Visualization of Speech Prosody and Emotion in Captions: Accessibility for Deaf and Hard-of-Hearing Users

Caluã de Lacerda Pataca
Computing and Information Science
Rochester Institute of Technology
Rochester, NY, USA
cd4610@rit.edu

Matthew Watkins
Roshan L Peiris
School of Information
Rochester Institute of Technology
Rochester, NY, USA
{mxw7981,rxpics}@rit.edu

Sooyeon Lee
Informatics/Ying Wu College of Computing
New Jersey Institute of Technology,
Newark, NJ, USA
sooyeon.lee@njit.edu

Matt Huenerfauth
School of Information
Rochester Institute of Technology
Rochester, NY, USA
matt.huenerfauth@rit.edu

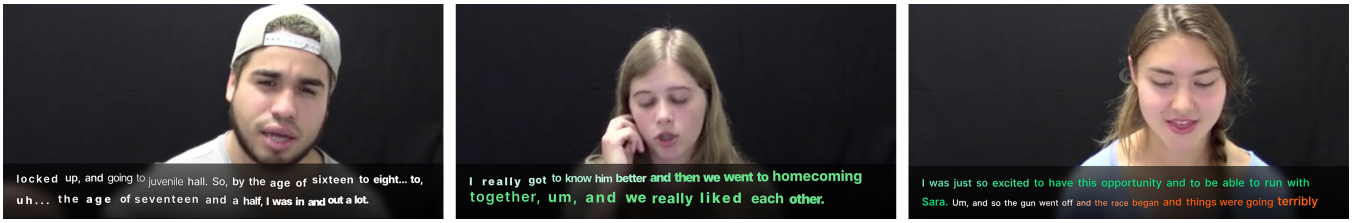


Figure 1: The three captioning models presented in this paper. Going beyond conventional models of captions that are limited to depicting only spoken words, the proposed models overlay visual depictions of prosody, prosody and emotions, and emotions.

ABSTRACT

Speech is expressive in ways that caption text does not capture, with emotion or emphasis information not conveyed. We interviewed eight Deaf and Hard-of-Hearing (DHH) individuals to understand if and how captions' inexpressiveness impacts them in online meetings with hearing peers. Automatically captioned speech, we found, lacks affective depth, lending it a hard-to-parse ambiguity and general dullness. Interviewees regularly feel excluded, which some understand is an inherent quality of these types of meetings rather than a consequence of current caption text design. Next, we developed three novel captioning models that depicted, beyond words, features from prosody, emotions, and a mix of both. In an empirical study, 16 DHH participants compared these models with conventional captions. The emotion-based model outperformed traditional captions in depicting emotions and emphasis, with only a moderate loss in legibility, suggesting its potential as a more inclusive design for captions.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581511

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**; *Empirical studies in accessibility*.

KEYWORDS

Accessibility, Emotion / Affective Computing, Individuals with Disabilities & Assistive Technologies, Empirical study that tells us about how people use a system

ACM Reference Format:
Caluã de Lacerda Pataca, Matthew Watkins, Roshan L Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of Speech Prosody and Emotion in Captions: Accessibility for Deaf and Hard-of-Hearing Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3581511

1 INTRODUCTION

The last decade has seen remarkable progress in automatic speech recognition, to the point that the word error rate for state-of-the-art systems has surpassed that of human-made transcriptions [32]. These systems' presence in a myriad of low-powered, ubiquitous, and portable personal devices has allowed for many use cases, planned or unforeseen [47, 51]. This includes their use for automatically generated captions during video conferencing meetings with Deaf and Hard-of-Hearing (DHH) individuals.

And yet, the way captions depict words has seen little change. Discussions of making captions more expressive have been ongoing

in the literature for at least 40 years, e.g., [55]. Although captions have moved beyond uppercase letters displayed white-over-black in coarse fonts [4], conventional captioning approaches are still based on a visually simple[1] depiction of speech. This effect is heightened in automatically generated captions, as seen in conferencing software: whereas a stenographer might give hints of non-speech and paralinguistic information in parentheses, automatic captions invariably present a neutral depiction of words spoken.

Human speech, however, is meaningful beyond just its words. Captions will typically not depict features such as prosody (*how loud, melodic, or fast does someone's voice sound?*), vocal quality (*does it sound old or young?*), a speaker's disposition (*is it tired or excited?*), or even their emotions (*does it carry anger or joy?*). While for asynchronous videos, it is possible for captions to be prepared by a human professional who could add parenthetical indications of this information, for synchronous video conferencing, software is needed that would analyze acoustic or lexical properties of the speech signal in real-time and then convey this information to users, e.g., through text styling.

As motivation and guidance for work in this area, it is necessary to understand whether the current limited approach to conveying speech in captions negatively impacts DHH users, particularly in communication settings between DHH and hearing individuals. If this is a problem, how could caption depiction be enhanced to embed these (paralinguistic) features of speech?

From this first question stems the first empirical study reported in this paper. In in-depth interviews with 8 DHH individuals, we probed their experience with automatic captions used for meetings with hearing peers, focusing on captions' non-depiction of prosody and emotions, the consequences of this, and workarounds. Our three main findings are that among other failings, (1) captions' depiction of words is felt as leaving out meaningful dimensions of speech, and (2) in automatically captioned meetings, this can lead to DHH individuals often feeling left out. This state of affairs has been naturalized to the point that (3) some interviewees seem to accept that there are types of conversations that they won't be able to participate in, as if an inherent reality of automatically captioned speech and not a consequence of the design of these systems.

For our second study, we investigated what dimensions of non-linguistic speech can most help DHH individuals decode a speaker's emotions and emphasis when they are depicted in captions. For this, we created three realistic prototypes of captioning systems that represented, beyond the linguistic content of speech, its prosody (shown in Figure 3b), emotions (shown in Figure 3c), and prosody and emotions simultaneously (shown in Figure 3d). As this paper focuses on live video communication contexts, the stimuli in this study consisted of videos of a single person speaking while looking into a camera, a perspective similar to videoconferencing. In an empirical, comparative study with 16 participants, we found that (1) the two captioning models that included emotions outperformed conventional captions in making a speaker's emotions, moods, and emphasis easy to identify. Also, (2) participants' interest in using these emotion-depicting captions (but not for the prosody-depicting variants) was comparable to convention captions for the workplace and personal settings.

There are two main empirical contributions of this work: We learned that captions' non-depiction of speech's emotion and emphasis limits how inclusive they can make video-conferencing communication between hearing and DHH individuals. Also, we present empirical evidence of benefits from conveying *emotional* information from the speaker in captions.

2 BACKGROUND AND RELATED WORK

Videoconferencing presents several challenges for DHH individuals [42]. These challenges include difficulties with sign language interpreters and how conferencing software deals with them, such as their placement among other users and the clarity of their signs [69, 78]. DHH users may also be ignored because of the aural-centric nature of the medium [67]. Other issues come from translating sound into visuals. The visual channel can easily be overloaded in multimodal settings. While, for instance, a hearing person can look at slides while listening to a speaker, a DHH person may need to switch back and forth between the two, leading to visual dispersion and loss of information [19, 43].

Some of these concerns are not exclusive to sign languages and are also felt in communication modalities such as lip-reading or closed-captioning [38]. Particularly with the latter, issues can stem from the lack of representation in captions of non-linguistic dimensions of speech, a topic that is understudied in the literature and to which our paper hopes to contribute.

Prosody, one such dimension, describes much of what we perceive as tone of voice [6]. It is a linguistic signal which produces a *procedural* encoding of information [82]. It guides a listener through a sentence, narrowing the search space of *plausible* meanings. It also helps convey a speaker's emotions, moods, and dispositions. As with facial expressions, a speaker conveys prosody through voluntary and involuntary processes, interpreted by a listener both instinctively and deliberately [5, 45].

Prosody does not negate linguistic semantics, but it can add meaning of its own. This can be seen when someone removed from linguistic understanding can still assign meaning to speech, e.g., a non-native speaker decoding moods and emotions from speech in a language they don't understand otherwise [22, 39].

2.1 Visualizing prosody and emotions through written text

Since typography is the medium through which captions convey their message, it is worth exploring ways by which it has been used to represent information beyond the written word. Various authors have explored ways of changing typography with new graphical elements [2, 25], expanded palettes of letters [77] and punctuation marks [16]– all intended as ways to convey emotions [61, 64] or prosody [2, 11, 18, 23, 60, 66, 83].

Some systems have extracted prosody from recorded speech to use as an input for typographic modulations, i.e., mapping acoustic features to visual changes in the text. These modulations can come from algorithmically processing audio signals [18, 23, 66, 83], or by manually manipulating typography according to an artist's interpretation of speech [44, 64].

[1] For instance, while the CEA-708 standard for digital TV supports many visual features, authoring tools still mostly support the limited, analog-era CEA-608 standard [46].

Emojis have also been explored to represent paralinguistic dimensions of speech in text [34, 44]. Hu et al. [34] mapped a mix of automatically extracted emotions from speech audio and semantic analysis of its content to one of four emojis, which were seen as reducing the loss of perceived emotions from sound to text. Lee et al. [44] worked with a film director to manually add emojis and colors to represent emotions in captions.

2.2 Strengths and shortcomings of automatically generated captions

From the outset, closed-captions[2] were seen as a way of making speech audio accessible for DHH individuals. Authors commented how DHH individuals could *'watch many of the same popular TV programs at the same time and with the same understanding as their neighbors'* [17]. While other benefits of using captions were found since then [30], that original vision was overly optimistic.

Captions lack expressive elements of speech, speaker identification, latency, etc. [42] Still, even the *linguistic* component of captions is not a settled issue. Measuring its quality is tricky, particularly with automatically generated captions. Many automatic speech recognition (ASR) quality metrics are quite technical, and rarely are they actually validated with DHH users [37, 76]. Aksënova et al. [1] discuss that word error rate has many overlooked subtleties, but by itself, it is an insufficient proxy for *perceived quality*, which can vary in the presence of regional language variation, non-native accents, different genders, ages, latency, domain-specific lexica, etc. In evaluating a state-of-the-art ASR system for Dutch, for example, Feng et al. [27] found significant differences in performance for speakers of different genders, ages, regional accents, and, markedly, non-native accents.

Considering automatically generated captions' use, particularly in the workplace, research has shown that current technologically-mediated communication practices between DHH and hearing individuals are cumbersome and need improvement [26]. Speaker's behavior plays a role in how well ASR and DHH individuals can follow speech. The use of ASR systems correlates with speakers' changing their vocal articulation [71]. Captioning styles complemented with graphical overlays have aided hearing speakers in adapting their speaking behavior, potentially boosting comprehension both by DHH audiences and ASR systems [52, 72].

Regarding the visual form of captions and changes to their design, preferences are widely spread, hinting at there being room for experimentation, but that some users might be resistant to changes. Scene occlusion and legibility are critical considerations. In small-group online meetings, Berke et al. [9] saw a tension between caption styles that either favor legibility (e.g., TV captions, with black boxes behind text) or less image occlusion (e.g., transparency behind letters), with divided preferences among participants.

Past experiments investigated visually overlaying additional dimensions of meaning on captioning systems. Berke et al. experimented with different visual strategies to represent the uncertainty that ASR systems assign to each predicted word. Participants generally preferred unmarked captions and, when captions depicted uncertainty levels, felt the 'markup style was too distracting.' [10]

[2] *Closed*, here, means captions that the viewer can choose whether to turn on or off, as opposed to *open* captions, which are fixed.

While some authors acknowledge that dimensions of speech such as prosody and emotions are not made accessible by traditional captioning styles [42, 52], few studies have tackled the issue. Those that do tend to overlook captioning, focusing on acoustic modeling and speech-modulated typography [60, 70, 83], with only a handful of authors specifically validating their approaches through DHH individuals' perspectives [44, 64]. Even considering the latter, the concepts explored were aimed at prerecorded media. To the best of our knowledge, no previous studies have investigated depictions of prosody and emotions with captioning approaches adequate for automatic speech recognition, which could be of use in videoconferencing settings.

2.3 Research questions

Knowing the importance of dimensions of speech such as prosody and emotions, and given that they are not depicted in captions, Study 1 investigated:

- RQ1.A In what ways do DHH individuals experience the absence of prosodic and emotional depictions in computer-generated captions, as are used in online meetings with hearing peers?
- RQ1.B How can their current experiences and workaround strategies inform the design of new captioning systems that depict prosody and/or emotions?

Results from this first study revealed that captions' inexpressive representations of speech render it ambiguous for DHH users. Participants' comments were less clear in indicating *what* non-textual dimensions of speech could be most helpful to alleviate these communication issues and in what settings their use would be most appropriate. Considering the different combinations of these dimensions, we built a set of caption prototypes that, in a second study, allowed us to investigate:

- RQ2.A How easy to identify are a speaker's emotions, moods, and emphasis when captions depict prosody and/or emotions beyond just words?
- RQ2.B In what settings is the use of these visual depictions of prosody and/or emotion felt as the most appropriate from the point-of-view of DHH individuals?

3 STUDY 1: INTERVIEW STUDY

To address RQ1.A and RQ1.B, Study 1 focused on how DHH individuals experience computer-generated captions when in remote meetings with hearing peers. More specifically, we were interested in uncovering the impact of these captions' lack of representation of paralinguistic dimensions of speech like prosody and emotions, and what our interviewees' experiences could inform us about the design of these representations.

3.1 Design of Study 1

We conducted a semi-structured interview study, divided into two parts: (1) an exploration of participants' experiences with remote meetings, including some questions about how speakers' emotions, moods, and other dimensions are handled; (2) new ideas for how caption design could be enhanced to include these dimensions — given the technical nature of a discussion about prosody-based and emotional models to represent speech, a brief introduction to these

topics was given,[3] after which we also asked about participants' experiences with and preferences for each.

After the semi-structured section of the interview was over, we showed three early design prototypes of captioning systems that depicted prosody and emotion. The goal was to gauge participants' initial reactions to each design.

3.1.1 Early designs for depicting prosody and emotion in captions. Thinking ahead to the possibility that we may need to create some prototypes for how to convey dimensions of speech through captions (to show to participants in study 2), we decided to use the final few minutes of study 1 to show participants a few videos with some initial designs. While the caption styles used in study 2 would be redesigned following our analysis of the interviews in study 1, we wanted to gather some initial reactions that might later inform our work in the subsequent study.

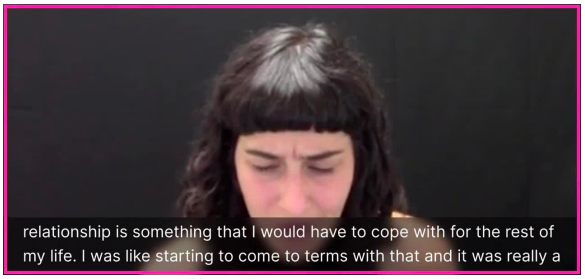
Given the sparse literature on the depiction of prosody and emotions in captioning systems, we consulted with design and typographic professionals to discuss these initial prototypes. This was based not on assuming they would be good judges of how the DHH community would respond to these designs — knowledge not yet well established — but, rather, because type design typically deals with achieving expressiveness within tight constraints (e.g. Erik Spiekermann's claim that 95% of a given font has to look like any other font, leaving type designers with only 5% to differentiate their work [35]), a setting not too dissimilar to ours. While our research goals were not aimed at establishing design recommendations for new caption styles, we still needed designs that were sufficiently expressive and to that end, these consultations could be one element helping us find *good enough* design parameters.

In conversations that lasted up to two hours, we showed and discussed an initial set of ideas with four domain experts in type and graphic design. We received assorted feedback: using visual particles seemed problematic since they would demand too much visual focus from the user, an already overloaded resource (especially for scrolling captions [62]). Experts also suggested that we increase the contrast between various visual properties to make them more apparent. After further refinement based on this feedback from experts, the three designs shown in Figure 2 came out as the most promising to be presented at the end of the interviews in study 1.

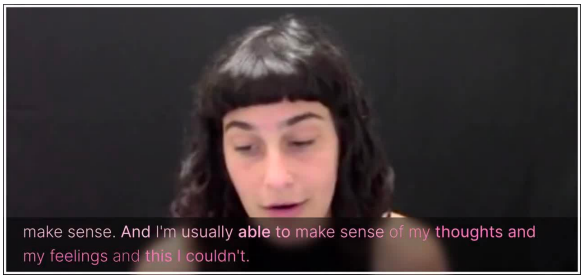
For these designs, we worked with videos available in the Stanford Emotional Narratives Dataset [59]. These are short videos of persons looking at the camera and telling stories with strong emotional valence (negative, positive, or both). Included with each are self-reported scores for valence and text transcriptions.

Automatic captioning systems used in live meetings typically present text one word at a time (*scrolling captions*), instead of as *block captions* [54]. To approximate this behavior in the prototypes, we divided these 5-second text blocks into even-sized chunks, each corresponding to a single word. Each was then assigned a value for valence, interpolated from the dataset, and loudness, measured from each word's timestamp.

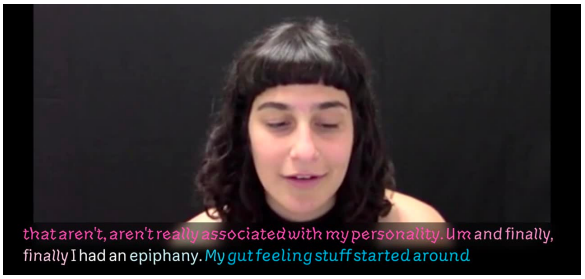
[3]For brevity's sake, we explained a simplified, one-dimensional version of the circumplex model of emotion, reduced to valence, i.e., words were said to be negative, neutral, or positive. The idea of contrasting which words are important was related to prosody, which we also didn't break down in constituent dimensions of loudness, pitch, and duration.



(a) Colored, ondulating border



(b) Changes to font weight and color



(c) Promphan [61]'s valence typography.

Figure 2: Three exploratory designs for an enhanced captioning system shown to DHH interviewees.

In terms of how valence and loudness were depicted, one design had a border whose thickness varied with loudness and whose color varied with valence (Figure 2a); in another, words had their font-weight related to loudness and color to valence (Figure 2b); in the last, a specially designed 'emotional' typeface with five unique, but related, letter shapes, going from very negative to neutral to very positive (Figure 2c); etc.

3.2 Participants and recruitment

The IRB-approved study ran from March to May 2022. One of the authors, who is Deaf, conducted eight semi-structured interviews with DHH individuals, sometimes in ASL, sometimes in English and ASL, the latter accompanied by an interpreter. Interviews took on average 51 minutes ($\sigma = 14'$). Participants were recruited through social media, particularly through DHH specific Reddit channels, and DHH specific mailing lists. $40 compensation was offered. The screening factor used was whether the person identified themselves as DHH and had had previous experience working with hearing colleagues sometime in the past five years. Out of the eight participants,

five identified themselves as female, two as male, and one as non-binary. Three identified themselves as hard-of-hearing, and five as deaf/Deaf. Their average age was 26 ($\sigma = 8$).

3.3 Analysis of Study 1

Research assistants fluent in ASL transcribed the eight interviews. On this qualitative data, we performed an iterative thematic analysis [15] employing both deductive and inductive approaches. Informed by the interview questions, we developed a framework that allowed high-level categorization of the transcripts and within this deductively formed structure, inductive bottom-up analysis was conducted through an iterative coding process. After refining the codes (e.g., *loss of affective information*) with several rounds of collating and grouping, a list of themes was generated (e.g., *Captions' dull ambiguity*). All authors reviewed them in meetings and finally synthesized them into four themes presented in the next section. A summary of participants' reactions to the prototypes is reported in subsection 3.4.5.

3.4 Findings of Study 1

With varying degrees, interviewees related having faced hardships when using automatic captioning systems to communicate with hearing peers. While some issues stem from failings of the current state of automatic speech recognition software, a lack of depiction of prosody and emotions emerged as a cause for *captions' dull ambiguity*. Since interviewees faced this on a nearly daily basis, *communication becomes an uphill battle*, with significant cognitive and emotional tolls.

Strategies to alleviate these ambiguities are diverse and include *reliance on multimodal signals* such as facial expressions, body language, and general engagement. Interviewees indicated that using these cues is not a straightforward, lossless process, and they were therefore favorable towards the promise of captions' depicting prosody and emotions. There was nuance to this preference: given how multifaceted these features can be and given the diversity of what is needed in particular settings, *different contexts call for different solutions*.

Where needed, quotes were edited for clarity and conciseness.

3.4.1 Theme 1: Captions' dull ambiguity. Captions' imperfections are felt in different ways. Automatic speech recognition capabilities in live captions have gotten better in recent years, but they still leave a lot to be desired. Agatha[4]: *'Sometimes it's really, really, slow. Someone speaks, and when a few seconds later the caption finally appears the speaker is already on the next topic.'* Eliah: *'Often, when people speak with accents captions will have a lot of mistakes.'* Otto: *'They're horrible, missing context and words. It takes a lot of work to understand exactly what they're saying.'*

Beyond how latency and imprecision can make the linguistic content hard to understand, there are also consequences related to how a shift in mood can go unnoticed. Alex tells us of a time when there was a quick shift from a casual to a serious topic that wasn't apparent right away, leading to them *'jumping in at the wrong time and causing my hearing colleagues to look down at me like I'm bad at reading people, which I'm not.'*

[4]All names are pseudonyms.

Human-made transcriptions of pre-recorded allow for greater accuracy, but even when written words perfectly match written counterparts, still, something seems missing. A common occurrence is failing to understand whether comments are serious or not. Alex says that since *'comedy relies heavily on tone, hearing people can understand immediately when something is a joke, but my friend, who is also DHH, has a hard time because they're missing that tone.'* Erin tells of a time when *'someone was telling a story that had a specific inside joke, and I had no idea what was going on because it was connected to the tone.'* Otto: *'Sometimes I'll realize it's a joke after they look at me and ask whether I understood it, and I was like "oh, I thought it was serious because the captions seemed serious."'*

Participants complained about the monotonous, droning quality of inexpressive captions. Alex tells that because of this they find it hard to focus on captions: *'I can find it easy to zone out because speech is not really... emphasized?'* Erin finds the contrast between captions and signing hard since she *'grew up accustomed to some use of body language, so it is hard to just watch and read captions all of the time.'* Otto: *'Facial and body language will show a lot of context, while captions are bland.'*

All of this gives captioned speech an unapproachable ambiguity that disproportionately affects DHH individuals. This is particularly true with dimensions of communication that are already inherently ambiguous, such as moods and emotions. Otto thinks that this disconnection is analogous to texting, which *'tends to be devoid of emotion. It's better to interact with the person, to see their real raw emotion, while texting hides it, making it hard to be emotionally transparent.'* For Agatha, when reading captions she tends to miss *'meaning or feeling behind the words.'* To deal with this, she usually has *'to read the full paragraph of what was said or have the picture of the speaker's face, but even then there's a delay in understanding.'*

3.4.2 Theme 2: Communication as an uphill battle. Working and studying among hearing peers, our interviewees relate recurrent feelings of isolation. The frequent shortcomings of captioning systems fall almost exclusively on their shoulders, leaving them forced to either speak up or face missing out on what's going on. Ira told us of how in meetings her peers can at times urge her to *'use captioning right away, but I feel awkward because I'm the only person using them. Sometimes I will miss something and feel awkward to ask hearing group mates to repeat themselves; it just feels weird.'*

Sometimes, it's only when they later read a meeting's transcript that what was said becomes clear. Agatha: *'I later understood, but I had to go back and read the transcript to fully understand.'* Eliah: *'It's nice that live captions' transcriptions can be saved as a transcript so I can catch up to what was said.'*

For some, this distance from peers has become naturalized. Erin: *'I am curious that I don't know what's happening and I just have to wait there. I know that I am frustrated but at the same time I know that I have to collaborate. I can't expect it to be easy to communicate all of the time.'* She later adds *'I tend to accept it because of work. Every weekend, they talk about parties and I accept that I am not part of that conversation and just leave it.'*

Some environments are more welcoming than others. Otto's manager makes a point of checking how their captions are coming out, saying *' "lemme check the captions... Oh no, I didn't mean that," and then repeating themselves until the captions are accurate.'* Eliah's

boss writes them a summary of what is being discussed because even with an interpreter and captions *'there's a lot of overlap and I can't really catch the specifics.'*

The flip side is that DHH persons depend on sometimes-lacking goodwill from their hearing colleagues to be included in the conversation. Ada: *'Often, my coworkers forget that I need a good environment before I can understand, so they'll be having a conversation with background noise, or not looking at me. I'll still try, but I'll feel alone and left out.'* Irene also faces issues with her coworkers' carelessness with how their environment can impact accessibility. When she raises the issue to leadership, they might try to do something, *'but the other members of the group are not as willing and, especially since COVID, have reached their limits.'* Otto: *'I try to be assertive, trying to talk to them, but even if I type in the chat some hearing people don't know how to use it or just ignore it and keep talking anyway. That means I can't do much about it.'* When she does intercede, Agatha feels that *'with the captions, I'm delayed, so if I had questions I need to ask them to go back on the conversation. I feel like it's annoying to my boss.'*

The emotional ambiguity of captions heightened these feelings of isolation. Eliah: *'With a large team, it's hard to see their faces and I usually depend on captions. I don't know their emotions, and I feel like I'm not there, not connected with them.'* Otto feels missing emotional representation always impacts him: *'In general communication, I can't really participate fully. The discussion can be work-related but there's also another discussion that's humorous, and I wouldn't understand. Most of them are laughing and I'm left out, unsure whether they're actually joking or not.'*

3.4.3 Theme 3: Reliance on multimodal signals. Interviewees related being very attuned to how people communicate with their facial expressions, body language, and general engagement. This, some said, is a way to tackle the shortcomings of captioning. Ira: *'When people are talking I can look and figure out what their thoughts are based on their behavior. With masks, I sometimes miss out on information, so I'll look at their eyebrows or eyes, but it's hard.'* Alex says that to gauge mood or emotion they *'have to look up from the captions at their expressions, body language, and how they react so I can tell what they mean,'* although *'that doesn't mean I capture all the information.'* In describing what makes a speaker's emotions easy to identify, Erin says that it comes out to *'body language; how they're shifting in their seat, how they're moving, their facial expressions, and mouth movement.'*

Cultural differences come into play here. Erin: *'Here in America you can definitely identify it easily, but in other countries, it's challenging.'* Irene: *'It is very hard to understand hearing people's body language and tone, especially through the computer. They tend to sit very relaxed with their hands on their face, or look neutral, while Deaf people are extremely expressive and clear.'*

Technology adds to the complexity of navigating this mosaic of affective signals, and this is present in Agata's comment that, *'with captions, sometimes I miss the facial expressions or emotions behind the words.'* The delay in captions makes Ada struggle with trying to listen and read at the same time: *'it's really hard: I have a choice of either listening to the person or reading the captions, but trying both simultaneously takes more work and won't help me.'*

3.4.4 Theme 4: Different contexts call for different solutions. Having introduced to the interviewees the idea that a speaker's tone of voice can serve to signify both different emotions and as a contrastive focus to emphasize certain words, we asked them what they'd think would be most important to represent in captions. Answers were varied and were tied to what interviewees felt was appropriate for different types of meetings.

Some, such as Otto, claimed that while both dimensions are important, in work environments one should prioritize prosody, *'because I need to understand information better, to pay attention to which word is important. Emotion is important, yes, but I'd rather hold off on that because it's more suitable for general communication.'* Eliah echoed this: *'We don't need to depend on mood because we're here for work. The working environment usually has a lot of discussions so it's important to have emphasis so that we can be involved, discuss more, and ask more questions as deaf people.'*

Others were undecided. Ada, for instance, said that while for her prosody would be generally more important, when their hearing is fatigued, *'I no longer can figure out valence myself, so it would then become the more important one.'* Alex: *'I think both should be included. Valence can show emotion, but not what's important; prosody emphasizes what's important, but not emotion, so how would I know?'*

Others preferred the representation of emotions. In Erin's case, the choice between prosody or valence was almost a tie, but *'emphasis I can figure out, while emotion is really nice to have on the screen so that I know what is going on.'* Ira: *'emotion is more important since it helps to visualize the full picture, which deaf people usually miss, while emphasis is just for a specific word.'* For Agatha, *'emotions add more depth to words,'* and are thus more important to be visualized.

3.4.5 Design recommendations from the pilot study. Reactions to the design prototypes shown after the interviews were mixed. There was an appreciation of the ideas explored, but not exactly their execution. This issue arose particularly when there was a perceived mismatch between what emotions/emphasis the captions were denoting and what participants were seeing from facial expressions in the video. Eliah: *'The woman on the video was showing distinct facial expressions, there wasn't much change in the border of the first design* [Figure 2a], *but then later on when she wasn't showing much the border became pink or blue.'*

The imprecise alignment between words and sounds did not go unnoticed. Irene: *'I would see the speaker take a breath but there was no break in the captions.'* The display of loudness also seemed misaligned. Ira: *'I liked the idea in the second design of bolding some words for emphasis, but it didn't seem to match the sentences.'*

Legibility was a major concern, with six out of the eight interviewees having mentioned it. Some of this could be related to the colors used: Erin: *'you get tired of reading, and then the colors start to change, it is confusing to try to understand the tone.'* She also mentioned having some degree of colorblindness, which made matters worse. The fonts used were also a source of concern. Ira: *'It was too busy. The font and color changes made it hard to read and look at the person's emotion.'*

Some participants did not notice the border changes in the first design, and some that did found it distracting. Erin: *'The border*

was awful. Its constant motion would give me a headache.' Alex: *'Zoom or Microsoft Teams already have a border around whoever is talking, so if you add an additional one tied to the captions it'll be extremely distracting.'* For Eliah, inversely, the border, which reminded him of a similar device used in the video-game 'The Sims,' was functional precisely because it didn't get in the way: *'I liked how the color change represented mood while staying out of the view of the captions.'*

Reactions to the typographic designs (designs two and three) were mixed. Agatha: *'I wish the third design* [Figure 2c] *had an easier font to read but I enjoyed the changing fonts because it helped to show the emotions.'* Ada: *'The best thing about the second design* [Figure 2b] *was maybe the change in font thickness, whether it's thin or thick to show the emphasis, I think that was helpful.'* Ira: *'Seeing the caption change color was interesting because it helped me separate one sentence from another, while also helping me understand how the person is saying specific phrases.'*

3.5 Discussion of Study 1

The first goal of the study was to find out in what ways DHH individuals experience the absence of prosodic and emotional depictions in computer-generated captions, as are used in meetings with hearing peers (RQ1.A). Our interviewees discussed the many dimensions in which speech accessibility solutions can fail them. Captions, in particular, have many shortcomings. Some come from known limits of current automatic speech recognition systems, which negatively affect DHH individuals' experience of captioning systems [42, 52], and include high-latency and difficulty with non-'standard' accents [1, 27].

Beyond these failings, however, we found that captions' depictions of words are felt as if lacking something, leaving out meaningful dimensions of speech. These elements are present acoustically, so their absence creates barriers for DHH individuals. Missing a shift in tone from a serious to a humorous conversation, for instance, was a frequent complaint — and an expected one, given that humor has prosodic markings [5].

Our interviewees deal with these challenges in a myriad of ways, but the strategies employed are not perfect. Reading and interpreting text perceived as dull has an additional cognitive toll, and is commonly thought of as *boring*. This finding agrees with studies that show that emotional stimuli draw more attention and are better remembered than neutral counterparts [45], an effect that extends to written text [41].

All of these issues leave interviewees feeling as if not part of the group when participating in meetings with their hearing peers. This is such a common occurrence that some have naturalized it as being an inherent aspect of such meetings, rather than a consequence of how their underlying technologies have been designed.

Our second goal was to understand how these strategies and experiences could inform the design of new captioning systems that depict prosody and/or emotions (RQ1.B). While participants agreed that including *some* non-textual dimensions of speech could help alleviate the ambiguities of ASR-generated captions, they diverged as to which of these dimensions would be most helpful: either emotional cues, prosodic cues, or both. A follow-up study investigating *how* these captions could look like could thus face a design space too vast to explore. A plausible alternative, then, was to first evaluate *what* non-textual dimensions are most effective to alleviate the communication issues that emerged from the interviews of Study 1, thus allowing future studies a narrower research scope while still measuring whether these expanded captions can help DHH individuals identify paralinguistic dimensions in speech. While a 'good enough' design style for the captions may be sufficient for the purposes of this 'what dimensions' study, its parameters must still be carefully considered. See subsection 4.1.2 for a detailed discussion of our approach to tackling this issue in Study 2.

In discussing the prototypes shown, responses reflected a diverse set of preferences, allowing some high-level recommendations: (a) Legibility is a notable concern, even when participants felt prosody and emotions were being well represented; (b) Even though participants will generally complement their understanding of captions with a multimodal apprehension of other signals, such as facial expressions and body language, peripheral visual elements used for representing prosody or emotions run the risk of being ignored.

4 STUDY 2: COMPARATIVE STUDY

To answer RQ2.A, Study 2 exposed participants to four different types of captioning systems, each designed as the representation of a different set of prosodic or emotional features, thus gauging how each approach changed users' understanding of what was being said. Additionally, to answer RQ2.B we measured participants' opinions about the ease of reading and appropriateness of each caption type for use in different settings.

4.1 Design of Study 2

The test was online and self-administered. An introduction was done between one of the researchers and the participant over email and/or teleconferencing, after which the link to the test was shared. On this website, there was an introduction about the goals and workings of the study, with examples and explanations for the four types of captions (for details, see subsection 4.1.1 onward).

Following a demographic questionnaire, eight videos were presented on separate pages, with no sound, and captioned in one of the four available styles. While Python scripts pre-processed the speech files to extract affective and acoustic cues, a Javascript pipeline running on a web browser handled the styling and animating of the captions. HTML video provides a series of native events fired at key moments of each line of text's life-cycle (*cueEnter, cueExit,* etc), and it was through overloading these events that we were able to customize each caption style. Although we found that even mid-end machines, such as a 2014 Macbook Pro, were able to render the captions in real-time and virtually flicker-free, we felt it safer to present them 'burned-in' (i.e., as open-captions) in the videos to account for participants' unpredictable computer settings.

The videos had an average duration of 50 seconds (σ = 15 s). As with Study 1, they had someone telling a personal story with strong emotional overtones. To counter how each story could bias participants' preferences, each video was generated in all four caption styles. While each participant saw the same eight videos, their order and caption style were randomized.

After watching each video, and immediately below it, participants indicated their agreement with the following statements

on a 7-point Likert scale: (1) I found the speaker's emotions and moods easy to identify; (2) I could easily tell which words were emphasized; (3) I would be interested in using this captioning style for *work meetings* in software such as Zoom, Google Meet, etc; (4) I would be interested in using this captioning style for *personal meetings*; (5) I found the captions easy to read.

After the eight videos, participants were asked a set of open-ended questions, which were chosen according to that participant's specific answers to the Likert-scale items previously, to add nuance to our understanding of those answers. For instance, if for a given video the participant gave a below mid-point rating to the scale *'I found the speaker's emotions and moods easy to identify,'* an image of that same video would resurface at this last step with the prompt: *'Could you elaborate on why you felt that the caption design shown was not helpful to understand the speaker's emotions and moods?'*

4.1.1 Extracting prosodic and emotional features. The four different sets of features represented were: (1) No prosodic or emotional features (C)—basically, a conventional captioning system depicting only words in a neutral fashion—shown in Figure 3a; (2) Only prosody (P), shown in Figure 4; (3) Only emotions (E), shown in Figure 3c; and (4) a combination of Prosody and Emotions (P+E), shown in Figure 3d. We describe how data for (2), (3), and (4) was acquired below, with the visual design for the representations discussed in subsection 4.1.2.

As in Study 1, we used videos from the Stanford Emotional Narratives Dataset [59]. Beyond the videos, the dataset includes a transcription of their speech, defined in 5-second blocks of words. To extract prosody and emotions, we needed timestamps of each word individually. To obtain these, we fed the transcriptions into a local instance of Gentle [58], a Kaldi-based force-alignment toolkit set at a word-based granularity level.

Extracting prosody from speech. We followed the extraction and processing procedures outlined in Pataca and Costa [60]. Loudness and pitch were extracted in the Praat software [14], patched through Python using the praatIO interface [49]. The algorithms used were root mean square for loudness and an auto-correlation for pitch, applied to the segmented words obtained through the forced alignment, from which we also determined word durations. We applied a rolling average with a 7-segment window size to loudness and pitch, whose values were then normalized to a range of 0 to 1 using the mid-point between local (15 segments) and global normalizations [60].

We had to modify how durations were normalized because Pataca and Costa [60]'s algorithm aimed at Brazilian Portuguese, and our experiment would be in U.S. English. In syllable-timed languages, as Brazilian Portuguese is in certain conditions [53], one can expect the duration of each syllable to be roughly similar, giving it a machine-gun-like rhythm. When there *are* changes in this average duration, one can assume a meaningful change in prosodic rhythm. U.S. English, on the other hand, is stress-timed, i.e., the regularity is in the rhythm between stressed syllables, giving the language a morse-code-like rhythm [57]. Here, syllables will naturally have different durations, so the same syllable-duration metric would not be as effective as a marker for prosodic rhythm.

To work around this, we used a text-to-speech synthesizer[5] to sound each word in the transcription. This synthetic word's duration was then used as a normalizing denominator for the duration of the actual spoken word, i.e., how faster or slower the actual spoken word was when compared to this 'neutral' synthetic counterpart. This new metric was thus used as a correlate of prosodic rhythm.

Extracting emotions from speech. We used the circumplex model of emotion [68], which decomposes emotions as the product of variables defining two dimensions: valence, set in the pleasure-displeasure axis, and arousal, set in the arousal-sleep axis. For example, excitement, in this model, is a high-pleasure, high-arousal emotion; depression is a low-pleasure, low-arousal one.

We ran the segmented audio recordings through a transformer-based neural network [80], which outputted values for valence and arousal. Notwithstanding its state-of-the-art accuracy, we chose this particular model because of two key characteristics: (1) it can generalize better across domains, i.e., even if not trained to our specific dataset[6] it is expected to lose less accuracy than alternative architectures; and (2) its accuracy suffers little loss between speakers of different genders, as were present in the dataset used.

The model was trained on sequences between 3 to 10 seconds long, so it would not be able to extract meaningful values from the audio sliced in too-short word-sized chunks. Based on a suggestion by one of the authors [74], we padded each word with its surrounding audio at a 3-second margin on each side.

4.1.2 Caption design. There were two main constraints to the visual choices we explored when designing the speech-modulated typography for the captions. First, we focused on purely typographic designs, which we saw in Study 1 were less distracting than independent interface elements. While this is not an empirically validated point, we use it here as a simplifying assumption to narrow our focus into a more manageable, albeit still vast, subset of visual approaches.

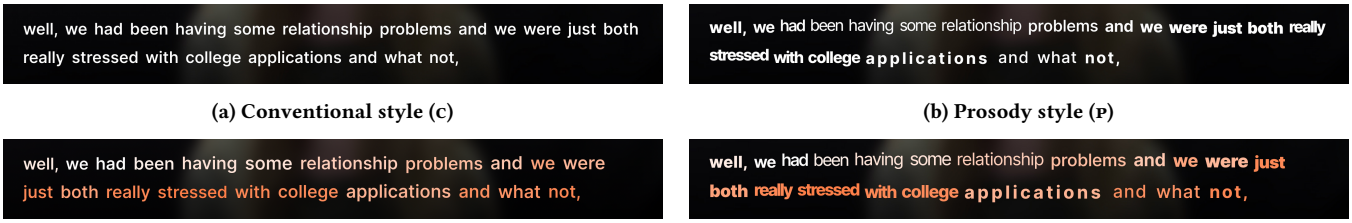
Second, we limited ourselves to typographic parameters that are freely combinable. Because we had a caption style where *both* prosody and emotions are shown, visual solutions applied to prosody and emotion must allow for a third, combined setting where all five different input dimensions (three for prosody and two for emotions) are applied simultaneously.

Depicting prosody: loudness, pitch, and duration. Unlike with emotions, there are quite a few competing models that represent prosody through typography [11, 18, 23, 60, 66, 70, 83]. We followed Pataca and Costa [60]'s model because of its flexibility in allowing the modulation of extra typographic parameters, as needed for depicting emotion.

Figure 4 shows an example of how we mapped loudness to font weight (thin to thick equating quiet to loud words), pitch to baseline shift (negative to positive vertical displacements equating low to high pitched words), and duration to letter-spacing (tight to spread out letters equating fast to slow sounding words).

[5] Using the Python library pyttsx3 [12], we created an instance of macOS' native speech synthesizer, NSSpeechSynthesizer, set with the default voice.

[6] Interestingly, while the only inputs used were the audio files, it has been shown that this particular architecture is implicitly able to derive affective information from linguistic elements present in speech, helping it beat the performance of explicitly multimodal neural networks [75]



(a) Conventional style (C)

(b) Prosody style (P)

(c) Emotion style (E)

(d) Prosody & emotion style (P+E)

Figure 3: The four caption styles used in the test. The C style has words with no additional styling (3a); The P style maps loudness to font-weight, pitch to baseline shift, and duration to letter-spacing (3b); The E style maps valence to color, with red meaning negative, white neutral, and green (not shown) positive, and arousal to font-size (3c); finally, the P+E style combines the five modulations from both the P and E styles (3d).

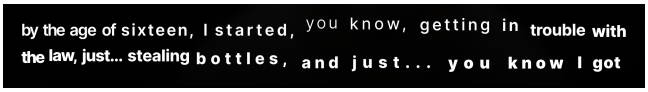


Figure 4: Example of the three prosodic features mapped as modulations of three typographic parameters.

Depicting emotions: valence and arousal. Challenges related to the ambiguity of captioned speech were a leading theme in Study 1. For Study 2, however, our design goal was not to create a model of captions that would remove all ambiguity in speech but, instead, one that would provide users with tools to better deal with it. Thus, while previous authors have worked with categorical models that define explicit emotions [34, 44, 63], our premise led us to work with a dimensional model.

While categorical emotions can map to the circumplex plane, these mappings may denote distinct 'categorical' emotions depending on the context (e.g., fear and anger are both emotions with negative valence and high arousal) [65]. In fact, in Russell's original paper many similar models are shown to have existed, all giving slightly different, albeit related, meanings to the two axes [68].

A design goal of embracing ambiguity is based on the assertion that to understand emotions in speech one must consider that meaning is not *only* found in an acoustic signal but also in how this signal is grounded in a particular socio-cultural context [13]. Leaving room for ambiguity can be an asset, i.e., a recognition that one's interpretation of something can vary depending on the subjects involved and the context Gaver et al. [29]. By embracing how the ambiguous nature of emotions gives form to open-ended visual representations, we align ourselves with Höök et al.'s design principle of non-reductionism, and Boehner et al.'s call for systems that support interpretive flexibility [13, 33].

All of this, however, does not equate to leaving individuals confused. As such, our design choices aim to be intuitive representations of both valence and arousal. The literature points to two common approaches: either using animation effects, especially if mimicking the bodily expressions of emotions [28, 50, 64], or directly manipulating type shapes, be it programmatically [48] or directly in their typographic designs [61].

Word animations would not work: they typically fragment the line of text, a troublesome characteristic considering that automatic speech recognition systems employ scrolling captions, i.e., the position of words is already continuously shifting as new lines appear, creating an unpredictable compounding motion with the affective animations. The embedding of valence in the type shape approach seems better suited for use in captions, particularly Promphan's *Emotional type* [61],[7] but falls short of our second constraint in designing captions for this study, i.e., the typographic parameters must be freely combinable. These two approaches, then, while competent in their representation of valence, could not be used.

The chosen typographic parameters were color, for valence, and font size, for arousal. We based our choice for color on ample evidence of how it can be used to represent moods and/or emotions [7, 20, 21, 31, 41, 44, 73]. While the use of red to represent negative valence seems a relatively straightforward choice [31, 41], we saw evidence for the use both of blue [73] and green [41] to represent positive valence. In Study 1 we pilot-tested a red-to-white-to-blue scale color scale to represent, respectively, negative to neutral to positive valence, but negative feedback from participants led us to settle on a red-to-white-to-green scale for the second study. While this color scheme is hard to distinguish for individuals with severe types of red-green color vision deficiency (protanopia and deuteranopia), we tinged the red with some yellow and the green with some blue to make them more discernible to individuals with the more common, milder deuteranomaly and protanomaly [56]. Figure 6 shows a simulation of how the palette weathers under different types of color vision deficiency[8], and Figure 5 shows an example of this color scheme applied to captions.

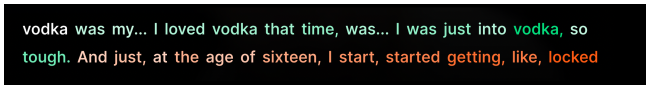


Figure 5: Font-color representing valence.

To the best of our knowledge, there are no direct examples of the representation of arousal by the modulation of typographic parameters. We opted to use font size because it has seen use as a

[7]The author kindly provided us with a revised version of the typeface published on her thesis. It was the basis for one of the prototypes shown to participants in Study 1 (see Figure 2c).

[8]Images created in the Coblis Color Blindness Simulator [81]

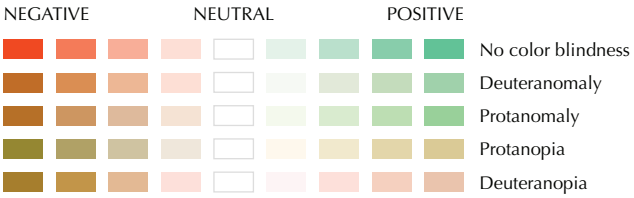


Figure 6: The valence color palette under simulation of various types of color vision deficiency.

representation for both changes in pitch [18] and loudness [66], features which have been associated with emotions of high (joy, anger) or low (sadness) arousal [24, 39]. An example of this modulation can be seen in Figure 7.

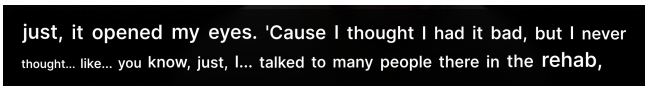


Figure 7: Font size being used to represent arousal.

4.2 Participants and recruitment

The IRB-approved study ran from July to August 2022. $40 compensation was offered. Recruitment was done through DHH specific social-media channels and mailing lists, screened for by identification as a DHH individual and not having participated in Study 1. A total of 16 individuals took the test. Eight identified as male and eight as female. Nine identified as being deaf/Deaf, and seven as hard-of-hearing. Their average age was 26 (σ = 5). Asked about how comfortable they were with reading and writing English, participants' median answer (on a Likert scale going from 1 to 7) was 7. Participants completed the test on average in 32 minutes (σ = 14).

4.3 Findings of Study 2

We conducted statistical significance testing on responses using a Kruskal-Wallis test, which was significant for all distributions. We then ran a post hoc Mann-Whitney U test between each caption type, with p-values adjusted using Holm-Šídák corrections. Figure 8 shows the distribution of answers, which comparisons were significant, and the median score for each scale.

4.3.1 Results of analysis. Median responses for the *clarity of emotions and moods* Likert scale for the C, P, P+E, and E styles were, respectively, 4, 4.5, 6, and 6, with statistically significant differences between the C and E styles (U = 295.0, p < 0.05[9]), P and E styles (U = 252.0, p < 0.01), and P+E and P styles (U = 734.0, p < 0.05).

Median responses for the *clarity of emphasis* Likert scale for the C, P, P+E, and E styles were, respectively, 3, 5, 5.5, and 5, with statistically significant differences between the C and P+E styles (U = 302.5, p < 0.05), and C and E styles (U = 269.0, p < 0.01).

Median responses for the *use in work meetings* Likert scale for the C, P, P+E, and E styles were, respectively, 6, 3, 3.5, and 5, with statistically significant differences between the C and P styles (U = 801.0, p < 0.01), C and P+E styles (U = 748.0, p < 0.01), and P and E styles (U = 287.0, p < 0.01).

Median responses for the *use in personal meetings* Likert scale for the C, P, P+E, and E styles were, respectively, 6, 3, 4.5, and 5, with statistically significant differences between the C and P styles (U = 790.5, p < 0.01), C and P+E styles (U = 735.5, p < 0.05), and P and E styles (U = 291.5, p < 0.05).

Median responses for the *legibility* Likert scale for the C, P, P+E, and E styles were, respectively, 7, 4.5, 5, and 6, with statistically significant differences between the C and P styles (U = 838.0, p < 0.01), C and P+E styles (U = 806.5, p < 0.01), C and E styles (U = 740.0, p < 0.01), and P and E styles (U = 311.5, p < 0.05).

In summary, captions that had an emotional component (P+E and E) significantly outperformed conventional captions in how they were able to help participants identify emphasis, but only the E style showed a significant improvement on how easy it made *emotions* and *moods* to identify. Traditional captions were perceived as more legible than all three other styles, including E, which outperformed P. Participants were less interested in using either the P+E or P styles than traditional captions for workplace or personal meetings — for E captions, the preference is smaller, but non-significantly so.

4.3.2 Open-ended comments. After watching the videos, participants were asked questions about what worked or didn't in the caption styles, with specific questions being chosen according to their most negative and positive answers to the Likert-type scales.

Regarding legibility, there were comments about how the new styles (P, P+E, and E) were harder to read than the more traditional C style. Prosodic representation, in particular, was condemned, with specific mention of how its use of baseline shift and changes in the spaces between letters made words 'wild' and hard to read. Some participants also had the impression that there was sometimes too much going on, making captions confusing, slower to read, or even headache-inducing.

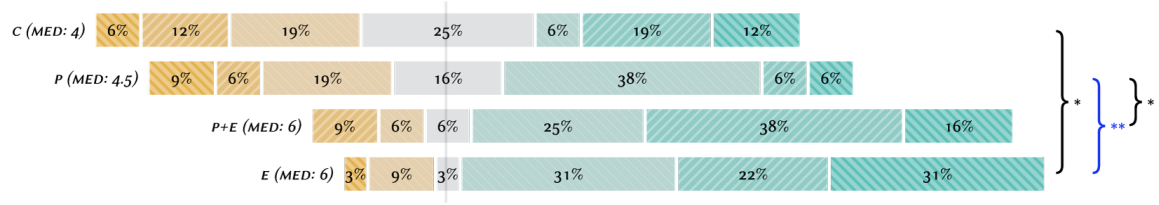
Speaking specifically to the typographic parameters modulated in the three new styles, some participants commented on how font weight and color worked effectively to represent a speaker's emotions and moods. Changes to font size were also positively cited as expressive modulations, with the caveat that at times they made captions too small to read comfortably. Lastly, and more rarely, a few participants felt changes in baseline shift and letter spacing negatively impacted legibility.

In terms of function, and particularly regarding styles P+E and E, the new captions received praise. They were said to work in terms of making the speakers' emotions and speech clearer. One participant imagined this reducing misunderstanding their friends. Another said that, while they are typically able to derive sufficient emotional understanding from facial expressions, the E-styled captions would be 'awesome' when the speaker's face is occluded. Discussing style E, one participant said it was easier to tell when the mood shifted between positive or negative feelings, which another participant said changed how they understood the stories in the videos.

More broadly, some comments pointed out that, even if the E style did not necessarily change their understanding, it was less bland than traditional captions. Similarly, P+E was felt as being more engaging and easier to follow along. Some participants claimed that

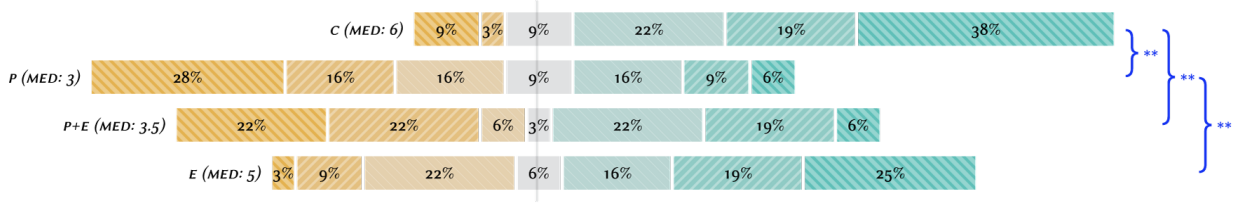
[9]P-values presented adjusted using Holm-Šídák corrections. Always two-tailed tests with n_1 and n_2 equal to 32.

(a) Key to the responses to the Likert-type scales.

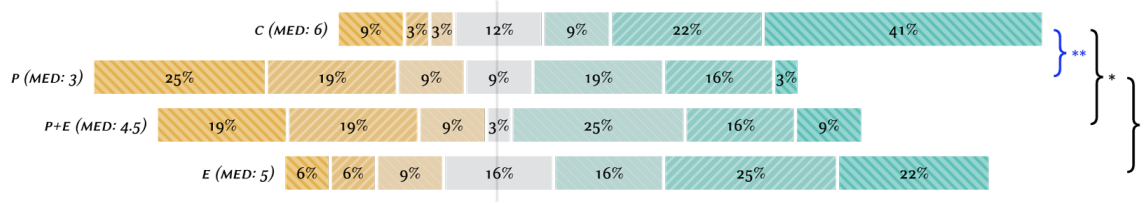


(b) Answers to the Likert scale presented immediately below each video with the following statement: *I found the speaker's emotions and moods easy to identify.*

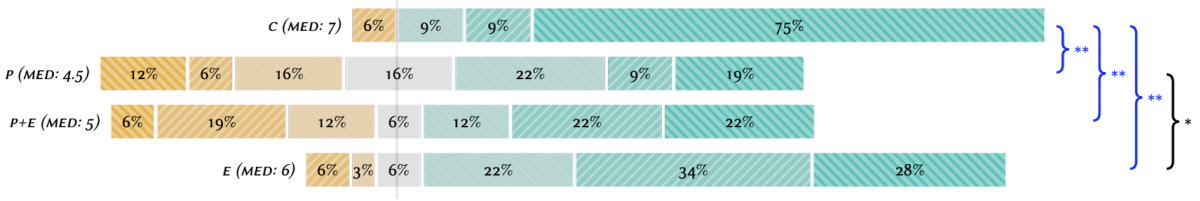
(c) As above, but for: *I could easily tell which words were emphasized.*



(d) As above, but for: *I would be interested in using this captioning style for work meetings in software such as Zoom, Google Meet, etc.*



(e) As above, but for: *I would be interested in using this captioning style for personal meetings in software such as Zoom, Google Meet, etc.*



(f) As above, but for: *I found the captions easy to read.*

Figure 8: Responses to the five Likert-type scales. Each row represents responses considering one out of the four caption styles. Caption styles are abbreviated as follows: Conventional (C), Only prosody (P), Prosody and emotions (P+E), and Only emotions (E). A blue ** marks a $p < 0.01$ significant Mann-Whitney U test between medians, while a black * marks a $p < 0.05$ comparison.

while they personally did not see a need for these new styles, they felt other DHH individuals could benefit from them.

4.4 Discussion of Study 2

Results revealed that the Emotion (E) caption style significantly outperformed conventional captions in participants' perceptions of how they expressed emotions and emphasis, with the Prosody + Emotion (P+E) variant also scoring higher in its depiction of emphasized words. This is an encouraging suggestion that the E style could help make accessible these important paralinguistic dimensions of speech.

Surprisingly, the Prosodic (P) style did not outperform traditional captions in representing emphasis. Given that it was based on a model that had been shown to successfully depict speech prosody with hearing individuals, it raises the question of how differently hearing and DHH individuals will interpret these captions.

We saw that both depictions with prosodic components had relatively low legibility scores, with two of its three typographic parameters (baseline shift and letter-space) being specifically denounced by some participants as culprits. Given that it also had low interest in use for personal or work meetings, the E style's performance for emotion and emphasis representation, coupled with its higher legibility and appeal, gives an unexpected but interesting answer to our RQ2.A and RQ2.B:

In seeking what dimensions of paralinguistic properties could represent a speaker's emotions and/or emphasis in captioned speech (RQ2.A), it seemed plausible to expect that the P style would score higher at representing emphasis and the E style at emotions. We found, however, that a choice between showing either emotions or emphasis, as came out of Study 1, may be unnecessary, with the E style capable of capturing and representing both dimensions.

As for RQ2.B,[10] while the E style did not outperform traditional captions in perceived suitability for work or personal meetings, it ranked significantly better than the P style for both settings. The assumption that there would be a divide between what was favored for work versus personal meetings did not pan out, with each style's preference score relatively consistent between the two settings (this effect could, however, be an artifact of how the videos we used skewed towards content one would expect to see in a personal conversation instead of a professional one).

5 LIMITATIONS AND FUTURE WORK

Study 1 showed that further work was needed both towards a better understanding of *what* dimensions of speech should be visualized and *how* to design those visualizations. With Study 2, we investigated the first path, but both are intertwined, i.e., to test *what* to represent, we needed to design the options somehow, and inversely, if we were to evaluate different caption designs, they would have to depict some model of these dimensions. As such, while we based our design choices on fair assumptions from research related to ours, given that the field is still sparse, further work is needed to investigate *how* to represent these dimensions systematically.

This design process needs to consider the perspective of DHH individuals, taking into account two considerations that emerged from Study 1 and 2: First, the color schemes used may leave individuals with color vision deficiency unable to distinguish between negative and positive valence words, as is seen in the last two rows of Figure 6. Future research should explore alternative color schemes and typographic modulations to make the style more accessible.

Second, legibility was a recurrent concern for participants from both Study 1 and 2. We used a Likert-rating scale to weed out acute issues with ease of reading with any of the three proposed caption styles. This was, however, a somewhat blunt instrument, ignoring aspects such as gaze time (which is already high for DHH individuals for conventional captions [36]), reading speed, cognitive workload, etc. A caption design's readability is related to users' demographics, personal preferences, and use cases, so a *one-size fits all* solution is probably not an ideal approach here [8]. Still, as these designs mature, further research should investigate their reading performance more thoroughly. This may include exploration of different granularity levels for the measurement and display of the non-textual dimensions of speech — while, like Rosenberger and MacNeil [66], we employed a word-level measure, this could be finer, e.g., syllabic [60, 66] or phoneme-grapheme mapping [83], or coarser.

From the perspective of the speaker, having an autonomous system that *proactively* codifies and depicts their speech based on an automatic analysis of their emotions carries the risk of a loss of autonomy, as described in Höök et al. [33]. This could be a sensitive issue and, as such, future studies should investigate how a user interface could represent and cede control to speakers about this emotion-sensing process, as it is ongoing.

We asked participants how clearly they could perceive emotions, moods, and emphasis in a captioned video. We did not measure, however, how helpful these represented dimensions were. Future studies could investigate whether the presence of these novel caption styles could alleviate ambiguity, especially considering the communication breakdown scenarios our interviewees described in Study 1: quick shifts in mood, inexpressive body language, and occluded faces, among others.

This last factor, of how a clear view of the speaker's face can influence how these captions are understood, might be an independent line of research of its own, e.g., [3], given that, in Study 2, videos showed speakers' faces clearly, a condition which plausibly could affect the interpretation of the captions themselves (e.g., similarly to how individuals with cochlear implants derive greater benefit from synchronous facial and speech channels than hearing individuals in emotion recognition tasks [79]).

While it was not a measured dimension, the fact that some participants of Study 2 commented about how the new captioning styles were more engaging is a compelling counterpoint to how some interviewees in Study 1 complained that traditional captions are *boring*. Considering that one of the comments specifically said that the new captions did not change their understanding of the story but were easier to follow along, we can envision that future studies could focus on measuring how immersive these captions are compared to past measures of traditional captions [40].

6 CONCLUSION

By examining the experience of DHH individuals with automatically captioned meetings, we found that, in many situations, they

[10] In what settings is the use of these visual depictions of prosody and/or emotion felt as the most appropriate from the point-of-view of DHH individuals?

feel left behind hearing peers and unable to participate fully in the conversations. We saw that the way captions typically leave out emotions and emphasis gives speech an ambiguity that can make it unapproachable and dull. Unlike factors such as latency or failures in word recognition, however, this is not perceived as if an imperfection of how captions are designed. As such, the challenges that arise in captioned meetings between DHH and hearing individuals are seen as inherent qualities of the medium, instead of objective—and addressable—shortcomings of how these accessibility tools work.

Next, we contrasted how three novel caption styles represented emotions and emphasis compared to traditional captions. We found that the best-performing option was based on the output of speech processed through a neural network that extracted emotional features from it. This approach had good legibility, albeit worse than conventional captions. Participants' willingness to use these captions for work or personal meetings is comparable to that of traditional captions.

Our work has investigated a rarely explored dimension of DHH experience with automatic captions, putting forward three novel approaches for modeling, processing, and depicting speech that may motivate the development of more inclusive captioning systems.

ACKNOWLEDGMENTS

This material is based upon work supported by the Fulbright Commission (Fulbright-CAPES Scholarship, ME / CAPES N^{o}8 / 2020) and the National Science Foundation under Grants N^{o} 1954284, 2125362, and 2235405.

REFERENCES

[1] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition?. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Association for Computational Linguistics, Online, 22–34. https://doi.org/10.18653/v1/2021.bppf-1.4

[2] Aviad Albert, Francesco Cangemi, and Martine Grice. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proceedings Speech Prosody*, Vol. 9. International Speech Communications Association, Poznan, Poland, 13–16.

[3] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Effect of Occlusion on Deaf and Hard of Hearing Users' Perception of Captioned Video Quality. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 202–220.

[4] Carl Armon, Dan Glisson, and Larry Goldberg. 1992. How Closed Captioning in the U.S. Today can Become the Advanced Television Captioning System of Tomorrow. *SMPTE Journal* 101, 7 (1992), 495–498. https://doi.org/10.5594/J02244

[5] Salvatore Attardo, Manuela Maria Wagner, and Eduardo Urios-Aparisi. 2011. Prosody and humor. *Pragmatics & Cognition* 19, 2 (2011), 189–201.

[6] Plínio A. Barbosa. 2019. *Prosódia.* Parábola Editorial, Brazil.

[7] Lyn Bartram, Abhisekh Patra, and Maureen Stone. 2017. Affective Color in Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1364–1374. https://doi.org/10.1145/3025453.3026041

[8] Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L. Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B. Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D. Sawyer, Shannon M. Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace, and Benjamin Wolfe. 2021. Readability Research: An Interdisciplinary Approach. *CoRR* abs/2107.09615 (2021), 85 pages. arXiv:2107.09615 https://arxiv.org/abs/2107.09615

[9] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312921

[10] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[11] Ann Bessemans, Maarten Renckens, Kevin Bormans, Erik Nuyts, and Kevin Larson. 2019. Visual prosody supports reading aloud expressively. *Visible Language* 53, 3 (2019), 28–49.

[12] Natesh M Bhat. 2021. Text-to-speech x-platform . https://pyttsx3.readthedocs.io/en/latest/

[13] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2005. Affect: From Information to Interaction. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1094562.1094570

[14] Paul Boersma. 2006. Praat: doing phonetics by computer. http://www.praat.org/. Accessed on August 24, 2022.

[15] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[16] Michele A. Buchanan. 2015. @ Face Value // Expanding Our Typographic Repertoire. *Communication Design* 3, 1 (Jan. 2015), 27–50. https://doi.org/10.1080/20557132.2015.1057373

[17] Doris C Caldwell. 1981. Closed-Captioned Television for Hearing-Impaired Viewers. *Media Information Australia* 19, 1 (Feb. 1981), 56–60. https://doi.org/10.1177/1329878X8101900113

[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* (Braga, Portugal) *(ARTECH 2019)*. Association for Computing Machinery, New York, NY, USA, Article 13, 8 pages. https://doi.org/10.1145/3359852.3359892

[19] Anna C. Cavender, Jeffrey P. Bigham, and Richard E. Ladner. 2009. ClassInFocus: Enabling Improved Visual Attention Strategies for Deaf and Hard of Hearing Students. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) *(Assets '09)*. Association for Computing Machinery, New York, NY, USA, 67–74. https://doi.org/10.1145/1639642.1639656

[20] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2021. Bubble Coloring to Visualize the Speech Emotion. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 361, 6 pages. https://doi.org/10.1145/3411763.3451698

[21] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2022. Designing voice-aware text in voice media with background color and typography. *Journal of the International Colour Association* 28 (2022), 56–62.

[22] Wellington da Silva, Plinio Almeida Barbosa, and Åsa Abelin. 2016. Cross-Cultural and Cross-Linguistic Perception of Authentic Emotions through Speech: An Acoustic-Phonetic Study with Brazilian and Swedish Listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 32, 2 (Aug. 2016), 449–480. https://doi.org/10.1590/0102-445003263701432483

[23] Caluã de Lacerda Pataca and Paula Dornhofer Paro Costa. 2020. Speech Modulated Typography: Towards an Affective Representation Model. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 139–143. https://doi.org/10.1145/3377325.3377526

[24] João Antônio de Moraes and Albert Rilliard. 2016. Prosody and Emotion in Brazilian Portuguese. In *Issues in Hispanic and Lusophone Linguistics*, Meghan E. Armstrong, Nicholas Henriksen, and Maria del Mar Vanrell (Eds.). Vol. 6. John Benjamins Publishing Company, Amsterdam, 135–152. https://doi.org/10.1075/ihll.6.07mor

[25] Jorge dos Reis and Valerie Hazan. 2012. Speechant: a vowel notation system to teach English pronunciation. *ELT journal* 66, 2 (2012), 156–165.

[26] Lisa Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. 2016. Deaf and Hard of Hearing Individuals' Perceptions of Communication with Hearing Colleagues in Small Groups. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 271–272. https://doi.org/10.1145/2982142.2982198

[27] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition. https://doi.org/10.48550/ARXIV.2103.15122

[28] Jodi Forlizzi, Johnny Lee, and Scott Hudson. 2003. The Kinedit System: Affective Messages Using Dynamic Texts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*.

Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/642611.642677

[29] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a Resource for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/642611.642653

[30] Morton Ann Gernsbacher. 2015. Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences* 2, 1 (Oct. 2015), 195–202. https://doi.org/10.1177/2372732215602130

[31] Sandrine Gil and Ludovic Le Bigot. 2016. Colour and emotion: children also associate red with negative valence. *Developmental science* 19, 6 (2016), 1087–1094.

[32] Awni Hannun. 2021. The History of Speech Recognition to the Year 2030. https://doi.org/10.48550/ARXIV.2108.00084

[33] Kristina Höök, Anna Ståhl, Petra Sundström, and Jarmo Laaksolaahti. 2008. Interactional Empowerment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 647–656. https://doi.org/10.1145/1357054.1357157

[34] Jiaxiong Hu, Qianyao Xu, Limin Paul Fu, and Yingqing Xu. 2019. Emojilization: An Automated Method For Speech to Emoji-Labeled Text. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313071

[35] Gary Hustwit. 2015. *Helvetica/Objectified/Urbanized: The Complete Interviews: The Design Trilogy Interviews*. Versions Publishing, London, UK.

[36] Carl J Jensema, Ramalinga Sarma Danturthi, and Robert Burch. 2000. Time spent viewing captions on television programs. *American annals of the deaf* 145, 5 (2000), 464–468.

[37] Sushant Kafle and Matt Huenerfauth. 2019. Predicting the Understandability of Imperfect English Captions for People Who Are Deaf or Hard of Hearing. *ACM Trans. Access. Comput.* 12, 2, Article 7 (jun 2019), 32 pages. https://doi.org/10.1145/3325862

[38] Yeon Soo Kim, Sunok Lee, and Sangsu Lee. 2022. A Participatory Design Approach to Explore Design Directions for Enhancing Videoconferencing Experience for Non-Signing Deaf and Hard of Hearing Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 47, 4 pages. https://doi.org/10.1145/3517428.3550375

[39] Maria Kraxenberger, Winfried Menninghaus, Anna Roth, and Mathias Scharinger. 2018. Prosody-based sound-emotion associations in poetry. *Frontiers in Psychology* 9 (2018), 1284.

[40] Jan-Louis Kruger, María T. Soto-Sanfiel, Stephen Doherty, and Ronny Ibrahim. 2016. Towards a cognitive audiovisual translatology. In *Reembedding Translation Process Research*. John Benjamins Publishing Company, Amsterdam, 171–194. https://doi.org/10.1075/btl.128.09kru

[41] Christof Kuhbandner and Reinhard Pekrun. 2013. Joint effects of emotion and color on memory. *Emotion* 13, 3 (2013), 375.

[42] Raja S. Kushalnagar and Christian Vogler. 2020. Teleconference Accessibility and Guidelines for Deaf and Hard of Hearing Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/3373625.3417299

[43] Walter S. Lasecki, Raja Kushalnagar, and Jeffrey P. Bigham. 2014. Helping Students Keep up with Real-Time Captions by Pausing and Highlighting. In *Proceedings of the 11th Web for All Conference* (Seoul, Korea) *(W4A '14)*. Association for Computing Machinery, New York, NY, USA, Article 39, 8 pages. https://doi.org/10.1145/2596695.2596701

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

[45] Einat Liebenthal, David A Silbersweig, and Emily Stern. 2016. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Frontiers in neuroscience* 10 (2016), 506.

[46] Jason Livingston. 2012. Closed Captioning Challenges for IP Video Delivery. In *The 2012 Annual Technical Conference & Exhibition*. SMPTE, Hollywood, CA, USA, 1–9.

[47] Fernando Loizides, Sara Basson, Dimitri Kanevsky, Olga Prilepova, Sagar Savla, and Susanna Zaraysky. 2020. Breaking Boundaries with Live Transcribe: Expanding Use Cases Beyond Standard Captioning Scenarios. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Virtual Event Greece, 1–6. https://doi.org/10.1145/3373625.3417300

[48] Catarina Maçãs, David Palma, and Artur Rebelo. 2019. TypEm: A Generative Typeface That Represents the Emotion of the Text. In *Proceedings of the 9th International Conference on Digital and Interactive Arts* (Braga, Portugal) *(ARTECH 2019)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/3359852.3359874

[49] Tim Mahrt. 2022. PraatIO. https://github.com/timmahrt/praatIO. Accessed on August 3, 2022.

[50] Sabrina Malik, Jonathan Aitken, and Judith Kelly Waalen. 2009. Communicating emotion with animated text. *visual communication* 8, 4 (2009), 469–479.

[51] James R. Mallory, Michael Stinson, Lisa Elliot, and Donna Easton. 2017. Personal Perspectives on Using Automatic Speech Recognition to Facilitate Communication between Deaf Students and Hearing Customers. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Baltimore Maryland USA, 419–421. https://doi.org/10.1145/3132525.3134779

[52] Emma J. McDonnell, Ping Liu, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2021. Social, Environmental, and Technical: Factors at Play in the Current Use and Future Design of Small-Group Captioning. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–25. https://doi.org/10.1145/3479578

[53] Alexsandro R Meireles, João Paulo Tozetti, and Rogério R Borges. 2010. Speech rate and rhythmic variation in Brazilian Portuguese. In *Speech Prosody 2010-Fifth International Conference*. International Speech Communication Association (ISCA), Chicago, USA, 1–4.

[54] Chris Mikul. 2014. *Caption quality: Approaches to standards and measurement*. Media Access Australia, Sydney, Australia.

[55] Virginia Murphy-Berman and Linda Whobrey. 1983. The Impact of Captions On Hearing-Impaired Children's Affective Reactions To Television. *The Journal of Special Education* 17, 1 (April 1983), 47–62. https://doi.org/10.1177/002246698301700107

[56] National Eye Institute 2019. *Types of color blindness*. National Eye Institute. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/types-color-blindness Accessed on August 3, 2022.

[57] Marina Nespor, Mohinish Shukla, and Jacques Mehler. 2011. *Stress-Timed vs. Syllable-Timed Languages*. John Wiley & Sons, Ltd, Oxford, UK, Chapter 48, 1–13. https://doi.org/10.1002/9781444335262.wbctp0048 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444335262.wbctp0048

[58] Robert M Ochshorn and Max Hawkins. 2015. Gentle: a robust yet lenient forced aligner built on Kaldi. https://lowerquality.com/gentle/

[59] Desmond Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12 (2019), 579–594.

[60] Calua de Lacerda Pataca and Paula Dornhofer Paro Costa. 2022. Hidden bawls, whispers, and yelps: can text convey the sound of speech, beyond words. *IEEE Transactions on Affective Computing* pre-print (2022), 1–1. https://doi.org/10.1109/TAFFC.2022.3174721

[61] Suksumek Promphan. 2017. *Emotional Type: Emotional expression in text message*. Master's thesis. Basel School of Design, Switzerland.

[62] Dhevi J Rajendran, Andrew T Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives* 21, 1 (2013), 5–21.

[63] Raisa Rashid, Jonathan Aitken, and Deborah I Fels. 2006. Expressing emotions using animated text captions. In *International Conference on Computers for Handicapped Persons*. Springer, Linz, Austria, 24–31.

[64] Raisa Rashid, Quoc Vy, Richard Hunt, and Deborah I Fels. 2008. Dancing with words: Using animated text for captioning. *Intl. Journal of Human–Computer Interaction* 24, 5 (2008), 505–519.

[65] Nancy A Remington, Leandre R Fabrigar, and Penny S Visser. 2000. Reexamining the circumplex model of affect. *Journal of personality and social psychology* 79, 2 (2000), 286.

[66] Tara Rosenberger and Ronald L. MacNeil. 1999. Prosodic Font: Translating Speech into Graphics. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania) *(CHI EA '99)*. Association for Computing Machinery, New York, NY, USA, 252–253. https://doi.org/10.1145/632716.632872

[67] Jazz Rui Xia Ang, Ping Liu, Emma McDonnell, and Sarah Coppola. 2022. "In This Online Environment, We're Limited": Exploring Inclusive Video Conferencing Design for Signers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 609, 16 pages. https://doi.org/10.1145/3491102.3517488

[68] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[69] Vít Rusňák, Pavel Troubil, Desana Daxnerová, Pavel Kajaba, Matej Minárik, Svatoslav Ondra, Tomás Sklenák, and Eva Hladká. 2016. CoUnSiL: A video conferencing environment for interpretation of sign language in higher education. In *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, Istanbul, Turkey, 1–8. https://doi.org/10.1109/ITHET.2016.7760711

[70] Tim Schlippe, Shaimaa Alessai, Ghanimeh El-Taweel, Matthias Wölfel, and Wajdi Zaghouani. 2020. Visualizing voice characteristics with type design in closed captions for arabic. In *2020 International Conference on Cyberworlds (CW)*. IEEE, IEEE, Caen, France, 196–203.

[71] Matthew Seita, Khaled Albusays, Sushant Kafle, Michael Stinson, and Matt Huenerfauth. 2018. Behavioral Changes in Speakers Who Are Automatically Captioned

in Meetings with Deaf or Hard-of-Hearing Peers. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 68–80. https://doi.org/10.1145/3234695.3236355

[72] Matthew Seita and Matt Huenerfauth. 2020. Deaf Individuals' Views on Speaking Behaviors of Hearing Peers When Using an Automatic Captioning App. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383083

[73] Hyeon-Jeong Suk and Hans Irtel. 2010. Emotional response to color across media. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 35, 1 (2010), 64–77.

[74] Andreas Triantafyllopoulos. 2022. Personal communication.

[75] Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. 2022. Probing Speech Emotion Recognition Transformers for Linguistic Knowledge. https://doi.org/10.48550/ARXIV.2204.00400

[76] Máté Ákos Tündik, György Szaszák, Gábor Gosztolya, and András Beke. 2018. User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning. In *Proceedings of Interspeech 2018*. ISCA, Hyderabad, India, 2628–2632. https://doi.org/10.21437/Interspeech.2018-1352

[77] Walda Verbaenen. 2019. *Phonotype. The visual identity of a language according to its phonology.* Master's thesis. PXL-MAD.

[78] Christian Vogler, Paula Tucker, and Norman Williams. 2013. Mixed Local and Remote Participation in Teleconferences from a Deaf and Hard of Hearing Perspective. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 30, 5 pages. https://doi.org/10.1145/2513383.2517035

[79] Celina Isabelle von Eiff, Sascha Frühholz, Daniela Korth, Orlando Guntinas-Lichius, and Stefan Robert Schweinberger. 2022. Crossmodal Benefits to Vocal Emotion Perception in Cochlear Implant Users. *iScience* 25, 12 (2022), 105711.

[80] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2022. Dawn of the transformer era in speech emotion recognition: closing the valence gap. https://doi.org/10.48550/ARXIV.2203.07378

[81] Matthew Wickline. 2001. Coblis – Color blindness simulator. https://www.color-blindness.com/coblis-color-blindness-simulator/

[82] Deirdre Wilson and Tim Wharton. 2006. Relevance and Prosody. *Journal of Pragmatics* 38, 10 (Oct. 2006), 1559–1579. https://doi.org/10.1016/j.pragma.2005.04.012

[83] Matthias Wölfel, Tim Schlippe, and Angelo Stitz. 2015. Voice driven type design. In *2015 international conference on speech technology and human-computer dialogue (SpeD)*. IEEE, IEEE, Bucharest, Romania, 1–9.