise Hickman and Shannon Finnegan. 2020. Captioning on Captioning. https://lux.org.uk/work/captioning-on-captioning/ Short film on LUX (UK) website. Accessed 2025-11-11.

[48] Megan Hofmann, Devva Kasnitz, Jennifer Mankoff, and Cynthia L Bennett. 2020. Living disability theory: Reflections on access, research, and design. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.

[49] Jessica A Holmes. 2017. Expert listening beyond the limits of hearing: Music and deafness. *Journal of the American Musicological Society* 70, 1 (2017), 171–220.

[50] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 1–14.

[51] Anthony Hunt, Helena Daffern, and Gavin Kearney. 2023. Avatar representation in extended reality for immersive networked music performance. In *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*. Audio Engineering Society.

[52] Industrial Designers Society of America. 2020. Music: Not Impossible. https://www.idsa.org/awards-recognition/idea/idea-gallery/music-not-impossible/ IDEA Award Gallery. Accessed November 11, 2025.

[53] Kosuke Itoh, Honami Sakata, Ingrid L Kwee, and Tsutomu Nakada. 2017. Musical pitch classes have rainbow hues in pitch class-color synesthesia. *Scientific reports* 7, 1 (2017), 17781.

[54] Aobo Jin, Qixin Deng, and Zhigang Deng. 2020. A live speech-driven avatar-mediated three-party telepresence system: design and evaluation. *PRESENCE: Virtual and Augmented Reality* 29 (2020), 113–139.

[55] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. 2016. Effects of avatar and background representation forms to co-presence in mixed reality (MR) tele-conference systems. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: modelling and simulating virtual humans and environments*. ACM, 1–4.

[56] Byungdae Jung, Jaein Hwang, Sangyoon Lee, Gerard Jounghyun Kim, and Hyunbin Kim. 2000. Incorporating co-presence in distributed virtual music environment. In *Proceedings of the ACM symposium on Virtual reality software and technology*. 206–211.

[57] Luis Jure. 2012. Pitch content visualization tools for music performance analysis. *International Society for Music Information Retrieval Conference* (2012).

[58] Jun Kato, Tomoyasu Nakano, and Masataka Goto. 2015. TextAlive: Integrated design environment for kinetic typography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3403–3412.

[59] Christine Sun Kim. 2020. Artist Christine Sun Kim Rewrites Closed Captions. Pop-Up Magazine video. https://www.youtube.com/watch?v=tfe479qL8hg Accessed 2025-11-30.

[60] Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee, and Jong C Park. 2022. Sign language production with avatar layering: A critical use case over rare words. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1519–1528.

[61] Rachel Kolb. 2017. Sensations of Sound. Interactive feature, The New York Times. https://www.nytimes.com/interactive/2017/multimedia/sensations-of-sound-vr-rachel-kolb.html Accessed 2025-11-30.

[62] Danai Korre. 2023. Comparing photorealistic and animated embodied conversational agents in serious games: An empirical study on user experience. In *International Conference on Human-Computer Interaction*. Springer, 317–335.

[63] Raja S Kushalnagar and Christian Vogler. 2020. Teleconference accessibility and guidelines for deaf and hard of hearing users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.

[64] ChungHa Lee and Jin-Hyuk Hong. 2025. musicolors: Bridging Sound and Visuals For Synesthetic Creative Musical Experience. *arXiv preprint arXiv:2503.14220* (2025).

[65] Daniel G Lee, Deborah I Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)* 5, 2 (2007), 11.

[66] Sebin Lee, Geunmo Lee, Seongkyu Han, Seunghwa Jeong, and Jungjin Lee. 2023. A simulcast system for live streaming and virtual avatar concerts. *Journal of the Korea Computer Graphics Society* 29, 2 (2023), 21–30.

[67] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1272–1279.

[68] Jing Li, Chuchu Wang, and Mo Chen. 2025. Effects of Driving Background Complexity and Interface Opacity on Visual Cognition in AR-HUD Systems. *Journal of the Society for Information Display* 33, 8 (2025), 919–936.

[69] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020).

[70] Chang Liu, Qunfen Lin, Zijiao Zeng, and Ye Pan. 2024. EmoFace: Audio-driven emotional 3D face animation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 387–397.

[71] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[72] Jiaju Ma, Anyi Rao, Li-Yi Wei, Rubaiat Habib Kazi, Hijung Valentina Shin, and Maneesh Agrawala. 2023. Automated conversion of music videos into lyric videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

[73] Anabel Maler. 2013. Songs for hands: Analyzing interactions of sign language and music. *Music theory online* 19, 1 (2013).

[74] Anabel Maler. 2015. Musical expression among deaf and hearing song signers. *The Oxford handbook of music and disability studies* 2015 (2015), 73–91.

[75] SM Mather and MD Clark. 2012. The effect of visual split attention in classes for deaf and Hard of Hearing students, Odyssey: New Directions in Deaf Education. 2012 13: 20–24.

[76] Richard E Mayer. 2005. Principles of multimedia learning based on social cues: Personalization, voice, and image principles. *The Cambridge handbook of multimedia learning* (2005), 201–212.

[77] Thomas Barlow McHugh, Abir Saha, David Bar-El, Marcelo Worsley, and Anne Marie Piper. 2021. Towards inclusive streaming: Building multimodal music experiences for the deaf and hard of hearing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[78] Imtiyaz Ali Mir, Moniruddin Chowdhury, Rabiul Md Islam, Goh Yee Ling, Alauddin ABM Chowdhury, Zobaer Md Hasan, and Yukihito Higashi. 2021. Relaxing music reduces blood pressure and heart rate among pre-hypertensive young adults: A randomized control trial. *The Journal of Clinical Hypertension* 23, 2 (2021), 317–322.

[79] JaeYoung Moon, Youjin Choi, Jin-Hyuk Hong, and Kyung-Joong Kim. 2025. Sign Dance Maker: A Generative Ai-Assisted Framework for Inclusive Music Performance Support for Sign Language Interpreters. *Available at SSRN 5245083* (2025).

[80] Jorge Mori and Deborah I Fels. 2009. Seeing the music can animated lyrics provide access to the emotional content in music for people who are deaf or hard of hearing?. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. IEEE, 951–956.

[81] Suranga Chandima Nanayakkara, Elizabeth Taylor, Lonce Wyse, and SH Ong. 2007. Towards building an experiential music visualizer. In *2007 6th International Conference on Information, Communications & Signal Processing*. IEEE, 1–5.

[82] Eva Naumann. 2025. Human-Like Avatar Embodiment: Advantage or Disadvantage in Digital Emotion Regulation Intervention? *Available at SSRN 5440041* (2025).

[83] Ferda Ofli, Cristian Canton-Ferrer, Joëlle Tilmanne, Yasemin Demir, Elif Bozkurt, Yucel Yemez, Engin Erzin, and A Murat Tekalp. 2008. Audio-driven human body motion analysis and synthesis. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2233–2236.

[84] Keita Ohshiro and Mark Cartwright. 2022. How people who are deaf, Deaf, and hard of hearing use technology in creative sound activities. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[85] Yifang Pan, Rishabh Agrawal, and Karan Singh. 2024. S3: speech, script and scene driven head and eye animation. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–12.

[86] Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. 2022. Vocal: Vowel and consonant layering for expressive animator-centric singing animation. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[87] Michael Pouris and Deborah I Fels. 2012. Creating an entertaining and informative music visualization. In *International Conference on Computers for Handicapped Persons*. Springer, 451–458.

[88] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. 2011. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*. 147–156.

[89] Pablo Revuelta, Tomás Ortiz, María J Lucía, Belén Ruiz, and José Manuel Sánchez-Pena. 2020. Limitations of standard accessible captioning of sounds and music for deaf and hard of hearing people: An EEG study. *Frontiers in integrative neuroscience* 14 (2020), 1.

[90] Filipa M Rodrigues, Ana Maria Abreu, Ingela Holmström, and Ana Mineiro. 2022. E-learning is a burden for the deaf and hard of hearing. *Scientific Reports* 12, 1 (2022), 9346.

[91] Justin Salamon and Emilia Gómez. 2012. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE transactions on audio, speech, and language processing* 20, 6 (2012), 1759–1770.

[92] Patricia A Shaw. 2008. Scat syllables and markedness theory. *Toronto Working Papers in Linguistics* 27 (2008).

[93] Tracey Skelton and Gill Valentine. 2003. 'It feels like being Deaf is normal': an exploration into the complexities of defining D/deafness and young D/deaf people's identities. *Canadian Geographer/Le Géographe Canadien* 47, 4 (2003), 451–466.

[94] Wenfeng Song, Xianfei Wang, Yang Gao, Aimin Hao, and Xia Hou. 2022. Real-time expressive avatar animation generation based on monocular videos. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 429–434.

[95] Katta Spiel, Kathrin Gerling, Cynthia L Bennett, Emeline Brulé, Rua M Williams, Jennifer Rode, and Jennifer Mankoff. 2020. Nothing about us without us: Investigating the role of critical disability studies in HCI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[96] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. 2024. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–9.

[97] Veronika Szucs, Beata Kovacs, and Balint Tasnadi. 2018. Music for seeing–visualization of sounds. In *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 000123–000128.

[98] Olivia Ting. 2024. Between Piano and Forte: Hearing with Aids. *Leonardo* 57, 2 (2024), 153–161.

[99] Pauline Tranchant, Martha M Shiell, Marcello Giordano, Alexis Nadeau, Isabelle Peretz, and Robert J Zatorre. 2017. Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in neuroscience* 11 (2017), 507.

[100] Tai-Chen Tsai, Yu-Hsuan Chen, Min-Chun Hu, and Tse-Yu Pan. 2024. DualStage: Enhancing Emotional Connections through Music Visualization for Synchronizing Live and Virtual Performances. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 771–775.

[101] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 448–458.

[102] Nancy Tye-Murray, Mitchell S Sommers, and Brent Spehar. 2007. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing* 28, 5 (2007), 656–668.

[103] Bavo Van Kerrebroeck, Giusy Caruso, and Pieter-Jan Maes. 2021. A methodological framework for assessing social presence in music interactions in virtual reality. *Frontiers in Psychology* 12 (2021), 663725.

[104] Quoc V Vy, Jorge A Mori, David W Fourney, and Deborah I Fels. 2008. EnACT: A software tool for creating animated text captions. In *Computers Helping People with Special Needs: 11th International Conference, ICCHP 2008, Linz, Austria, July 9-11, 2008. Proceedings 11*. Springer, 609–616.

[105] Xueyang Wang, Sheng Zhao, Yihe Wang, Howard Ziyu Han, Xinge Liu, Xin Yi, Xin Tong, and Hewu Li. 2025. Raise Your Eyebrows Higher: Facilitating Emotional Communication in Social Virtual Reality Through Region-Specific Facial Expression Exaggeration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[106] Yubo Wang, Fengzhou Pan, Danni Liu, and Jiaxiong Hu. 2023. Music-to-facial expressions: emotion-based music visualization for the hearing impaired. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 16096–16102.

[107] Sijing Wu, Yunhao Li, Weitian Zhang, Jun Jia, Yucheng Zhu, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. 2025. SingingHead: A large-scale 4D dataset for singing head animation. *IEEE Transactions on Multimedia* (2025).

[108] Hiromu Yakura and Masataka Goto. 2020. Enhancing participation experience in vr live concerts by improving motions of virtual audience avatars. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 555–565.

[109] Suhyeon Yoo, Georgianna Lin, Hyeon Jeong Byeon, Amy S Hwang, and Khai Nhut Truong. 2023. Understanding tensions in music accessibility through song signing for and with d/Deaf and Non-d/Deaf persons. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[110] Suhyeon Yoo, Khai N Truong, and Young-Ho Kim. 2025. ELMI: Interactive and Intelligent Sign Language Translation of Lyrics for Song Signing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[111] Kyohei Yoshikawa, Takashi Machida, Kiyoshi Kiyokawa, and Haruo Takemura. 2004. A high presence shared space communication system using 2D background and 3D avatar. *IEICE TRANSACTIONS on Information and Systems* 87, 12 (2004), 2532–2539.

[112] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*. Springer, 1–19.

[113] LUIZA ZAN and STELA DRĂGULIN. 2022. Vocal Depersonalization in Scat Singing. *Studia UBB Musica* 67, 1 (2022).

[114] Eduard Zell, Carlos Aliaga, Adrian Jarabo, Katja Zibrek, Diego Gutierrez, Rachel McDonnell, and Mario Botsch. 2015. To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–12.

[115] Fan Zhang, Molin Li, Xiaoyu Chang, Kexue Fu, Richard William Allen, and RAY LC. 2025. " Becoming My Own Audience": How Dancers React to Avatars Unlike Themselves in Motion Capture-Supported Live Improvisational Performance.. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.

[116] Shu Zhang, Xinge Liu, Xuan Yang, Yezhi Shu, Niqi Liu, Dan Zhang, and Yong-Jin Liu. 2021. The influence of key facial features on recognition of emotion in cartoon faces. *Frontiers in psychology* 12 (2021), 687974.

[117] Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 conference papers*. 1–13.

[118] Kyrie Zhixuan Zhou, Weirui Peng, Yuhan Liu, and Rachel F Adler. 2024. Exploring the Diversity of Music Experiences for Deaf and Hard of Hearing People. *arXiv preprint arXiv:2401.09025* (2024).

A Appendix: Supplementary Figure and Tables

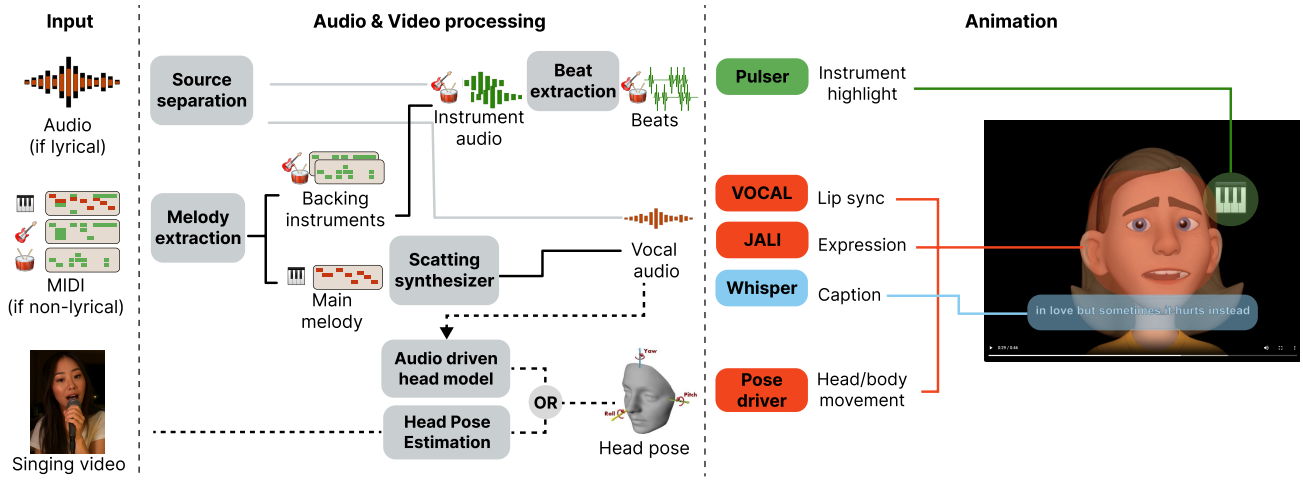

Figure 12: Pipeline for generating Avatar animation, instrument highlight and caption for lyrical music (based on song audio, highlighted by grey lines) and non-lyrical music (based on MIDI input, highlighted by black lines).

Table 5: List of songs used for Study 1

Valence	Lyrical	Artist – Title	Genre	YouTube MV/PV
Positive	No	Antonio Vivaldi – Spring (The Four Seasons)	Classical	https://www.youtube.com/watch?v=6LAPFM3dgag
Positive	No	Koji Kondo – Mario Theme	Techno	https://www.youtube.com/watch?v=_9bB7r0M9kg
Positive	Yes	Pharrell Williams – Happy	Pop	https://www.youtube.com/watch?v=ZbZSe6N_BXs
Positive	Yes	Mark Ronson (ft. Bruno Mars) – Uptown Funk	Funk	https://www.youtube.com/watch?v=OPf0YbXqDm0
Negative	No	John Williams – Star Wars (The Imperial March)	Classical	https://www.youtube.com/watch?v=-bzWSJG93P8
Negative	No	Antonio Vivaldi – Winter (The Four Seasons)	Classical	https://www.youtube.com/watch?v=Yu6Hr9kd-U0
Negative	Yes	Adele – Someone Like You	Pop	https://www.youtube.com/watch?v=hLQl3WQQoQ0
Negative	Yes	Lady Gaga, Bruno Mars – Die With a Smile	Pop	https://www.youtube.com/watch?v=kPa7bsKwL-c

Table 6: Avatar demographics for Study 2

Title	Avatar
Spring (The Four Seasons)	Male / Black
Mario Theme	Female / Asian
Happy	Male / Black
Uptown Funk	Male / White
Imperial March	Male / White
Winter (The Four Seasons)	Female / Black
Someone Like You	Female / White
Die With a Smile	Female / Asian

Table 7: Conditions in the perception test (Study 2 - Phase 1)

Visuals	Audio
Spring	Correct Song
Spring	Different Part
Spring	Different Song
Imperial March	Correct Song
Imperial March	Different Part
Imperial March	Different Song
Happy	Correct Song
Happy	Different Part
Happy	Different Song
Someone Like You	Correct Song
Someone Like You	Different Part
Someone Like You	Different Song