Simplified Audio Production in Asynchronous Voice-Based Discussions

Venkatesh Sivaraman
Bexley High School
Columbus, OH
venkats@mit.edu

Dongwook Yoon
Cornell University
Ithaca, NY
dy252@cornell.edu

Piotr Mitros
edX
Cambridge, MA
pmitros@edx.org

ABSTRACT

Voice communication adds nuance and expressivity to virtual discussions, but its one-shot nature tends to discourage collaborators from utilizing it. Text-based interfaces have made voice editing much easier, especially with recent advancements enabling live, time-aligned speech transcription. We introduce SimpleSpeech, an easy-to-use platform for asynchronous audio communication (AAC) with lightweight tools for deleting and inserting content, adjusting pauses, and correcting transcript errors. Qualitative and quantitative results suggest that novice audio producers, including high school students, experience decreased mental workload when using SimpleSpeech to produce audio messages than without editing. We also found that the linguistic formality of SimpleSpeech is between that of traditional oral and written media. Our findings on formality and workload characteristics of editable speech input help define best practices and use cases for the design and application of such systems.

Author Keywords

Speech editing; transcription-based editing; asynchronous audio communication.

ACM Classification Keywords

H.5.2. User Interface: Voice I/O

INTRODUCTION

Asynchronous audio communication (AAC) is rapidly becoming available to mass audiences through social platforms such as WhatsApp, iMessage, and Facebook. While text is still by far the most prevalent (and often most convenient) mode of communication on the Internet, audio is desirable in many situations because of its ability to convey more expression and nuance than text. For instance, Mayer's work on multimedia learning [**?**] indicates that audio communication is particularly useful for managing cognitive load where material is primarily conveyed visually (e.g., by pointing), and in situations where an emotional connection between speaker and listener is desirable. AAC therefore holds considerable potential for improving online education, where voice communication has been demonstrated to improve student-student and student-instructor engagement as well as a sense of the instructor's social presence [15, 21, 27]. Our research is primarily motivated by the implications of AAC as a peer-to-peer educational tool, though it is also applicable to other collaborative settings, such as instructor authoring.

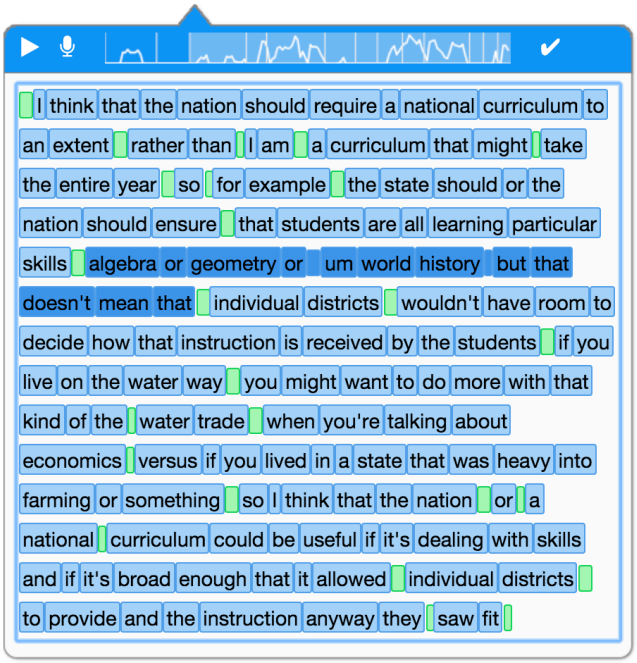

Figure 1. The user interface for SimpleSpeech presents an automatically-generated transcription of a voice comment, which can then be manipulated through normal word-processing operations, such as selection, deletion, and insertion. Word and pause tokens are mapped with audio via time alignment data.

The problem with replacing textual communication with speech, however, is that speakers may face difficulty articulating their ideas vocally. For instance, Marriott and Hiscock's study using Wimba voice boards for discussion forums found that students overwhelmingly preferred text over speech comments, in part because it required them to speak fluently without making errors [20]. Since this problem affects students even in physical classrooms, it could certainly prevent some learners from participating in online oral discussions. AAC platforms in such situations, then, must somehow compensate

CHI'16, May 7-12, 2016, San Jose, California, USA.

ACM ISBN 978-1-4503-3362-7/16/05...$15.00.
DOI string from ACM form confirmation

for the linearity and immutability of audio on the production side.

Our solution is to provide lightweight, easy-to-use editing tools based on automatic speech recognition (ASR)-generated transcripts. Many prior studies have utilized transcription to assist in audio editing [6, 22, 31], but only recently has fast, live editing become possible through advances in ASR technology [7, 23]. We implemented an easy-to-use audio production tool, SimpleSpeech, that allows users to delete and insert segments of the recording in real-time, simplifying quick word-level editing even when transcription errors are present.

Qualitative evidence indicates that SimpleSpeech's simplified interface gave users enough control over the editing process and enabled them to produce more polished audio comments, but that these benefits came at the price of elevated mental workload and formality. Further quantitative investigation on these themes showed that the workload of recording voice messages was in fact significantly less with editing functionality, demonstrating that SimpleSpeech would be a valuable enhancement to online audio communication platforms. Finally, linguistic characteristics of messages created using AAC are also discussed in comparison to other forms of communication, leading to new considerations and insights on optimal applications of this technology.

RELATED WORK

The linear, sequential nature of voice communication not only impedes skimming and navigation capabilities [10], but can also hamper the speech *production* process. Voice production is a temporarily linear process which demands that the speaker thinks and speaks simultaneously [19, 34]. Therefore, cognitive load arises from the fact that one has to keep speaking to prevent undesirable long pauses. In addition, mistakes in recorded speech are harder to revise than textual typos, mainly due to the lack of lightweight voice editing software [19]. Building on the qualitative implications of these previous works, our study presents a quantitative analysis of these burdens when the voice production system includes lightweight editing features.

Because lower-level audio waveform editing is an onerous task, speech manipulation tools have been developed that present audio in semantically meaningful higher-level chunks, such as words and phrases. Acoustic detection of the presence of speech provides binary visual guidance through which users can edit or index the speech recording [1, 14]; on the other hand, a pure acoustic approach has limited recognition granularity. Time-aligned automatic speech recognition (ASR) has now become a popular tool to achieve the word-level structuring of speech [24, 32]. Compared with acoustic structuring, ASR presents semantic information at higher resolution, but has also suffered from high computational load and delay. However, recent technical developments have made ASR faster and more accurate, and we take full benefit of this real-time transcription capability.

Since speech transcription elicits the contents of the recording, researchers have utilized it to assist in visual skimming and navigation. MedSpeak [18] and SCANMail [8] are well recognized as precursors of such systems that use time-alignment data of the transcript for indexing audio. Since transcription errors tend to obstruct visual comprehension, Vemuri et al. suggested a novel visualization of the transcript that adjusts transcription brightness to the word's ASR confidence score [28].

On the production side, there have been several systems that use a time-aligned transcript for editing audio [22, 31, 33] or video [3, 6]. Among them, Whittaker and Rubin's editing system leveraged users' familiarity with text-editing interfaces to adopt audio editing within that paradigm. Since we targeted non-professional users, our interface also took the text-like approach, but emphasizing a *live* production process and going beyond editing previously-transcribed speech.

Just as with listeners, ASR errors can be detrimental for understanding and skimming audio contents during editing [11]. In the MedSpeak interface, Lai et al. provided a separate graphical window for fixing transcription errors [18]. In a speech production system like SimpleSpeech, though, users could easily get lost between the audio editing and transcription correction modes, so we chose to guide the user's attention through these modes via the movement of the editing caret into a quasi-modal interface.

Pauses in speech deliver nuanced meaning such as hesitation or emphasis, so easy and powerful manipulation of pause duration is important. A system called SpeechSkimmer automatically condenses pauses for fast auditory skimming [2]. Other previous systems supported pause editing via a designated button [3] or specialized tags [22]. Rubin et al.'s system used the period key as a shortcut to insert pause tags, but the duration of the gap was preset and required the use of a separate menu. Our approach is based on a traditional text editor, where users adjust lengths of pauses by adding and deleting spaces with the space bar and delete keys. The length of the pause corresponds to the number of spaces.

Perhaps most importantly, studies of the mental and linguistic effects of AAC discourse are limited despite the variety of AAC system designs discussed above. Those studies that exist are focused on the larger category of computer-mediated communication (CMC), which is dominated by textual media such as SMS, email, or Facebook posts. For example, Kiesler, Siegel, and McGuire [16] found more equalized group participation and more uninhibited expression of opinions in synchronous text-based CMC than in face-to-face discussions. Asynchronous CMC, similar to a discussion board, induces more prosocial behavior and, in fact, more informal communication styles over time than face-to-face [30]. On the other hand, formality and politeness in emails has been shown to increase as the social distance, status gap, and importance of a request increase [5]. How these findings about textual media relate to spoken modes of CMC has not been heavily investigated, however.

DESIGNING SIMPLESPEECH

Building on the capabilities developed in these prior studies, SimpleSpeech is a web-based application for recording

and editing short voice messages in a discussion setting (see Fig. 1). The design of SimpleSpeech is inspired by the transcription-based speech editing systems developed by Rubin [22], Yoon [33], and Whittaker [31], which also use ASR transcription as a semantic and visual proxy for audio. Conceptual modifications were necessary to suit the "live" editing paradigm, in which the user records and edits spoken comments in a unified workflow.

We followed an iterative procedure to improve the design and interactions of SimpleSpeech. After building an initial prototype of the application, an informal pilot test was conducted with 5 participants. Each user was given a brief introduction to the software and shown how to use the basic features, then given the scenario of creating an audio response to a written claim on an online forum. The prompts used in the tests were adapted from the GRE Pool of Issue Topics. This user feedback helped us improve the capabilities of SimpleSpeech in several ways:

- *Pause manipulation*. An important finding in the pilot study was the importance of being able to introduce and adjust pauses between words, not just to remove them. These gaps in the audio help make natural-sounding cuts between audio clips as well as punctuate claims (e.g., the end of a sentence). The original system only allowed the user to delete pauses, so we added a spacebar action to insert a fragment of silence.

- *Live insertion*. Users often misspoke a word or phrase in the middle of a message. Since re-recording the entire message was a large burden, a partial deletion and re-recording capability was required. Our interface maintains the text-based principle that the cursor position uniformly indicates the focus of editing. Therefore, in SimpleSpeech, clicking the Record button naturally inserts a new stream into the existing audio at the point corresponding to the cursor position.

- *Quasi-modal transcription editing*. In consideration of the recipient, the pilot testers wanted to correct ASR errors. We found that an interface with an audio-editing mode for revising speech and a text-editing mode for correcting mistranscriptions confused users. This led us to design a small, separate quasi-modal editing box (see Fig. 2) that pops up during transcription editing. The movement of the cursor to the modal box clearly indicates that the system is in a separate text-editing mode, accessed by pressing the Return key.

Our text-based approach to speech editing requires a reliable transcription as well as time intervals corresponding to each word. Both of these requirements are fulfilled by the IBM Watson Developer Cloud speech-to-text transcription service, which is reported to have a word error rate of around 10% [25].

QUALITATIVE EVALUATION

The interaction paradigm of SimpleSpeech was tested in a qualitative assessment to determine (1) the practicability of a lightweight text-based audio editor, (2) the effects of minor transcription errors on audio consumption and production, and (3) the implications of being able to edit audio in an asynchronous online discussion.

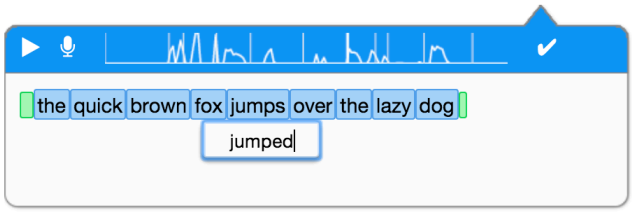

Figure 2. To keep the user interface from becoming cluttered with secondary functionality, the transcription editing feature was implemented as a modal interaction. The pop-up box shown above gives clear visual indication to the user that they are no longer directly editing the audio.

Participants were introduced to the functionality of the system, then given two untimed tasks. First, to simulate an asynchronous audio discussion, the test users were asked to listen to an audio comment left by the previous tester and create an original audio response. Next, they received a different, textual prompt and created an audio comment which would be consumed by the next user. In both cases the user was asked to edit his or her recording to be polished and clear. The participants were interviewed at the end of the test; these interviews were transcribed with conversational elements filtered out. Themes were extracted from the remaining sentences via open coding, followed by flat coding to sort statements among the themes. Cohen's κ between the two coders was .78, indicating the reliability of the categorized themes.

The sample for the study consisted of 9 test subjects (4 male, 5 female, mean age 22 yrs; henceforth denoted $P_1, P_2, \ldots, P_9$). All participants were native English speakers. Two individuals, P_2 and P_3, were media editors who provided technical feedback and a comparison to professional audio editing. The remainder were interns at edX, an educational technology non-profit, and students at RSI, a summer program for gifted high school students.

Results

The coding process revealed some themes that replicated the findings of Whittaker and Amento [31], as well as some novel results. For instance, the consensus was that a text-based editing paradigm provides sufficient control to render waveform manipulation unnecessary (P_4, P_5, P_6, P_8), as well as being"more accessible, more doable" than pure waveform editing (P_3, P_7). On the consumption side, the transcript was helpful in allowing users to "see all the points [the speaker was] making instead of having to remember them" (P_3, P_4, P_6). Overall the presence of errors was not a heavy distractor from the content of the messages, echoing findings from Whittaker's voicemail study [8], although one user described a greater focus on the audio because of these errors (P_8).

Although users were mostly accustomed to using the transcript for consuming the messages (P_6), they also pointed out that the audio gave them the ability to "feel the emotions

coming across, so ... you can still relate to what the persons feeling" (P_1, P_5, P_8). This result indicates that text played the most significant role in *factual* comprehension, while audio was most useful for *tonal* comprehension. Since both modalities seem necessary to the purpose of SimpleSpeech, we decided to maintain the combined paradigm throughout the study.

In addition to these transcript-related results, the qualitative study also yielded the following new themes:

The primary use of lightweight voice editing is to make fine-grained rather than large-scale adjustments. The most commonly-used manipulation during the qualitative study was the removal of disfluencies (P_1, P_2, P_4, P_5, P_7), followed by pause deletion (P_2, P_3, P_5, P_6, P_8). Only P_1 and P_8 edited large chunks of audio by deleting or rerecording, and P_8 reported doing so only to improve the smoothness of a smaller change in a sentence. Perhaps because SimpleSpeech was presented as a tool to be briefly used to "clean up" recordings, participants focused on removing the "embarrassing" and "awkward" sounds (P_1, P_5).

The linearity of audio leads to a pressure to organize one's thoughts during recording. P_4, P_7, and P_9 described a "psychological sort of ... need to get it all out, and the fact that it won't necessarily be as organized there." Another tester, P_5, had "a tendency to get like a blank slate" in which he "couldn't think of anything to say." The elevated mental task load that P_5 describes could be inherent in oral discussion; P_9 noted that "[it] might just be the fact that I was recording," and that "editing would make it nicer." Users likely reported this pressure despite the capability to edit their recordings due to the previous theme of lightweight editing, and also because this pressure arose in the moment of recording rather than throughout the entire message-composition workflow.

Awareness of the recipient and the editability of the audio drive up the quality of contributions. Four users mentioned the formality of their recordings (P_1, P_5, P_7, P_9), which they attributed to "an expectation" to edit, given that "someone else would know that I had that opportunity" (P_8). As P_9 explained her motivation to edit:

> Personally I'm editing to express myself a little more in a polished way when I'm writing.... especially if I know someone else is going to review it and be able to respond, I want to make sure I'm as clear as possible and as concise in a way that doesn't really come across when I'm talking.

Listening to another participant before initiating their own comment may have been a factor in determining the users' performance (P_9), as well as the presence of editing tools: "Since you have the ability to edit things, it feels like you're talking to somebody who's prepared a point or a conversational view" (P_5). The negative viewpoints of a few users on the editing ability were generally expressed in similar terms, so we considered this theme important for the success of AAC.

QUANTITATIVE EVALUATION

Our second, quantitative experiment was conducted in order to validate the three original themes identified in the qualitative evaluation above. In doing so, these measurements collectively assess the efficacy of SimpleSpeech and, indirectly, the usefulness of audio editing tools in general for educational discussions.

Procedure

Two between-subject dimensions were studied: students versus teachers, as well as the formality of initial stimulus recordings (see "Stimuli and Formality Measures" below). In addition, the dimension of no-editing versus editing was studied on a within-subject basis. Participants in the study were given two task parts in random order: recording messages without editing functionality (the **No Editing**, or **NE** task) and using SimpleSpeech (the **Editing**, or **E** task). Each task consisted of "discussion threads," in which users read a prompt statement, listened to another person's opinion on the issue, then produced an original response. Participants responded to two threads for each task, for a total of four messages of about one-minute duration each. (Before starting the E part, participants were given a standardized tutorial to learn how to edit using SimpleSpeech.)

After each task, the NASA Task Load Index (NASA-TLX) questionnaire was used to quantify the pressure or mental task load of producing a voice message [12]. NASA-TLX is a subjective analytical tool that measures task load along six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants in the study completed the TLX ranking and weighting procedure after each task to obtain comparisons between the no-editing and editing situations.

The quantitative study was conducted at a small suburban public high school in the midwestern U.S. with 28 volunteer participants (16 students, ages 16-18, and 12 teachers; 13 male, 15 female). This location was ideal for the study because the sample contained a variety of learning and speaking styles as well as different aptitudes for technology and discussion.

Stimuli and Formality Measures

The criterion used for formality was the F-score, a measure of contextuality introduced by Heylighen and Dewaele in 2002 [13]. The F-score is a purely textual metric based on the frequencies of various parts of speech in a text: nouns, adjectives and prepositions decrease contextuality and increase the F-score since they are independent of the circumstances around the text, while verbs, adverbs, pronouns, and interjections increase contextuality and decrease the F-score.

The initial stimulus recordings for each of the prompt statements were generated by a group of five initial volunteers, who were asked to plan and edit some of the comments and improvise on the others. After splitting the resulting messages by formality, the average F-score was 53.7 for the Group A messages and 49.4 for the Group B messages ($p = .11$ by unpaired t-test), reflecting the likely greater contextuality of the recordings produced on-the-fly. Group A

stimuli also tended to use longer words than those for Group B (4.62 versus 4.38 letters, $p = .047$), to be more concise (113 versus 193 words, $p = .0095$), and to have higher speaking rates (149 compared to 134 words per minute, $p = .076$). After obtaining and categorizing these messages, the voices were anonymized by adjusting the pitch randomly.

Since the qualitative study had indicated that prior exposure to other individuals' messages could affect the perceived formality in the discussion, half the participants (Group A) listened to only formal recordings, while the other half (Group B) listened to only informal ones. We hypothesized that the participants in Group A would produce more formal messages due to the stimuli they received.

Results

We analyzed the effects of the live voice editing features by analyzing system logs, speech contents, and the task load survey data. The results fell into the following three categories:

Utilization of SimpleSpeech Features

As in the qualitative study, most participants appreciated and took advantage of the ability to edit their messages. They found the interface intuitive and natural, presumably due to familiarity with text-editing interfaces.

The amount of editing that users engaged in varied widely: on average, about 17.7 edits were made to each comment ($SE = 2.3$, including inserting a new recording, inserting a pause, deleting words, or deleting a pause). Of these changes, the vast majority were subtractive: 6.6 word deletions ($SE = 1.6$) and 6.3 pause deletions ($SE = 1.1$) per message. This was consistent with the previously-observed inclination to remove disfluencies and "awkward" hesitations from the recordings. Both teachers and students exhibited this impetus with no significant difference (12 deletions for students, 16.2 for teachers, $p = .36$ by unpaired t-test), so correcting misspeeches was fairly universal across participants. Insertions of any kind were less common, at 1.1 per message ($SE = 0.3$), probably because users tended to correct themselves *post hoc* during the initial take and delete the mistakes afterward. Four users went without performing any edits at all on at least one message; three out of these were teachers, and their messages were generally already fluent.

One error that several participants made was to use the Delete key on tokens to fix transcription errors, which resulted in the permanent deletion of that audio token. However, emphasis in the tutorial that the Delete key deleted the audio permanently did help other participants avoid making this mistake. Another misconception we observed in a few participants was a tendency to treat SimpleSpeech as a dictation tool. These users paused for long periods of time during recording sessions and neglected to play back the messages during editing. Furthermore, their inclination after stopping a recording session was to go back and correct transcription errors so that the visual representation made sense.

Effect on Task Load

Since the NASA-TLX scale is subjective, it does introduce variability between participants due to the differences between their perceived skill at the task [12]. For instance, one participant could rate the recording task at a 3 out of 20, while another could rate the very same task at a 15. Therefore, the strongest comparisons of task load were made in the within-subject dimension, which was the ability or inability to edit.

Overall, the students reported significantly *lower* levels of mental task load or pressure during the E task than the NE task (8.70 compared to 10.8, $p = .011$ by paired t-test). The values for the individual components of the TLX, shown in Table 1, yielded the following contributory dimensions on the TLX questionnaire:

- *Temporal demand*. Students rated the temporal demand at 7.81 for the E task, significantly less than the NE rating of 10.5 ($p = .037$ by paired t-test). As described by the TLX form, temporal demand refers to "time pressure due to the rate or pace at which the tasks or task elements occurred" [12]. Students verbally described the increase in time demand reported on the TLX in terms of having to think of words quickly, with the knowledge that every second not filled with speech would be an embarrassing silence.

- *Performance*. Students felt more concern about the quality of their messages in the NE task, with a marginally significant difference of 10.0 compared to 8.25 for the E task ($p = .083$ by paired t-test). Just as the participants in the prior qualitative study had articulated a desire to make their messages better for the sake of their listeners, the students also evidently wanted to improve their recordings in the NE task. The inability to do so resulted in elevated task load due to performance, while for the E task the stress was lower because they were afforded the chance to correct their mistakes.

- *Effort*. Similarly to performance, students reported having to work significantly harder in the NE task to complete it to their desired level (rated 11.6 compared to 9.06 in the E task, $p = .014$ by paired t-test). This increased effort could correspond to the additional mental activity which had to be expended in order to generate speech fluently and without excessive hesitation.

While the teachers also reported slightly lower average workload levels in the E task, as shown at the right of Table 1, this difference was not significant. In fact, 7 of the 12 participating teachers actually rated the E task as requiring a higher workload than the NE task. This subset of the teachers, 5 of whom were in Group A, reported an average task load greater in the E task than the NE task for *all* dimensions. The reason for this rating, these teachers explained, was that the availability of the editing tools caused them to feel more worried about their performance (likely due to P_8's "expectation to edit"). They were thus made to expend more effort to preserve the existing fluidity of their messages.

Overall, the fact that the differences in perception of workload varied so much among teachers indicates that they were not as heavily affected by the ability to edit as the students,

	Students (N = 16)		Teachers (N = 12)	
Task	*E*	*NE*	*E*	*NE*
Mental Demand	9.56 (1.0)	11.1 (1.0)	11.4 (1.3)	10.8 (1.4)
Physical Demand	3.69 (0.6)	2.63 (0.5)	4.00 (1.3)	2.83 (0.6)
Temporal Demand	7.81 (1.1)	10.5* (1.0)	7.50 (1.4)	10.0 (1.3)
Performance	8.25 (0.7)	10.0^{+} (0.6)	8.50 (1.4)	9.67 (1.3)
Effort	9.06 (1.1)	11.6* (0.9)	9.83 (1.7)	10.4 (1.4)
Frustration	7.75 (1.1)	8.88 (1.0)	8.42 (1.6)	10.0 (1.7)
Total (weighted)	8.70 (0.7)	10.8* (0.6)	9.47 (1.2)	10.6 (1.3)

Table 1. The unweighted mental work load ratings reported by students and teachers from recording voice messages. *E* and *NE* refer to the tasks in which editing was allowed and disallowed, respectively. Each value ranges from 1 to 20, indicating the subjective level of task load that participants rated for each metric; standard error is given in parentheses. ($^{+}$ – $p < 0.10$, * – $p < 0.05$, paired two-tailed comparison of *E* and *NE*)

who clearly appreciated the security that SimpleSpeech offered.

Speech Formality Between Subject Groups

Contrary to the hypothesis that prior exposure to audio messages would affect the formality or linguistic traits of new messages, the F-scores of the participants' output were unrelated to the group they were in, as shown in Table 2. However, the students' speaking rates in Group A were significantly faster than those in Group B (p = .016 by unpaired t-test). Moreover, students in Group B spoke significantly *slower* than teachers exposed to the same stimulus recordings (p = .0016 by unpaired t-test), but there was no corresponding difference in Group A. This may reflect the auditory component of formality observed by the qualitative test users, if the faster speaking rate was interpreted by students as thought-out or engaging speech.

Interestingly, the same teachers who reported higher task load in the editing task also produced more formal messages than the other teachers (mean F-score 56.6 compared to 53.0, p = .12 by unpaired t-test) and with longer words (4.58 compared to 4.35 letters, p = .0082). In fact, the student participant group also contained members who rated the E task as more demanding than the NE task, though fewer in number (4 out of 16); these students produced much more formal messages than their peers as well (F-score 57.9 compared to 53.5, p = .027 by unpaired t-test). These participants could have had more experience speaking extemporaneously or felt less inclined to speak conversationally, ultimately leading to SimpleSpeech not being as useful to them.

On the whole, the formality of the recordings was not affected by the stimulus message or even whether the participant was a teacher or a student. Considering that the F-score measures contextuality between the speaker and the audience, the principal sources of variation in F-score must have been personal aptitude and preference for the medium and the scenario of an online forum discussion.

FORMALITY COMPARISON

Contextuality in the online voice-based discussion scenario could be highly indicative of AAC's potential applications, and to our knowledge this trait has not been studied extensively. We therefore conducted a comparison between the voice messages composed during this study and several publicly-available corpora. To our knowledge, no corpus exists that was collected in an audio-text situation like SimpleSpeech, so contrasts in formality inevitably arise from the differences in medium.

The SimpleSpeech text, representing AAC, contained 14,569 words from 112 messages. For written documents, we used several sections of the well-known Brown corpus to compile general categories of text: nonfiction, fiction, and technical writing (consisting of government documents, scientific articles, and news) [17]. We obtained chatroom text from the `nps_chat` corpus, face-to-face conversation data from the `webtext` corpus, and telephone data from the `switchboard` corpus, all available as part of the Natural Language Toolkit (NLTK) [4]. Finally, we also analyzed email communication in non-spam messages from the Enron corpus [26], as well as a corpus of Twitter posts [9].

The results of this comparison, shown in Fig. 3, illustrate the middle-ground that AAC takes relative to oral and written media. The least formal and most contextual corpora were those based on oral communication (with the notable exception of web chat messages), while the most formal and least context-dependent were the written texts, including email and Twitter posts. We will note three possible explanations for the formality of each medium based on the ordering of the corpora:

Speaker-audience relationship. Since the F-score is inversely related to contextuality, it is reasonable that the chat and telephone corpora had the lowest F-scores because the participants knew each other and were conversing on a one-to-one basis. On the other hand, the written forms of communication (with the exception of email) were more formal because the audience was defined more loosely and not necessarily acquainted with the speaker. AAC using SimpleSpeech was more closely related to the latter condition (as an online forum discussion), which probably contributed to its greater formality compared to the other spoken corpora.

	Students		Teachers	
Group	*A*	*B*	*A*	*B*
Formality (F-score)	55.8 (1.1)	53.4 (1.3)	54.7 (1.6)	55.5 (1.5)
Word length	4.40 (0.06)	4.44 (0.05)	4.50 (0.06)	4.47 (0.06)
Disfluencies (per 100 words)	1.59 (0.35)	2.38 (0.41)	1.27 (0.27)	1.41 (0.38)
Word count	101 (6.5)	140^{+} (13)	155 (19)	130 (9.4)
Speaking rate (words/min)	130 (5.0)	115* (3.8)	137 (6.8)	138 (4.9)

Table 2. Various metrics describing the formality of the audio messages produced by each participant group. Group A listened to more formal initial stimulus recordings than Group B. There were few significant differences in these criteria between the groups, indicating that formality was dependent on the general context of AAC as well as the speaker's preference, especially for teachers. Standard error is given in parentheses. ($^{+}$ – $p < 0.10$, * – $p < 0.05$, unpaired two-tailed comparison of groups A and B)

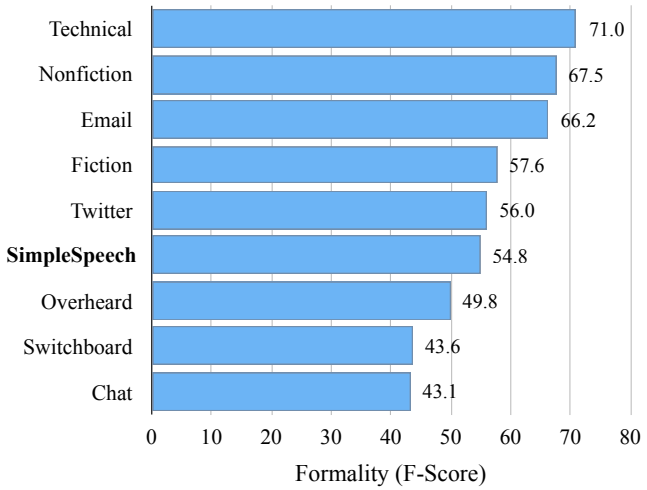

Figure 3. The formality of corpora in different genre and media. The messages produced using SimpleSpeech during the quantitative study are intended to reflect general AAC discussion characteristics, and seem to be more formal than other spoken forms of communication but not as formal as email.

Immediacy of communication. The tendency to speak or write more contextually when the recipient must respond immediately helps explain why the online chat text, though written, was more contextual and less formal than the oral corpora. It also justifies the fact that the email corpus was more formal than all of the other direct communication media. Again, AAC falls toward the more formal end of this spectrum because there is little temporal proximity between the speaker and the audience.

Tendency toward verbosity. Media that pressured the creator to be brief or precise were more formal and less contextual. For instance, writing technical documents requires the preferential use of nouns over pronouns to maximize clarity. Twitter messages are, of course, limited to 140 characters, leading to a greater concentration of meaning that favors less contextual words. For AAC, the ability to edit could potentially influence the contextuality if discussion members were pressured to trim down their recordings. For our study, however, the F-score was not affected by verbosity, because edits were more concentrated on removing disfluencies and misspeeches than on improving concision.

DISCUSSION

In this study two forms of responding to pressure in a communication task were measured: the mental workload involved in completing the task and the degree of formality in the messages created in the task. Using this information, we will evaluate the strengths and weaknesses of SimpleSpeech as a tool for enabling AAC as well as the viability of AAC in educational and collaborative contexts.

Imbalance Between Speaker and Listener

As Grudin notes, it is critical for collaborative software to spread the burden of usage equally on its constituent members. For instance, he cites email as a medium in which "everyone generally shares the benefits and burdens equally" [10]. On the other hand, voice applications create inequality between speaker and listener since the former must expect that the latter will listen thoroughly and carefully to the message, a relatively slow task compared to reading.

However, the premise of SimpleSpeech is that the bias toward the speaker is reversed. ASR transcription can greatly facilitate the listener's task, as has already been demonstrated [8, 29], bringing the workload down and closer to that of reading. Meanwhile, students who record messages could experience a *greater* workload relative to writing because of the linearity of audio, which prevents them from correcting mistakes after the fact and thereby elevates the pressure to do well the first time.

SimpleSpeech was demonstrated to be a useful counterbalance in situations where the speaker's workload is elevated. In the qualitative study, some users noted the pressure "to have organized thoughts" and to "sound composed more" during recording, but that "editing would make it nicer because you can go back and fix the mistakes" (P_2). Furthermore, the level of control was just right for most users: since they focused on deleting the disfluencies and pauses in their speech, the word-tokenized editor for the most part provided exactly the information needed to quickly delete undesirable sounds. For the few users who did want to edit on a larger scale, the audio insertion feature was deemed helpful as well.

In the quantitative evaluation, we found strong evidence to support the use of SimpleSpeech, especially for students. There was a significant decrease in task load on students when given the capability to edit, even in spite of the added time required to listen to the message and perform the editing. On the other hand, teachers did not find SimpleSpeech editing as useful, probably because they already had higher confidence in their recordings as being of acceptable quality. One teacher reasoned that he was "already used to hearing [his] own voice" from lecturing, a medium where statements cannot be retracted as easily as with SimpleSpeech. However, many teachers did use the editing tools, even though their workload levels were not significantly different with or without this opportunity. This would indicate that the editing tools are a valuable option for producers to have, but users should not be obligated to use them.

Implications for Formality in AAC

Our comparative analysis of formality using Heylighen et al.'s F-score [13] is unique in its application to mixed-media discourse. While the corpora against which SimpleSpeech messages were tested were generated through either purely-verbal or purely-written means, users in this study were required to consider both the textual, semantic meaning of the transcript and the auditory, connotative meanings expressed by the voice. As a result, our comparison of contextuality scores inherently lacks the auditory component that may be communicated through slang, dialect, or speaking rate. We do not know of any one single measure that can capture these phenomena across different media. However, our findings are useful for describing the formality of the textual message content, which is arguably the primary purpose of the communication as well as the aspect on which message consumers focused most.

The textual formality of the SimpleSpeech discussions was not significantly affected by the experimental conditions, indicating that for the most part, users are likely to adopt their own style for audio messages suitable for a discussion environment. Nevertheless, it is critical to the success of general-purpose AAC that the formality of discussion be controlled to some extent, so that collaborators feel willing to participate. For students this impetus toward quality is not as problematic, since they felt more relaxed rather than more stressed with editing functionality. If discussion quality were driven up by artificial means, however, such as by grading students on the eloquence of their comments or evaluating employees on the basis of their online interactions, then individuals might gravitate toward socially "safer" modes of communication over which they feel more control (namely, text). Proper acquaintance with audio editing capabilities is essential for AAC's survival under these pressures toward high-quality production.

To provide an example, peer feedback within an online course would be an ideal use of AAC using SimpleSpeech because students could send messages to a well-defined audience, thereby compensating for the additional formality imposed by the spatiotemporal distance between the participants. The editing tools would also drive students to produce better discussion input, increasing productivity and enhancing the learning experience. On the other hand, enabling editing for personal communication, such as WhatsApp voice messages, may reduce the desired informal speaking style of the platform. Since the contextuality demanded by each situation is different, future audio-based collaboration platforms must consider the factors presented here and tailor their functionality accordingly.

CONCLUSIONS

SimpleSpeech's intuitive design alleviates the pressure associated with the linearity of audio because users have the ability to easily fix errors after the fact. Furthermore, we designed SimpleSpeech to clearly distinguish audio and text modalities, from the visual cues provided by the waveform to the quasi-modal interface for correcting transcription errors. Students' use of these editing tools resulted in greater feelings of comfort when producing comments than without SimpleSpeech functionality. The true utility of this software, then, was to (at least partly) un-linearize audio, even making it more text-like.

Because studies of the linguistic and social characteristics of computer-mediated communication have been mostly limited to textual interactions, we also explored the formality and contextuality of AAC. Our finding that it was roughly in between spoken and written media is not discouraging *per se*; however, the relatively formal characteristics of AAC must be taken into account before such a system is implemented in practice. Nevertheless, we feel that the small-scale edits that users engaged in during this study are reassuring for potential applications of AAC. Removing disfluencies and pauses allows users to feel comfortable with their recording while maintaining the spontaneity of thought in a spoken message.

The results of the qualitative study point to new directions for improving SimpleSpeech. For instance, on initial exposure to the application users initially tended to focus preferentially on the text instead of on the voice. Slightly different visual layouts of the application, such as overlaying or juxtaposing the transcription on a more prominent waveform, could help users understand better that the text is a secondary tool. Another possible feature could be automating certain edits, such as removing disfluencies and hesitations, to improve efficiency and edit quality even further. Additionally, the findings in our quantitative study revealed promising trends concerning the benefits of AAC for online discussion, but may need a larger pool of test participants to attain statistical significance.

Our hope in developing SimpleSpeech is that asynchronous audio communication will gain greater usage in education and other collaborative settings. With the combination of ASR transcription for listeners and low-barrier editing tools for speakers, voice-based communication tools can engage students and improve the quality of collaboration on the Web.

ACKNOWLEDGMENTS

We would like to thank all the interns, high school students, and teachers who participated in this study. We owe thanks to the Center for Excellence in Education, whose Research

Science Institute summer program at MIT provided the opportunity to conduct this research. We also thank Nicholas Chen, Abigail Sellen, and François Guimbretière for design inspiration.

REFERENCES

1. Stephen Ades and Daniel C Swinehart. 1986. *Voice annotation and editing in a workstation environment*. XEROX Corporation, Palo Alto Research Center.
2. Barry Arons. 1993. SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology (UIST '93)*. ACM, New York, NY, USA, 187–196. DOI: http://dx.doi.org/10.1145/168642.168661
3. Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Trans. Graph.* 31, 4, Article 67 (July 2012), 8 pages. DOI: http://dx.doi.org/10.1145/2185520.2185563
4. Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.
5. Thomas Cho. 2010. Linguistic Features of Electronic Mail in the Workplace: A Comparison with Memoranda. *Language@Internet* 7, 3 (2010).
6. Juan Casares et al. 2002a. Simplifying Video Editing Using Metadata. In *DIS '02 Proceedings*. London, 157–166.
7. Janet M. Baker et al. 2009. Developments and directions in speech recognition and understanding, Part 1. *Signal Processing, IEEE* 26, 3 (2009).
8. Steve Whittaker et al. 2002b. SCANMail: a voicemail interface that makes speech browsable, readable and searchable. *CHI Letters* 4, 1 (2002).
9. Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification using Distant Supervision*. Technical Report. Stanford.
10. Jonathan Grudin. 1988. Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. *ACM Conference on Computer-Supported Cooperative Work* (1988).
11. C Halverson, D Horn, C Karat, and John Karat. 1999. The beauty of errors: Patterns of error correction in desktop speech systems. In *Proceedings of INTERACT99*. 133–140.
12. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, P. A. Hancock and N. Meshkati (Eds.). North Holland Press, Amsterdam.
13. Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: an empirical measure. *Foundations of Science* 7 (2002), 239–340.
14. Debby Hindus and Chris Schmandt. 1992. Ubiquitous Audio: Capturing Spontaneous Collaboration. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work (CSCW '92)*. ACM, New York, NY, USA, 210–217. DOI: http://dx.doi.org/10.1145/143457.143481
15. Philip Ice, Reagan Curtis, Perry Phillips, and John Wells. 2007. Using asynchronous audio feedback to enhance teaching presence and students' sense of community. *Journal of Asynchronous Learning Networks* 11, 2 (2007).
16. Sara Kiesler, Jane Siegel, and Timothy W. McGuire. 1984. Social Psychological Aspects of Computer-Mediated Communication. *Amer. Psychologist* 39, 10 (1984), 1123–1134.
17. Henry Kučera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence.
18. Jennifer Lai and John Vergo. 1997. MedSpeak: Report Creation with Continuous Speech Recognition. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, New York, NY, USA, 431–438. DOI: http://dx.doi.org/10.1145/258549.258829
19. Philip Marriott. 2002. Voice vs text-based discussion forums: An implementation of Wimba Voice Boards. In *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Vol. 2002. 640–646.
20. Philip Marriott and Jane Hiscock. 2002. Voice vs Text-based Discussion Forums: An Implementation of Wimba Voice Boards. In *Proc. E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, M. Driscoll and T. Reeves (Eds.). Chesapeake, VA.
21. Jody Oomen-Early, Mary Bold, Kristin L. Wiginton, Tara L. Gallien, and Nancy Anderson. 2008. Using asynchronous audio communication (AAC) in the online classroom: a comparative study. *Journal of Online Learning and Teaching* 4, 3 (2008).
22. Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-Based Tools for Editing Audio Stories. In *UIST '13*. 113–122.
23. George Saon, Hong-Kwang J. Kuo, Steven Rennie, and Michael Picheny. 2015. The IBM 2015 English Conversational Telephone Speech Recognition System. *Interspeech* (2015).
24. Christopher Schmandt. 1981. The Intelligent Ear: A Graphical Interface to Digital Audio. In *Proceedings, IEEE International Conference on Cybernetics and Society, IEEE*.

25. Hagen Soltau, George Saon, and Tara N. Sainath. 2014. Joint training of convolutional and non-convolutional neural networks. In *Proceedings of the IEEE Intl. Conference on Acoustic, Speech and Signal Processing*. Florence, 5572–5576.
26. Will Styler. 2011. *The EnronSent Corpus*. Technical Report 01-2011. University of Colorado at Boulder Institute of Cognitive Science, Boulder, CO.
27. Chi-Hsiung Tu and Marina McIsaac. 2002. The Relationship of Social Presence and Interaction in Online Classes. *Amer. Journal of Distance Education* 16, 3 (2002).
28. Sunil Vemuri, Philip DeCamp, Walter Bender, and Chris Schmandt. 2004a. Improving Speech Playback Using Time-compression and Speech Recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 295–302. DOI: http://dx.doi.org/10.1145/985692.985730
29. Sunil Vemuri, Philip DeCamp, Walter Bender, and Chris Schmandt. 2004b. Improving Speech Playback Using Time-Compression and Speech Recognition. *CHI Letters* 6, 1 (2004).
30. Joseph B. Walther. 1995. Relational Aspects of Computer-Mediated Communication: Experimental Observations over Time. *Organization Science* 6, 2 (1995), 186–203.
31. Steve Whittaker and Brian Amento. 2004. Semantic Speech Editing. *CHI Letters* (2004), 527–534.
32. Lynn Wilcox, Ian Smith, and Marcia Bush. 1992. Wordspotting for Voice Editing and Audio Indexing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. ACM, New York, NY, USA, 655–656. DOI: http://dx.doi.org/10.1145/142750.150715
33. Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: blending ink, speech, and gesture to support collaborative document review. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 481–490.
34. Dongwook Yoon, Nicholas Chen, Bernie Randles, Amy Cheatle, Corinna E. Loeckenhoff, Steven J. Jackson, Abigail Sellen, and François Guimbretière. 2016. Deployment of a Collaborative Multi-Modal Annotation System for Instructor Feedback and Peer Discussion. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM.