

UC Berkeley

UC Berkeley Previously Published Works

Title

TalkLess: Blending Extractive and Abstractive Summarization for Editing Speech to Preserve Content and Style

Permalink

<https://escholarship.org/uc/item/9vr9n5x5>

Journal

PROCEEDINGS OF THE 38TH ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY, UIST 2025

Authors

Benharrak, Karim

Peng, Puyuan

Pavel, Amy

Publication Date

2025-09-28

DOI

10.1145/3746059.3747795

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

TalkLess: Blending Extractive and Abstractive Summarization for Editing Speech to Preserve Content and Style

Karim Benharak
karim@cs.utexas.edu
The University of Texas at Austin
Austin, TX, USA

Puyuan Peng
pyp@utexas.edu
The University of Texas at Austin
Austin, TX, USA

Amy Pavel
amypavel@eecs.berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

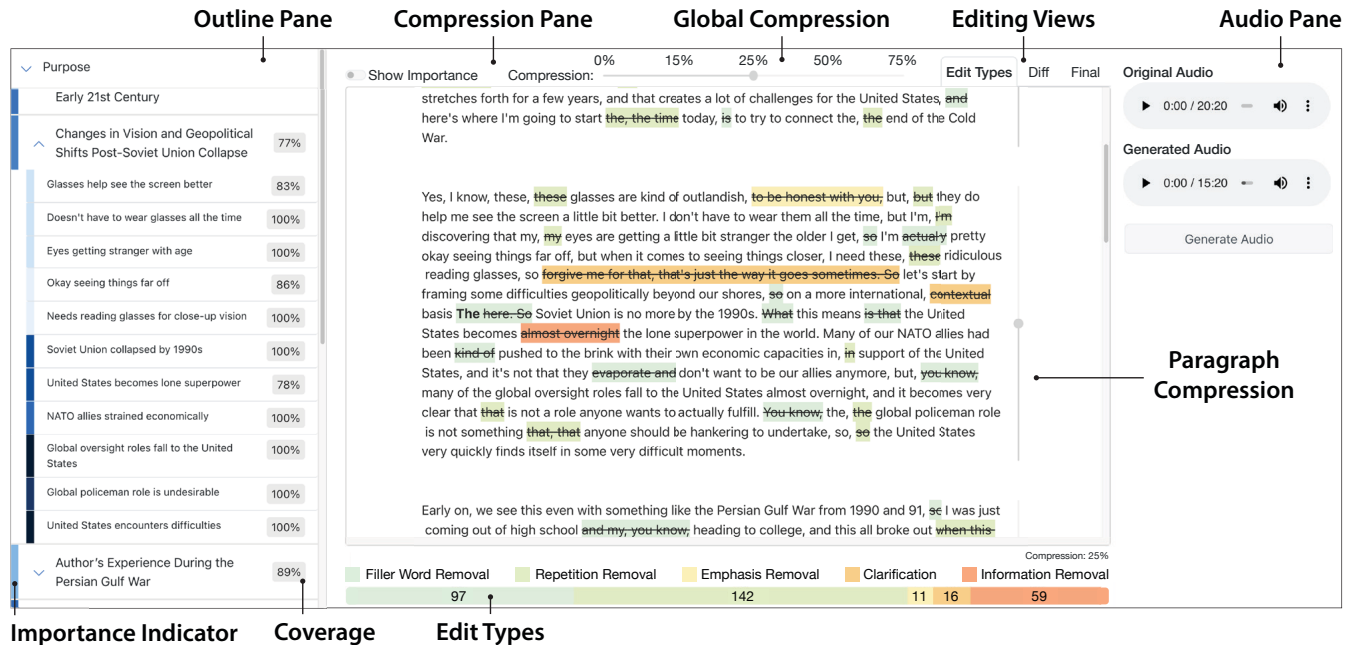


Figure 1: TalkLess's interface lets creators edit their speech by skimming and browsing the outline pane to determine regions to cut, using the compression pane editing the transcript directly or by selecting a global- or paragraph-level compression amount, and listening to the original or generated audio aligned to the transcript in the audio pane.

ABSTRACT

Millions of people listen to podcasts, audio stories, and lectures, but editing speech remains tedious and time-consuming. Creators remove unnecessary words, cut tangential discussions, and even re-record speech to make recordings concise and engaging. Prior work automatically summarized speech by removing full sentences (extraction), but rigid extraction limits expressivity. AI tools can summarize then re-synthesize speech (abstraction), but abstraction strips the speaker's style. We present TalkLess, a system that flexibly combines extraction and abstraction to condense speech while preserving its content and style. To edit speech, TalkLess first generates possible transcript edits, selects edits to maximize compression, coverage, and audio quality, then uses a speech editing model to translate transcript edits into audio edits. TalkLess'

interface provides creators control over automated edits by separating low-level wording edits (via the compression pane) from major content edits (via the outline pane). TalkLess achieves higher coverage and removes more speech errors than a state-of-the-art extractive approach. A comparison study (N=12) showed that TalkLess significantly decreased cognitive load and editing effort in speech editing. We further demonstrate TalkLess's potential in an exploratory study (N=3) where creators edited their own speech.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools.**

KEYWORDS

Speech, Audio Editing, Summarization, Creativity Support Tools

ACM Reference Format:

Karim Benharak, Puyuan Peng, and Amy Pavel. 2025. TalkLess: Blending Extractive and Abstractive Summarization for Editing Speech to Preserve Content and Style. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*, September 28–October 1, 2025, Busan,



This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '25, September 28–October 1, 2025, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2037-6/2025/09
<https://doi.org/10.1145/3746059.3747795>

Republic of Korea. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3746059.3747795>

1 INTRODUCTION

Millions of people create and consume audio-driven media such as podcasts and lectures. However, editing speech recordings to be clear, concise, and engaging is tedious and time-consuming [7]. For example, podcast creators spend 2-7x the recording length editing the speech (e.g., 7 hours to edit 1 hour of audio [94]). They carefully cut unnecessary audio including pauses, filler words (e.g., “um”, “uh”, “you know”), repetitions (at a word-, sentence-, or paragraph-level), and irrelevant content, all while making sure to not introduce incoherent or unnatural transitions by adding noticeable cuts [95]. Transcript-based editing tools [3, 9, 22, 51, 71, 76] let creators edit speech by deleting text or generating new words [22], but reading raw speech transcripts and deciding what to cut while preserving important information and the speaker’s style (e.g., variations in pitch, pacing, and tone to capture listener attention [73, 85]) remains challenging. Thus, creators without time or resources to edit (e.g., educators, amateur podcasters, or journalists) must publish less polished audio, or choose not to post their work altogether.

Prior automated approaches use **extractive** speech shortening which extracts segments from the original audio to preserve important content and speaker style [22, 69, 95]. For example, ROPE [95] selects important transcript sentences, and Rescribe [69] selects words from sentences (e.g., “An orange cat sat.” to “A cat sat.”). However, rigid extraction in ROPE [95] and Rescribe [69], often preserves repetitions, creates abrupt transitions, and leaves unresolved references (e.g., may keep “*she presented the results*” but remove who “*she*” is) [98]. The limitations of extraction are also well explored for text summarization [14, 56, 98, 101]. In response, a longstanding solution for text summarization is to use **abstractive** summaries that write new sentences and words to summarize sentences and can achieve fluent summaries with high content coverage (e.g., changing “*There is one event in 1990. There is another in 1992.*” to “*There were events in 1990 and 1992.*”). Prior work has explored using abstraction to create text summaries of speech transcripts to ease reading [54] or write from speech-to-text [58], but prior work has not yet applied abstractive summarization directly to edit speech. One approach may be to create an abstractive speech summary with an LLM, then synthesize the speech (e.g., with ElevenLabs [24] voice cloning). But, re-synthesizing speech strips away the speaker’s linguistic (e.g., word choice) and paralinguistic (e.g., emphasis with pitch) style that carry the speaker’s identity, message, and delivery [45]. Editors can also combine extractive and abstractive summarization by manually inserting synthesized segments into existing speech using transcript-based editors that allow inserting new words for synthesis through voice cloning [6, 23, 44, 70]. However, manually combining abstractive and extractive summarization is tedious.

We present TalkLess, a transcript-based editing system to condense speech that contributes a speech summarization algorithm and editing interface to condense speech. TalkLess’s algorithm flexibly combines extraction and abstraction to condense speech by jointly considering the speech transcript and audio through balancing preserving content and speaker style with maintaining

coherent and natural-sounding audio. TalkLess transcribes, aligns, and segments the original audio, generates potential transcript edits using an LLM, then selects a set of edits that maximize compression and content coverage without compromising audio quality. Finally, TalkLess translates the selected transcript edits into audio edits and synthesizes transitions between audio cuts and new words. Editing raw unscripted speech requires many low-level edits to remove filler words and repetitions, and requires high-level edit decisions, as editors often want to intentionally exclude irrelevant content depending on the editing purpose. TalkLess contributes a speech editing interface that separates the two concerns by automating low-level edits to clean up the speech and support editors in making high-level content decisions (Figure 1). To make low-level editing more efficient, TalkLess’s *Compression Pane* provides dynamic compression controls at both the transcript and segment level to automatically shorten speech, and visualizes filler word and repetition removals to support quick review of low-risk edits. To support high-level editing, TalkLess visualizes information removed by automated edits and indicates content importance based on the editing purpose (e.g., post a class lecture to YouTube) in the *Outline Pane* to help editors identify what content to keep or remove (e.g., a creator’s sidenote about his glasses is light blue).

The *Audio Pane* lets creators listen to the original or generated audio.

We evaluate TalkLess with a technical and results evaluation, an authoring user study, and an exploratory user study. Our technical evaluation reveals that TalkLess retains more content and reduces more speech errors than a state-of-the-art extractive audio shortening baseline [95], and participants in our results evaluation significantly preferred listening to TalkLess’s shortened speech for 15% and 25% compression compared to the baseline. Our user evaluation (N=12) revealed that all editors preferred using TalkLess to edit lecture speech recordings to an extractive baseline system, and experienced significantly lower cognitive load as well as significantly higher creative flexibility when editing speech recordings with TalkLess. In our exploratory user study (N=3) creators used TalkLess to edit their own speech and compare their results with fully re-synthesized versions. Creators expressed concern about losing vocal authenticity with re-synthesis and preferred TalkLess to preserve their original speaking style.

In summary, our work contributes:

- An automatic speech shortening algorithm that combines extraction and abstraction to condense speech by jointly considering the speech’s transcript and audio.
- TalkLess, a speech editing interface that supports editors by separating editing concerns into automated low-level edits to clean up speech and information highlighting to leave high-level content edits to the editor.
- A results evaluation and two user studies demonstrating why editors prefer TalkLess’s results and its interface for editing others’ and their own speech.

2 BACKGROUND

TalkLess provides a method and interface to shorten speech with a transcript and thus relates to prior work on speech editing tools, text summarization, and automatic speech shortening.

TalkLess	Prior Extractive Approach [95]
I was nervous at first, like, you know, because I hadn't done anything like this before. But I kept going. And then, um, one day it just clicked. which That moment gave me the confidence to keep pushing forward.	I was nervous at first, like, you know, because I hadn't done anything like this before. But I kept going. And then, um, one day it just clicked. That moment gave me the confidence to keep pushing forward.

Table 1: Compression with TalkLess (left) vs. a prior extractive approach [95] (right). TalkLess removes speech errors and condenses content across sentence boundaries, the prior approach leaves speech errors and drops the speaker's main point.

2.1 Transcript-based Editing Tools

Traditional audio editing workflows rely on timeline-based editors [2, 4], where editors manually remove unnecessary filler words, repetitions, and unimportant content by making precise waveform-level cuts. However, editors must listen closely to the audio to determine what to cut and carefully place cuts so as not to introduce audio errors, which is a time-consuming and tedious task.

To make audio editing easier and more efficient, prior work introduced transcript-based editing, which allows users to edit audio or video similar to a text document by time-aligning the words in the speech transcript with words in the audio [6, 9, 13, 22, 29, 39, 53, 75, 76, 82, 91]. However, reading long, unstructured transcripts remains cognitively demanding as raw speech transcripts include disfluencies and are often unstructured.

Prior systems aim to support editors in surfacing potential edits by highlighting segments with important content [91, 95, 96], narration errors (e.g., filler words, unclear pronunciation, bad flow) [22, 75], or grouping thematically coherent transcripts into chunks for further editing [53]. For example, Truong et al. [91] highlight memorable moments in social conversation transcripts, and ROPE [95] highlights relevant sentences to include in the final edited audio to support editors in preserving important content. However, while these systems assist editors in locating content to preserve or speech to remove, they either require the user to manually cut their speech [53, 91], which is tedious and time-consuming, or give little to no editing flexibility, for example, by only letting editors toggle sentences to include [95]. While Descript [22] automatically detects and removes filler words and word repetitions, it is limited to a fixed set of filler words, cannot detect semantic content repetitions and content to cut to make the audio more concise.

Beyond highlighting edits, prior work also aimed to support users in browsing and skimming long transcripts by providing section summaries [68], transcript segments based on topics [28, 68, 92], and outlines [41] which can be used to find content to cut. Similarly, to make reading and navigating speech transcripts easier, we segment the transcript into semantically meaningful segments and provide an outline based on information included in the speech. We expand on prior transcript navigation aids and visualize edit types to make skimming and reviewing edits easier.

2.2 Text Summarization

Text summarization condenses text content while preserving the core meaning, and is widely categorized as either extractive [27, 34, 59, 100] or abstractive [31, 36, 57, 64, 67] summarization. Extractive summarization condenses content by selecting a subset of existing sentences or words from the original text without modification and is popular in reading-focused applications where preserving

original phrasing matters [17, 35]. To facilitate text skimming GP-TSM [35] generates multiple levels of extractive summaries through LLM-based sentence compression, then overlays these summaries on top of each other to visually occlude less important words. Marvita [17], a human-AI reading support tool, uses extraction by automatically choosing a summative subset of text for users based on their time budget and questions they want to answer. However, extraction misses opportunities to improve flow and conciseness [98], especially when dealing with informal, repetitive, or unstructured text like speech transcripts. In contrast, abstractive summarization can be more flexible, concise, and coherent through the generation of new text [98]. For example, while extractive summarization might select the sentences “*The meeting started at 9 AM. We discussed the budget. The budget was over by 20%.*”, abstractive summarization generates “*The 9 AM meeting revealed a 20% budget overrun.*” Prior tools use abstractive summarization to transform documents [5, 21], audio [54, 58, 81, 95, 102], or video [6, 96] into text summaries. Text simplification also uses an abstractive method to replace complex language with simpler alternatives [20, 103]. However, abstraction sacrifices original phrasing and style for improved conciseness and coherence [98]. We flexibly combine extractive and abstractive text summarization for audio summarization to condense speech while preserving its original content and style.

2.3 Audio Summarization

Prior work explored abstractive and extractive text summarization for audio summarization to let audience members consume, skim, and navigate audio recordings efficiently. Prior work explored two main approaches to summarize audio: generating text summaries from audio transcripts [54, 81, 102] or directly editing audio to create shortened audio output [6, 22, 69, 95]. For example, Li and Chen et al. [54] support people quickly skim long audio by first transcribing the recording, then generating high-level text summaries that users can drill into progressively longer and more detailed summaries. However, abstractive text summaries of audio can only be converted back to audio via full speech re-synthesis [6, 44]. For example, Lotus [6] transforms long-form videos into short-form videos by transcribing the audio, generating an abstractive transcript summary with an LLM, then synthesizing the text summary using voice cloning [24, 70]. Although full re-synthesis through voice cloning produces speech with high speaker similarity, it strips linguistic (e.g., word choice) and para-linguistic features (e.g., volume, pace, tone, pitch, pauses), and thus speaker identity [45, 52].

Prior automatic approaches that directly edit the audio use extraction to preserve speaker identity by reusing existing speech segments through filler word removal [22], word-level edits [69], or sentence-level extraction [95]. Descript [22] automatically removes

a set list of filler words (e.g., “um”, “uh”) from speech recordings. Rescribe [69] reduces speech length by automatically removing less important sentences and words within sentences while preserving grammar. ROPE [95] extracts speech based on transcript sentences similar to an abstractive transcript summary to preserve core content. Further, extractive approaches can cause grammatical and audio errors when combining disjoint sentences, abrupt topic transitions, and miss opportunities to shorten across sentences [30, 95, 98]. Our method directly edits the audio for audio summarization and combines extraction and abstraction to preserve original speaker style while maximizing content condensation by generating mostly-extractive summaries that use abstraction and thus synthesis only when necessary.

3 TALKLESS

We present TalkLess, a system that supports automatic shortening of raw speech recordings. We designed TalkLess according to design goals that we derived from our review of prior work.

3.1 TalkLess Design Goals

Creators and editors spend time and effort to edit raw speech recordings to make them clear and engaging. We aim to design a transcript-based editing system that supports editors and creators in editing raw speech recordings (e.g., class recordings, podcast episodes, or interviews) using text to make them more concise while preserving their original content and style:

G1: Support surfacing and removing unnecessary speech. Not all speech carries equal communicative importance. Creators thus remove unnecessary speech, but distinguishing essential from non-essential speech is tedious and time-consuming [7]. While scripted speech or written text is carefully crafted to convey a message, unscripted or lightly scripted speech frequently contains disfluent speech [45, 47, 76] and tangential content [76, 95]. For example, Jurafsky et al. [45] reveal common disfluencies: filler words (e.g., “But, uh, that was absurd”), word fragments (e.g., “A guy went to a d-, a landfill”), repetitions (e.g., “it was just a *change of, change of* location”), and restarts (e.g., “it’s -, I find it very strange”). Prior work also showed that creators want to remove tangential content (e.g., a side story, meta discussions) from speech [76, 95]. We aim to support editors in surfacing and removing unnecessary speech.

G2: Preserve important information. After editing speech, the speech should retain the core message. We thus aim to achieve high coverage of important information, similar to prior work [95]. We prioritize removing speech that is unlikely to change the meaning of the speech and use abstraction to achieve higher coverage for a similar length [65]. We then let creators manually make any substantial changes to the information provided.

G3: Preserve speaker style. Speakers vary in how they use linguistic (language use and construction) and para-linguistic (volume, pace, tone, pitch, pauses) variations in their speech [15, 45]. Speakers consciously and unconsciously use such variations to guide audience attention (e.g., raise volume for a key point) [25, 33] and reinforce their identity [15, 79] as such styles inform how speakers

are perceived [26, 77, 78]. For example, speakers may pronounce “something” as “somethin” due to regional ties, frequently use a specific set of words (e.g., “wonderful” correlates to agreeable personalities [99]), or use selected vocabulary for core concepts (e.g., “affordance”). We aim to preserve speakers’ linguistic and para-linguistic styles while shortening to retain their identity.

G4: Support granular control and efficient review of automated edits. Compared to manual editing, automating speech edits shifts the burden from producing to reviewing the edits. We aim to reduce the effort required to review edits by providing flexibility in the level of edits to review (e.g., global, paragraph, fact), and supporting skimming the edits with visualizations.

G5: Avoid audio errors. Editing speech occasionally introduces distracting audio errors. For example, removing a word risks introducing artifacts (e.g., cutting in the middle of the word, disrupting background noise) [30, 69, 95] and the loss of speaker style (G3), for example due to unnatural pacing or missing breaths. We aim to avoid such audio errors.

3.2 TalkLess Interface

TalkLess’s editing interface consists of a *Compression Pane* that lets creators edit speech wording, an *Outline Pane* that lets creators edit speech by content, and an *Audio Pane* that lets creators preview their results (Figure 1).

3.2.1 Compression Pane. In TalkLess’s compression pane, editors can dynamically shorten their speech on the transcript-level or paragraph-level across four target compressions (15%, 25%, 50%, and 75% compression). Editors choose a compression using the slider on the top to globally shorten the transcript across paragraphs. Choosing a target compression on sliders next to each paragraph compresses only the respective paragraph. The two-level compression approach lets editors control what parts of the transcript to compress more or less, depending on their editing purpose (G1).

TalkLess’s compression pane features three transcript views to refine and review system edits: *Edit Types*, *Diff*, and *Final* (Figure 2). Each view allows manual transcript-based editing on the original speech transcript. While system edits are accepted by default and rendered in the compression pane, editors can manually refine and override system edits by toggling words they want to keep or remove from the audio, or inserting new words to be synthesised between existing words. While all views render cuts and insertions made in the audio, each view visualizes system edits at different levels of detail (G1, G4, G5).

While some edits are less risky and can be removed from the speech without significant effect (i.e., removal of filler words or repetitions), others are more risky as they may change the meaning or remove content from the original speech (i.e., clarifications or information removal) [45]. The **edit types view** supports editors in understanding the effect of edits by automatically classifying edits into 5 types: (ordered by risk level): *Filler Word Removal* (G1), *Repetition Removal* (G1), *Emphasis Removal* (G3), *Clarification* (G3), and *Information Removal* (G4). *Filler Word Removal* removes unnecessary words or phrases that do not add meaning to the sentence, such

Final View

we take a radiopharmaceutical, a radioisotope, called technetium, which changes shape, and produces a photon, a little pocket of light, that we can measure. we combine technetium with HMPAO, a medicine easily taken up by brain cells We inject the combination into your arm, called a first pass extraction, with 70% taken up in your brain in about two minutes. the hardest part is

Deletion Insertion

Diff View

so-what-we-do-is we take a radiopharmaceutical, so-you-take a radioisotope, we-take one-we-use, it's called technetium, which and technetium has self-esteem problems, it doesn't like being who it is, and it changes shape, and when it changes shape, it produces a photon, or a little pocket of light, that we can measure. So we combine technetium with HMPAO, a medicine that's easily taken up by brain cells We inject in the combination brain, combine them, inject them into your arm, and it's called a first pass extraction, with so 70% of it is taken up in your brain in that first pass, so within about two minutes. And then, so, it's the hardest part is of the procedure,

Deletion Insertion

Edit Types View

so-what-we-do-is we take a radiopharmaceutical, so-you-take a radioisotope, we-take one-we-use, it's called technetium, which and technetium has self-esteem problems, it doesn't like being who it is, and it changes shape, and when it changes shape, it produces a photon, or a little pocket of light, that we can measure. So we combine technetium with HMPAO, a medicine that's easily taken up by brain cells We inject in the combination brain, combine them, inject them into your arm, and it's called a first pass extraction, with so 70% of it is taken up in your brain in that first pass, so within about two minutes. And then, so, it's the hardest part is of the procedure,

Filler Word Removal Repetition Removal Emphasis Removal Clarification Information Removal

Figure 2: A speech snippet displayed in all three views within the compression pane: The final view displays the final edited transcript with rendered cuts, the diff view renders both inserted and deleted parts of the transcript, and the edit types view supports skimming and reviewing of automated edits through edit types.

as “um,” “like,” “you know,” or “basically.” These words are often used as verbal pauses and can be omitted without changing the overall content. **Repetition Removal** removes repeated words, phrases, or ideas that add no new value to the speech to increase clarity and conciseness. **Emphasis Removal** removes words or phrases used to stress or exaggerate a point. These may include words like “really” and “very”. **Clarification** makes an unclear or ambiguous phrase more understandable by providing additional context or rephrasing the phrase. **Information Removal** removes details, facts, or sections of the content that are either irrelevant to the main information or not relevant in a different context.

The **diff view** is similar to the edit types view, but displays cuts in gray and insertions in light purple to highlight the potential audio effect rather than the semantic effect of the edits (G5).

Editors use the **final view** to read the final transcript to assess the speech flow after the edits occur (G4, G5). The final view is the default view and displays the final edited transcript, hiding any removed text from the original transcript and indicating cuts in gray (similar to prior transcript-based audio editing tools that display cuts only [22, 76]).

3.2.2 Outline Pane. The outline pane lets editors navigate, locate, and edit content included in the original speech transcript (G1, G4). The outline pane shows content that occurs in the original speech and groups the content into semantically meaningful groups.

Each group is linked to a paragraph in the *Transcript View* and can be further broken down into more fine-granular information that occurs within the paragraph. Editors can collapse or expand these groups for quick navigation. Hovering over any element in the outline makes the *Transcript View* scroll to and highlight the respective paragraph or aligned low-level outline element.

To support editors in skimming for content to remove or keep, each element in the information sidebar consists of a blue stripe on the left indicating its importance to the overall speech (G1). The importance is visualized from least important (light blue) to most important (dark blue). Editors can guide the importance visualization by typing in a speech purpose at the top of the outline pane. For instance, when re-purposing a class lecture to be uploaded on YouTube, the editor may specify “lecture uploaded to YouTube for a general audience” as the purpose. The system then re-evaluated each sidebar element’s importance according to the specified purpose. Each sidebar element shows a percentage indicator depicting how many of the keywords of each information element are still included within the final edited transcript to help editors assess what information is being lost as a result of their edits or what information could further be removed (G5).

3.3 Algorithmic Methods

TalkLess uses a *speech shortening pipeline* that transcribes, aligns, and segments the original speech recording, then generates shortened segment candidates, and after choosing and concatenating the best candidates, automatically edits the audio using an audio editing model. TalkLess *classifies edit types* based on the generated results to make edits easier to skim and review (G4). Finally, we *extract information* that occurs in the original speech recording and present each bit of essential information together with a computed importance score to the user within the outline sidebar (G2).

3.3.1 Automatic Speech Shortening Pipeline. Our automatic speech shortening and editing pipeline consists of 2 steps: After transcribing the input audio and aligning each spoken word with the transcript, we create a shortened target transcript during the **transcript shortening** step, then cut and synthesize based on the transcript to receive the final shortened audio during the **automatic audio editing** step (Figure 3).

Transcript Shortening. After pre-processing the audio, we prompt GPT-4o [66] to segment the transcript into semantically distinct segments $S_1, S_2, \dots, S_n$ based on the speech’s content. Early experiments demonstrated our pipeline works best when compressing smaller segments rather than the full transcript (similar to prior text summarization approaches [35, 55]).

For each target compression $\tau \in \{0.15, 0.25, 0.5, 0.75\}$ and segment S_i , we generate a set of 25 shortened candidate transcripts $C_{i,1}, C_{i,2}, \dots, C_{i,25}$ by providing GPT-4o [66] with each segment and a prompt that describes eight desired characteristics for shortened transcripts (A.1.1) based on our design goals, e. g., removing filler words and repetitions (G1), preserving original information (G2), style (G3), and unique words (G3), limiting word insertions (G3), and not changing word spellings (G3, G5).

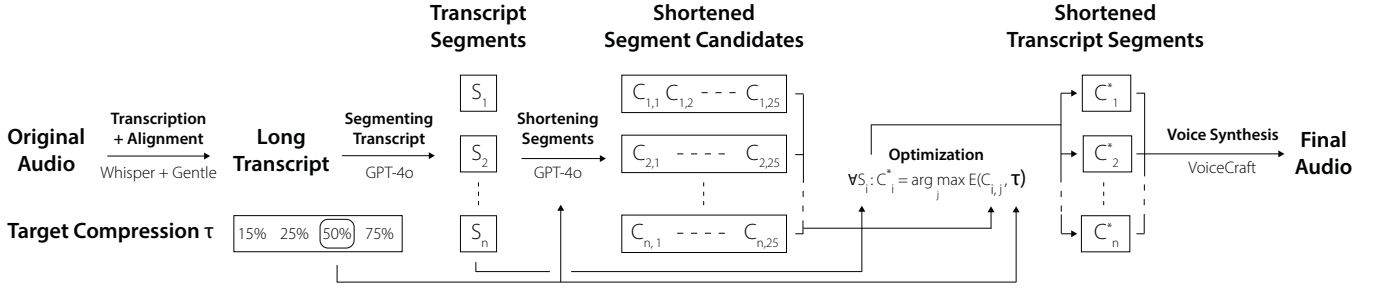


Figure 3: TalkLess takes an original audio and a target compression ratio as input. TalkLess first transcribes and aligns the audio into a transcript. The long transcript is then segmented into transcript segments $S_1, \dots, S_n$. For each segment S_i , TalkLess generates 25 shortened segment candidates $C_{i,1}, \dots, C_{i,25}$. For each segment S_i , we choose the best candidate C_i^* based on compression, number of edits, edit lengths, and coverage. Finally, TalkLess takes the best candidates $C_1^*, \dots, C_n^*$ as input to its voice editing model (VoiceCraft [70]) to generate the final shortened audio.

Early experiments revealed that prompting an LLM to generate transcripts that follow our design goals (G1-G5) and target compressions (15%, 25%, 50%, 75%) produces unreliable results and thus creates an unpredictable user experience (Table 3). Hence, TalkLess uses an optimization-based approach to generate reliable results.

To automatically select the optimal candidate for each segment, we score each candidate transcript $C_{i,j}$ based on a target compression τ using a candidate evaluation function $E(C_{i,j}, \tau)$ that combines four metrics based on our design goals:

$$E(C_{i,j}, \tau) = \lambda_1 \cdot E_{comp}(C_{i,j}, \tau) + \lambda_2 \cdot E_{edits}(C_{i,j}) + \lambda_3 \cdot E_{len}(C_{i,j}) + \lambda_4 \cdot E_{cov}(C_{i,j})$$

We empirically determined weights between 0 and 1: $\lambda_1 = 0.4$, $\lambda_2 = 0.15$, $\lambda_3 = 0.1$, $\lambda_4 = 0.35$. To receive the desired compression level, we compute $E_{comp}(C_{i,j}, \tau)$ which is the difference between the candidate and target compression τ , calculated by computing the ratio of the word length of the shortened candidate $C_{i,j}$ to the word length of the original segment S_i :

$$E_{comp}(C_{i,j}, \tau) = 1 - \left| \frac{\text{length}(C_{i,j})}{\text{length}(S_i)} - \tau \right|$$

For example, for 75% compression, $E_{cov}(C_{i,j})$ may generate:

The quick ~~brown~~ fox jumped over the ~~sleeping~~ ~~lazy~~ dog ~~sleeping quietly~~ in the ~~sunny~~ garden.

To minimize audio errors, we compute $E_{edits}(C_{i,j})$, which is the number of edits required to get from the original audio to the final edited audio (G5). We use the Needleman-Wunsch [84] algorithm to receive a list of edit instructions that describe the differences between the texts of $C_{i,j}$ and S_i . These instructions are then grouped into chunks of consecutive edits, with each chunk representing a single edit operation as performed by the voice editing model. The total number of edit operations is normalized by the maximum possible number of edits, defined as half the word count (*i.e.*, every second word is edited) of the original segment (*e.g.*, “The quick brown fox jumps over the lazy dog” has at most 4 edits):

$$E_{edits}(C_{i,j}) = 1 - \frac{\# \text{ of edits}}{(\text{length}(S_i)/2)}$$

Adding $E_{edits}(C_{i,j})$ now reduces the number of edits required:

The quick brown fox jumped over the ~~sleeping~~ ~~lazy~~ dog ~~sleeping quietly~~ in the sunny garden.

We avoid long speech synthesis that risks stripping speaker identity (G3) and introducing artifacts (G5), by computing $E_{len}(C_{i,j})$ which is the mean length of insertions within a segment $C_{i,j}$:

$$E_{len}(C_{i,j}) = 1 - \frac{\sum (\text{length of insertions})}{\# \text{ of insertions}}$$

Adding $E_{len}(C_{i,j})$ now removes the insertion:

The quick brown fox jumped over the lazy dog ~~sleeping quietly~~ in the sunny garden.

To maximize preserving important information (G2), we compute $E_{cov}(C_{i,j})$ by matching each sentence $s \in S_i$ to the most similar sentence $c \in C_{i,j}$, using a sentence transformer model (all-mpnet-base-v2 [40]), and then averaging the best matches to get a final coverage score:

$$E_{cov}(C_{i,j}) = \frac{1}{|S_i|} \sum_{s \in S_i} \max_{c \in C_{i,j}} \text{sim}(s, c)$$

After adding $E_{cov}(C_{i,j})$, the information that the dog was asleep is now included:

The quick brown fox jumped over the ~~lazy~~ dog ~~sleeping quietly~~ in the sunny garden.

Finally, for each segment S_i , we select the candidate with the highest score $C_i^* = \arg \max_j E(C_{i,j})$ and concatenate all C_i^* to construct the final shortened transcript.

Automatic Audio Editing. We generate the final shortened audio based on the original audio and the shortened transcript. We use the list of edit instructions received from applying the Needleman-Wunsch algorithm [84] between the original transcript and the final shortened transcript. TalkLess supports three types of edits: (1) **deletions**, or removing segments from the original audio, (2) **insertions**, or adding new segments to the original audio, and (3) **replacements**, or replacing existing segments with new ones.

For every **deletion** (“...to figure out ~~how to weather or~~ how to maintain...”), we identify the start and end timestamps of the segment to remove and cut it from the audio. To avoid noticeable cuts in the audio, we synthesize a smooth cut using VoiceCraft [70], an

open-source speech editing model. We use the 10 seconds of the original audio that surround the edit as input to the model, as we found this is the ideal length to match the speaker’s original speech characteristics and not fall out of the model’s training distribution. Because natural speech blends adjacent words, we re-synthesize half of the word before and after the cut. To reduce synthesis artifacts, we align transition start points with natural pauses when possible, and choose transitions shorter than 0.6 seconds, or the shortest among 5 generations, whichever is first (G5).

We synthesize each **insertion** (“...everyone *has* to sort of make things work...”), using the same approach as generating transitions for deletions. To avoid bad generations, we predict the expected generation length based on the mean phoneme length surrounding the edit, then choose the closest after generating 5 generations.

We define **replacements** as insertions directly followed by deletions (“...books, but *try* they’re trying to send money...”) and treat them like insertions after removing the deleted part from the audio. Each edit generation and the preceding original transcript segment that remains unchanged are appended to the previous edit output, gradually creating the final shortened audio.

3.3.2 Classification of Edit Types. For each segment and compression level, we use the list of required edits received by applying the Needleman-Wunsch algorithm [84] within the Speech Shortening Pipeline (Section 3.3.1), to classify each of the five edit types (G4). For each edit, we use GPT-4o [66] to classify its edit type by prompting it with the text before and after the edit and definitions of edit types (see prompt in A.1.2).

3.3.3 Extracting Information and Evaluating their Importance. For each segment of the transcript, we extract the atomic pieces of information into bullet points using a GPT-4o prompt (see A.1.4). This approach is inspired by prior work on automatic outline generation [32] and atomic fact extraction [49, 90]. We create information groups by generating a summary of all information pieces included within a transcript segment. For each information piece, we evaluate its importance by prompting GPT-4o to generate an importance score between 1 and 10 based on the information piece, the original transcript, and a user-specified purpose (G2). This prompt can be found in A.1.3. As LLM-generated scoring tends to mostly output scores in the range of 7 to 10, we redistribute the scores to be between 1 and 10. Each information group is assigned the mean importance score of all the information pieces it includes.

3.4 Transcript Shortening Ablation

We compared and analyzed shortened transcripts (*i. e.*, target compression deviation, portion of words synthesized, and coverage) generated by two methods: (1) using an LLM only with our prompt (see A.1.1) and (2) using our same LLM prompt combined with our optimization-based approach. We generated 25 transcripts with each method for 4 audio files and all 4 target compressions (15%, 25%, 50%, 75%). Our experiment revealed that the LLM-only outputs deviated from the compression target by 0-73 percentage points ($\mu = 22$, $\sigma = 5$). In contrast, TalkLess deviated from the compression target by 1-3 percentage points ($\mu = 2$, $\mu = 1$). LLM-only also risked synthesizing more transcript words, ranging from 0-43% ($\mu = 10\%$, $\sigma = 0.4\%$) compared to 0-22% ($\mu = 3\%$, $\sigma = 1\%$) with TalkLess,

Evaluation Metric	Description
SPEECH DISFLUENCIES	The combined count of <i>filler words</i> (e.g., “uh”, “um”) and <i>repetitions</i> (e.g., “[...] in my, in my perspective [...]”), counted using a speech transcript linked to the audio [22].
COHERENCE ERRORS	Edit-induced errors that disrupt natural flow (e.g., abrupt transitions through cuts in the middle of a word) and inconsistencies in speech (e.g., missing co-references), counted by listening to the audio.
COVERAGE	Percentage of content from the original transcript retained in the shortened version, measured by computing cosine similarities between sentence embeddings of the transcript and summary, selecting the highest similarity per transcript sentence, then averaging these scores.
STYLE PRESERVATION	Cosine similarity between content-independent style representations [97] for transcripts generated by TalkLess, the baseline, and synthesized abstractive summaries.
AUDIO PREFERENCE	Human annotator’s preference for one audio sample over another.

Table 2: We use six evaluation metrics to evaluate results generated by TalkLess and the baseline across all target compressions (15%, 25%, 50%, 75%).

and missing more information, ranging from 33% to 93% coverage ($\mu = 62\%$, $\sigma = 1\%$) compared to 65-96% coverage ($\mu = 84\%$, $\sigma = 4\%$) with TalkLess (Table 3).

3.5 Implementation

We implemented a frontend and a backend for TalkLess. The frontend was implemented using *React.js*. The backend consists of a Python Flask API, which contains most of the pipeline steps and handles their procedural flow. We used rev.ai [72] and Gentle forced-alignment [60] to transcribe and align the input audio. We also hosted the open-sourced VoiceCraft [70] speech editing model on a server with an A40 GPU (48 GB VRAM).

4 TECHNICAL & RESULTS EVALUATION

We evaluated the shortened audio results generated by TalkLess. We compared our results with those shortened using ROPE [95], serving as a baseline of previous automatic audio shortening methods.

4.1 Method

We selected a total of 4 speech recordings, across 3 categories: podcast, interview, and lecture. We selected the state-of-the-art approach for extractive audio summarization, ROPE [95], as our baseline. We re-implemented ROPE as it is not open-source and replaced ROPE’s abstractive summarization pipeline step with state-of-the-art abstractive summarization through GPT-4o.

We used five evaluation metrics (Table 2) to evaluate results generated by TalkLess and the baseline for all target compressions (15%, 25%, 50%, 75%): SPEECH DISFLUENCIES, COHERENCE ERRORS, COVERAGE, STYLE PRESERVATION, and AUDIO PREFERENCE. We compute compression rates based on the number of words ratio between the original and the shortened transcripts (*i. e.*, a 20% compression means that the text has been shortened by 20%). We also include the original audio for SPEECH DISFLUENCIES and COHERENCE measures. To measure how well TalkLess and the baseline preserve linguistic style compared to full abstraction combined with voice cloning, we include a synthesized abstractive summary for each compression to

evaluate STYLE PRESERVATION. We generate abstractive summaries by prompting GPT-4o [66] to shorten the transcript to the same target compression ratios without the desired characteristics used in TalkLess's approach.

We recruited 12 human evaluators from our organization to rate their AUDIO PREFERENCE between results generated by TalkLess and the baseline based on all 4 original audio files in an online survey. We counterbalanced the order of comparisons among all evaluators who were unaware of the condition. We analyzed all results using pairwise t-tests and repeated-measures ANOVAs.

4.2 Results

Overall, shortened results generated by TalkLess had significantly higher coverage (15%, 25%, 75%), fewer speech disfluencies and coherence errors, and were significantly more preferred (15%, 25%) than the baseline. Across all conditions, TalkLess generated audio with a mean deviation of 3% ($\sigma = 3.66\%$) from the specified target compression rate. Generated results contained synthesized insertions with a mean length of 1.45 words ($\sigma = 0.03$) and no insertion was longer than 5 words. TalkLess's results contained fewer speech disfluencies than those of the baseline across all target compressions (Figure 4). The baseline results contain more than twice as many disfluencies as TalkLess's when the compression rate is low, and only reach a comparable count when the compression rate is very high. This is because TalkLess's word-level compression allows targeted removal of speech disfluencies while sentence-level compression in the baseline preserves speech disfluencies in extracted sentences. The baseline produced significantly more coherence errors than TalkLess across all target compressions ($p < .01$, Figure 4 right).

Figure 5 shows that results generated by TalkLess covered significantly more information of the source recording than the baseline for 15%, 25%, and 75% target compression rates ($p < .05$).

TalkLess's results were significantly more preferred than those of the baseline for 15% and 25% target compressions ($p < .05$, Figure 6). TalkLess's results were more preferred than those of the baseline for higher compressions, yet there was no significant difference. Hence, our approach is more suitable for information-rich scenarios where one needs to have a low target compression between 0% and 25% to retain most of the information in the original speech. Investigating preferences for individual audios revealed that TalkLess is preferred more often when the original audio contains more disfluencies, incoherence errors, and is more informal. Thus, TalkLess may be more useful for spontaneous and unscripted speech.

We found a significant difference in METHOD ($F_{1,07,5,37}, p < .01$) on text style cosine similarities between transcripts generated by each method and the original transcript across compression targets (Figure 11). Across all compression targets, TalkLess's ($p < .05$) and the baseline's ($p < .01$) transcript style were significantly closer to the original transcript style than abstractive transcripts.

5 USER EVALUATION

We evaluated TalkLess's editing interface to investigate its ability to facilitate shortening speech recordings.

5.1 Method

5.1.1 Participants. We recruited 12 participants (8 male and 4 female, ages 21 to 39) from within our organization and from outside using email lists. All participants had prior experience in audio

and/or video editing ($\mu = 3.42$ years, $\sigma = 2.64$ years), and 4 (P2, P5, P6, and P11) had professional experience in audio editing.

5.1.2 Baseline and study design. We conducted a within-subjects study where we compared shortening raw speech recordings with TalkLess's interface to a baseline interface. The baseline interface is a restricted version of TalkLess's interface where editors can only choose from global compressions generated by the state-of-the-art extractive audio summarization approach (*i. e.*, ROPE [95]) and do manual refinements in the *diff view* and *final view* only (Figure 13). ROPE provides a strong baseline as it extracts important moments, encourages coherence with clause-level edits, and follows compression targets. We also evaluated an extractive approach using an LLM that used our base prompt (A.1.1) but modified to enforce strict extraction with no inserted or modified words. However, unlike ROPE, LLM-only extraction was unreliable as it did not adhere to extraction constraints or target compressions (similar to LLM-only in Table 3). We updated ROPE's abstractive summary generation step to use GPT-4 [66] and modified ROPE's interface to allow word-level edits and dynamic compression between the same target compression rates as in TalkLess (Figure 13). The study has 4 conditions (TALKLESS vs. BASELINE) $\times$ (AUDIO 1 vs. AUDIO 2). For AUDIO 1 and AUDIO 2, we selected two history lecture recordings from the same speaker with similar length, as those can be consumed without visuals and contain raw unedited speech [11, 12]. We counterbalanced conditions among all participants to alleviate ordering effects.

5.1.3 Procedure. During the study, participants completed a demographic profile questionnaire and a consent form (5 minutes). We then introduced the study and gave participants a tutorial for both interfaces (20 minutes). The participants complete two editing tasks (2 $\times$ 20 minutes = 40 minutes). After each editing task, participants completed a questionnaire. The between-task questionnaire includes Likert rating scales extracted from the NASA TLX [38], the Creativity Support Index (CSI) [18], and the System Usability Scale (SUS) [10]. After completion, we interviewed participants (30 minutes). Participants were compensated with \$30 via PayPal or Venmo and each study session lasted approximately 90 minutes.

5.1.4 Analysis. We analyzed task load, creativity support, and usability ratings using paired Wilcoxon Signed Rank tests. We apply Greenhouse-Geisser correction when the equal-variances assumption is violated (Mauchly's test $p < .05$). We analyzed participants' interaction data using multiple repeated-measures ANOVAs.

5.2 Results

All participants preferred TalkLess over the baseline interface and expressed enthusiasm for using it in the future to polish audio to make podcasts and recorded presentations easier to follow (P1, P4, P7, P10), re-purpose audio to adapt it to new audiences (P3, P4, P8), or summarize audio to create concise summaries of lengthy lecture recordings or information-rich content (P4, P7). Participants noted that TalkLess saved time and effort, allowing for an efficient workflow of "*first record, then edit to ideal transcript*" (P3). Non-professionals also saw value in TalkLess for improving clarity when editing their own recorded speech (P3, P7, P10).

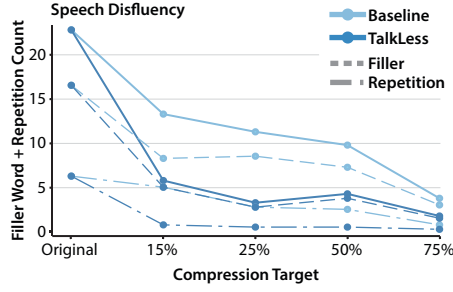


Figure 4: SPEECH DISFLUENCIES and COHERENCE ERRORS. Dotted lines represent filler words and repetition counts.

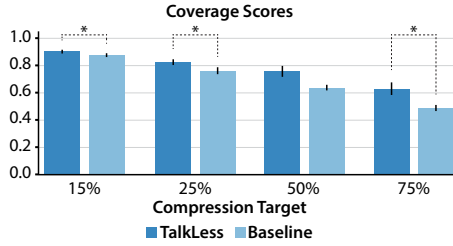
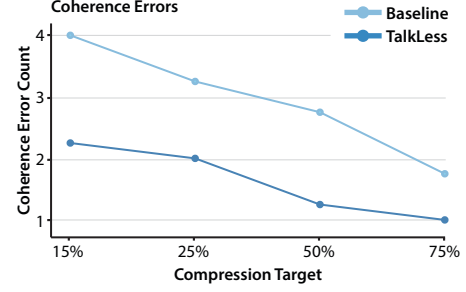


Figure 5: Comparison between coverage for each compression level for TalkLess and the baseline. Error bars represent a 95% confidence interval. * $p < .05$

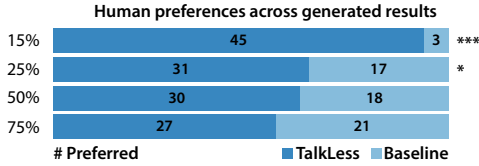


Figure 6: Human preferences. The numbers represent the number of times a method is preferred for each compression. * $p < .05$, *** $p < .001$

5.2.1 User Interactions & Subjective Ratings. Participants had similar compression rates with both methods ($\mu = 0.58$, $\sigma = 0.11$ with TalkLess compared to $\mu = 0.58$, $\sigma = 0.15$ with the baseline). However, participants manually edited significantly fewer words in TalkLess ($\mu = 477.55$, $\sigma = 394.58$) than in the baseline interface ($\mu = 807.45$, $\sigma = 352.75$) ($p < .05$). Participants frequently switched between views in TalkLess to verify and perform edits (Figure 12). While the diff view was initially active in TalkLess, all but one participant switched to the edit type view within the first third of the session. Near the end of the study, 5 participants switched to the final view to verify their final edited transcript. Although participants started on the diff view, they spent significantly more time on the edit types view ($\mu = 12.88$, $\sigma = 2.92$) than on both the diff view ($\mu = 2.35$, $\sigma = 3.28$; $p < .001$) and final view ($\mu = 1.35$, $\sigma = 2.17$; $p < .001$) (Figure 8).

Participants reported significantly lower mental demand, temporal demand, effort, and frustration when working with TalkLess compared to the baseline interface ($p < .05$, Figure 7 left). Using TalkLess, participants found it significantly easier to explore many ideas, felt more engaged in the activity, felt that what they were

able to create was more worth their effort, and were significantly more expressive than in the baseline interface ($p < .05$, Figure 7 middle). Participants rated TalkLess to be significantly easier to use and faster to learn than the baseline ($p < .05$, Figure 7 right). Participants want to use TalkLess significantly more frequently in the future ($p < .01$). We report the means and standard deviations for task load, creativity support, and usability ratings in Table 4.

5.2.2 TalkLess vs. Baseline Compression. All participants preferred the compression of TalkLess over the compression of the baseline. For instance, due to TalkLess’ ability to “connect sentences” (P4) or “make it more concise” (P12). P3 preferred TalkLess’s compression because it is “very common that half [of a sentence] is useful and the other half is not.” Participants found the initial compression levels in the baseline interface less helpful and all participants wished for more flexible edits in the baseline interface, as P4 described:

“The system excluded many sentences here, but it’s still very relevant... I am thinking how I can revert it back. It is hard to notice all filler words and delete them.”

P1 remarked, “I cannot add words instead of other words,” when attempting to replace a repeated term with a pronoun, noting that this limitation “makes it hard to retain content.” P11 adds that they would also insert new voice by using speech generation services, however, highlighting that “fine-tuning it” to fit into the existing speech is time-consuming. P5 comments on the ability for inserting new words within TalkLess:

“[Editing in TalkLess] is more satisfying because of the larger amount of control when editing the transcript.”

Reading their final edited transcript within TalkLess’s *Final View*, P1 reported on the edited transcript coherence, “there is this good flow.” P11 stated that TalkLess’s shortening is a “sharp tightening” where “you are not missing any content.”

P2 described that their usual process consists of “very low-level editing” and that they required time to get familiar with more high-level edits that TalkLess provided, as they “thought in too fine-granular edits”. P10 mentioned that TalkLess would be more helpful for editing unscripted speech than well-scripted speech, such as a politician’s speech. In addition to editing audio, 5 Participants (P1, P5, P9, P10, P12) mentioned they want to use TalkLess for consuming audio, for example, to “re-watch past lectures.” (P12).

5.2.3 Editing Workflows. All participants followed a consistent workflow across both interfaces. Participants first chose an initial global compression rate, then skimmed and verified system edits

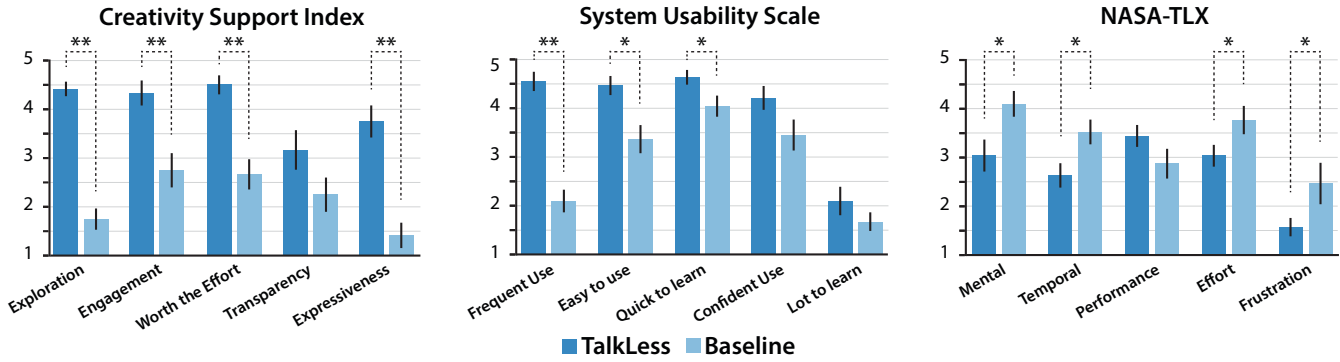


Figure 7: Participant ratings for questions selected from Creativity Support Index, System Usability Scale, and NASA TLX for TalkLess and baseline. Error bars represent a 95% confidence interval. $*p < .05$, $p < .01$**

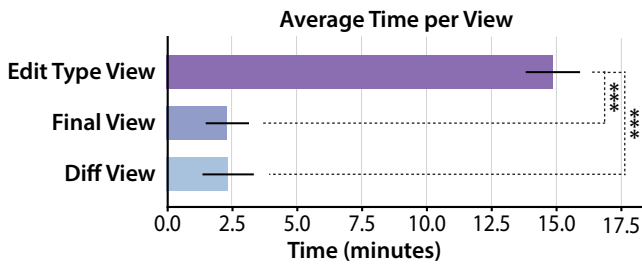


Figure 8: Users spent significantly more time on the Edit Types view than on the Final and Diff view when working with TalkLess. $*p < .001$**

segments by segment, and finally manually edited the speech to adjust content based on the editing purpose. 6 participants used the segment-level compression for local adjustments based on the segment's content relevance to the new purpose. In the baseline interface, participants manually skimmed the transcript word by word, often describing this as “tiring” (P6). In contrast, TalkLess revealed new editing strategies in participants' workflows.

Choosing an initial compression. All participants first decided on an initial global compression. In the baseline interface, participants had to skim the transcript for each compression to assess its quality. P6 expressed particular frustration, describing the compressions as difficult to review: “I don't really know what it does”.

In TalkLess, participants used the edit types view to quickly skim and review system edits. For instance, P3 noted that the edit types view allowed them to quickly “get an overview” of the edits and that it helped to choose an initial compression. P4 found the edit types view particularly useful to find a balance between compression and preserving content: “A 30% to 50% compression seems like a good balance [...] to keep the information removal very low”. P3 mentioned understanding the speech first before editing took “less effort” when using TalkLess's outline pane compared to the baseline interface.

Participants employed different strategies when choosing an initial compression. Participants either began with a high compression so that they could focus on reviewing edits and re-including critical content that the system removed or participants started

with a low compression to focus on finding further content that can be removed according to their editing purpose.

Verifying system edits. After choosing an initial compression, participants verified whether the system edits aligned with their editing goals. P11 mentioned that edit verification is part of their typical workflow and that they usually in Descript to later decide whether to remove it or not. Participants typically explored and verified edits linearly by going through the speech transcript from start to end, edit by edit. However, the strategies for this process differed between the two interfaces.

In the baseline interface, participants skimmed the entire transcript word by word to verify edits. Participants reported “jumping back up and down a lot [between words] intense” (P6). In contrast, when using TalkLess, participants prioritize edits based on perceived risk and importance to manage cognitive load when verifying system edits. Participants navigated paragraph by paragraph using the outline pane's higher-level content information, then focused on specific edits using the edit types view. P10 describes their strategy when using TalkLess's outline pane:

You can look at it [the outline], compress, and check if the sentence that got compressed is related to the main topic of the paragraph or not. And when you think it's okay, then you can listen through it, and [...] make sure it all sounds correct [...] like sculpting something.

P6 compared TalkLess to the baseline interface and highlighted that the color-coded edit types reduced their mental effort, as they could “quickly filter” low-risk edits (green) while focusing their attention on more critical edits (red) such as information removal. While some participants, like P9, initially found the color-coding “distracting,” they eventually appreciated it for verifying edits. P4 the edit types view the “most helpful part:”

It reduced my effort [...] I only need to focus on information removal and for the filler word removal I just need to take a glance and skim it.

Performing manual edits. All participants performed manual edits in both interfaces to ensure that the final edited speech aligned with their editing purpose.

Participants described manual editing in the baseline interface as “tiring” (P6) because of the need to go linearly through the whole transcript to locate and remove speech disfluencies or unnecessary content. For instance, P4 noted that “[they] went through everything” and “manually did all the work” when using the baseline interface.

In contrast, participants used TalkLess to perform edits at multiple levels. Participants used the edit type and diff views to perform edits on the word level, the outline pane to perform content-based edits, and the compression sliders for segment-level edits. Participants appreciated the segment level compression sliders because they allowed them to compress whole segments based on their content relevance. P6 noted that the outline pane “strategically breaks long sentences down,” into “thought threads.” Participants used the outline pane to identify less-important sections of the transcript that can be edited. P4 compared their strategy using the outline pane to their strategy within the baseline interface:

“I look at parts with lighter blue and remove the non-important parts first [in TalkLess] instead of going through the paragraphs linearly [in the baseline].”

Participants highlighted the alignment between TalkLess’s importance indicators and their own judgment. P3 felt that the system’s suggestions helped validate their manual edits: “I find it cool that the importance agrees with me when deleting.” However, while some participants, like P4 and P7, actively used the importance toggle to find additional content to cut by rendering less important content in gray, others engaged with this feature less frequently (Figure 12).

5.2.4 User Concerns. All participants preferred TalkLess over the baseline interface. However, some participants had difficulties transitioning between TalkLess’s views (P3, P9, P12). For instance, P3 mentioned that it was “difficult to switch” between views when losing their place in the transcript after switching from the diff view to the final view. We have since improved the interface to maintain cursor and scroll position when shifting views. Although participants appreciated having more editing control in TalkLess, participants raised concerns about being able to change a speaker’s speech by making them “say something they never said.” (P10). P5 and P10 highlighted the ethical implications of editing speech to potentially misrepresent the speaker’s intent or meaning, such as removing words of the professor’s speech to “make it feel like [the professor] didn’t care for [their] class” (P10). Although TalkLess’s outline pane is aligned with the content in the compression pane to make editors aware of modifications in narratives, future iterations can flag edits that substantially change the original text’s meaning.

6 EXPLORATORY STUDY

The comparison study demonstrated that editors preferred TalkLess over an existing extractive audio summarization baseline [95] for creating concise speech recordings. Prior work indicates that synthesized speech can be indistinguishable to listeners from human speech for unfamiliar speakers [61, 70, 74]. Thus, creators could theoretically use abstraction only for rewriting, then synthesize their entire audio without impacting audience perception. However, it remains unclear whether creators themselves are comfortable with such extensive synthesis of their own speech. As a result, we compare TalkLess with the strongest abstractive audio summarization baseline we identified (GPT-4 [66] + ElevenLabs [24]).

We conducted a 60 minute exploratory study with three creators C1–C3¹ who had more than 1 year of audio editing experience (for social media videos, podcasts, or class projects). Creators first recorded a 2–3 minute speech recording², then edited their speech using TalkLess. We provided TalkLess result and a fully synthesized version of the same result using state-of-the-art voice cloning [24].

When comparing TalkLess’s result with the fully synthesized speech for the same transcript, all creators strongly favored TalkLess’s result as it preserved most of their authentic speech characteristics. Creators described fully synthesized results as robotic (C1, C2, C3), monotone (C1, C2), and lacking personality (C2). C1 noted “the way I talk is something I want to include,” and mentioned that fully synthesized speech mispronounced uncommon terms in their speech, such as names of climbing walls. C3 said that for synthesized speech “I know it’s not me,” and they preferred TalkLess because it preserved a “sense of authenticity” as “the idea is to reuse and not regenerate” speech. While creators preferred TalkLess’s result and found most of TalkLess’s edits to be seamless, all creators noted at least one audible cut. All creators liked the automatic edits suggested by TalkLess and made only minor edits (all inserted and removed at least one word). C1 expressed a desire for additional controls, such as the ability to manually insert pauses, and C3 described that their openness to synthesized edits depended on the context — synthesis was appropriate for casual scenarios and audience who did not know their voice (e.g., posting to TikTok) but not for integrity-sensitive scenarios (e.g., formal class assignments). Overall, all creators wanted to use TalkLess in the future to edit their speech recordings to balance authenticity with conciseness.

7 DISCUSSION

7.1 Reflection on Results

Our evaluations demonstrate that TalkLess’s algorithm and interface effectively address our design goals.

TalkLess Algorithmic Results: Compared an extractive baseline [95], TalkLess produces shortened speech that preserves significantly more content for 15%, 25%, and 75% compression targets and removes significantly more speech disfluencies while having significantly fewer coherence errors (**G1** – surface and remove unnecessary speech, **G2** – preserve important info, **G5** – avoid audio errors). TalkLess preserves transcript style similarly as well as our extractive baseline and significantly better than a fully abstractive baseline (**G3** – preserve speaker style). Human evaluators preferred TalkLess’s outputs significantly more than baseline outputs for target compressions of 15% and 25%, while preferences were similar for higher compression rates that demanded more cuts and insertions. Creators in our exploratory study unanimously preferred editing their own speech with TalkLess over complete re-synthesis, favored re-using original speech for authenticity (**G3**).

TalkLess Interface Results: All participants in our user evaluation preferred TalkLess over the baseline text-based editor with extractive summary results. With the baseline, participants edited

¹We recruited audio creators from our professional networks. One creator was male, two were female, and creators’ ages ranged from 21–23.

²Creators selected a topic from a list of ten (see A.2)

by manually skimming through the transcript word by word, which was perceived as tedious and time-consuming. Participants thus rated TalkLess as significantly easier and faster to use, less mentally demanding, and more engaging than the baseline due to TalkLess's visual skimming and reviewing aids that aligned with participants' editing strategies (G1, G2, G4). The edit types view, importance indicators, and outline pane specifically allowed participants to quickly identify and prioritize edits such that they could focus on meaningful content modifications rather than low-level manual edits. TalkLess thus significantly reduced manual edits TalkLess (average of 478 manually edited words in TalkLess compared to 807 words in the baseline) for a similar level of compression.

Use Considerations: Our evaluations also revealed ethical concerns that TalkLess has the potential to unintentionally alter a speaker's intended meaning when inserting or rearranging content, and context-specific acceptance or rejection of voice synthesis.

7.2 Limitations

We used ROPE [95] for sentence-level extraction in our baseline interface (Figure 13). While a well-engineered LLM-based extractive approach could surpass our baseline, developing such a system was beyond the scope of this work. Although VoiceCraft [70] represents the state-of-the-art in speech editing, it operates at low fidelity (16kHz sample rate) and can have unstable performance, such as dropped or distorted words during generation, which may negatively affect audio quality at higher compression rates (G5). Future speech editing models with higher fidelity and stable generation would increase the quality of results produced by TalkLess. TalkLess segments the original transcript into segments and compresses segments individually. To prevent missing global context and important cross-segment content relationships, we plan to explore hierarchical summarization techniques [55] in the future. Finally, more optimal transcript edits may be identified by combining edits across edit candidates or generating more edit candidates and we will expand our candidate generation approach in the future.

7.3 Extensions

We demonstrate two extensions of TalkLess that preview application scenarios and future directions.

Extractive-Abstractive Speech Editing Spectrum. Prior work explored the spectrum between extractive and abstractive methods for text [98]. Our work opens the opportunity to explore a similar spectrum for extractive and abstractive speech editing to balance preserving linguistic and para-linguistic speaker style with optimizing information compression, fluency, and clarity. To support flexible editing strategies, we extended TalkLess with a slider embedded directly within each paragraph in the compression pane (Figure 9). Editors can select among four predefined editing modes: *Extractive* (Figure 9.A) to preserve maximum authenticity, *extractive-abstractive* (Figure 9.B) to balance condensing information and preserving speaker style, *abstractive-extractive* (Figure 9.C) that uses an abstractive summary for maximum content coverage but re-uses original speech to preserve speaker style (e.g., Lotus [6] uses mostly abstraction to maximize compression but uses extraction when needed to preserve speakers' talking heads), and *abstractive*

Extractive-Abstractive Speech Editing Spectrum

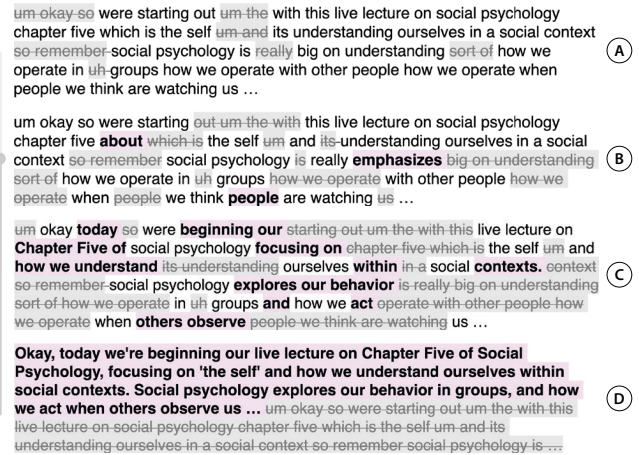


Figure 9: Extractive and abstractive speech editing creates a spectrum that allows partial blending between the two contrary editing types. Editors can choose between extractive (A), extractive-abstractive (B), abstractive-extractive (C), and abstractive (D) editing.

(Figure 9.D) which synthesises the full abstractive summary. Each mode is implemented by explicitly instructing the LLM to do edits according to the chosen mode and adjusting the scoring algorithm to consider the balance between extractive and abstractive edits.

Future work could automatically recommend optimal slider settings based on speech characteristics (e.g., spontaneous vs. scripted speech) and contextual constraints (e.g., time limits), or user-specific editing styles learned from past interactions. Additionally, we plan to investigate when editors prefer voice re-synthesis versus re-recording or re-using audio to let our system automatically adjust abstractive-extractive sliders based on the editing context.

Original:

So **remember**, a lot of these **concepts** with the **self** are **uh really** basic.

Before considering emphasis:

A lot of **concepts** with the **self** are basic.

After considering emphasis:

Remember, these **concepts** with the **self** are **uh really** basic.

Figure 10: We extended TalkLess to also preserve emphasized speech when shortening speech recordings.

Editing for Preserving Emphasis. Speakers use para-linguistic cues (e.g., volume, tone, and pauses) to emphasize key points in their speech [15, 25, 33, 45, 63]. To help speakers preserve these often intentional communicative elements, we extended TalkLess with an emphasis-preserving editing module that automatically detects and prioritizes emphasized speech. We used a neural speech prominence estimator [63] to assign each spoken word an emphasis value from 0 (no emphasis) to 1 (high emphasis), then modified the TalkLess scoring function to penalize edits that would remove

speech with high emphasis. We updated the interface to visualize emphasis with a yellow gradient to help editors surface emphasized words when reviewing edits (Figure 10). While our extended approach preserves intended emphasis (“*Remember*”), it also preserves speech errors (“*uh really*”) as speakers often unintentionally emphasize filler or transition words when they hesitate. Future work can distinguish between intentional and unintentional paralinguistic patterns to preserve emphasis only where it is meaningful.

7.4 Future Work

Implications for Speech Editing. TalkLess combines extractive speech editing [69, 95] with synthesized abstraction using voice cloning [24, 70], enabling efficient shortening of raw, unscripted speech. However, this approach may be less effective for scripted speech, where disfluencies and tangents are less common, and high-level extraction alone may suffice. That said, even scripted speech often includes paralinguistic speech errors, such as mispronunciations or unintended emphasis. To fix such errors, we suggest future work to explore tools that go beyond transcript-based editing to offer editors fine-grained control over paralinguistic features to give editors control over not just what is said, but how it is said. Another direction is to explore how speech style varies based on audience context [25, 33]. Editors often adapt content to fit different audiences [8, 19, 42, 62], and a similar adaptation likely exists in speech. Future systems could support audience-aware speech editing to suggest edits depending on the listener.

Generalization to Other Media. TalkLess’s approaches hold promise for applications beyond speech, such as video and text editing. Many creators repurpose long-form videos (e.g., on YouTube) to produce short-form videos (e.g., on TikTok) with a similar goal of condensing content as presented in prior work [6, 96]. Prior work in video summarization [37, 86] compresses a long series of frames into short ones, either by removing single frames or whole video segments [43, 93]. As these cuts are based only on the visuals, this leads to unnatural cuts in the audio track. Future work can explore how audio and video compression approaches can be combined to address these issues. TalkLess extends transcript-based speech editing [22, 51, 76], which brings speech editing closer to writing. This suggests opportunities for future work to integrate elements from prior work on writing assistants into work on transcript-based audio and video editing systems and vice versa. For instance, TalkLess’s editing types could support efficient review of AI writing suggestions for shortening drafty text, particularly when combined with executable, verifiable edits [50]. In addition, TalkLess’s outline pane with importance scores could assist writers in quickly identifying segments or sentences to cut in long text documents, similarly to how editors in our studies used the outline pane for segment- and sentence-level cuts. We expect future systems to narrow the gap between transcript-based speech editing and writing systems.

Editing for Perfection versus Authenticity. TalkLess makes a tradeoff between polishing the speech transcript through abstraction and preserving the original speaker style through extraction. There exists a general trade-off that editors need to make between

using editing tools to reach perfection (e.g., auto speech enhancement [1, 80, 88, 89], automatic filler word removal [23, 71], condensing speech [6, 95]) and avoiding editing for authenticity to preserve unique context and identity. In professional contexts such as small business ads or lectures, users may prioritize polishing for perceived professionalism and clarity and thus desire the removal of disfluencies and filler words. Conversely, in personal or intimate contexts, users may want to preserve their authenticity by retaining personal filler words that reflect speaker personality [52]. Thus, to preserve more of the speaker’s personality, future iterations of TalkLess can avoid removing unique speech patterns by extending the scoring function to reward generations that preserve a higher count of unique speech patterns based on user-defined regular expressions where creators can define what words to preserve. Similarly, future work can explore using LLMs to implicitly learn speaker-specific speech patterns based on the transcript to prevent the removal of those for authenticity.

Ethical Considerations. TalkLess allows editors to efficiently shorten audio recordings. However, it introduces ethical risks, particularly in high-stakes scenarios such as journalism, legal proceedings, and academic communication where accuracy and authenticity are important. The ability to rewrite, condense, or rephrase speech raises concerns about editors being able to intentionally or unintentionally misrepresent a speaker’s original intent or remove important contextual information. Prior work on deepfakes has shown how synthetic audio can undermine trust in media [83]. For example, when editing quoted speech material, even subtle changes can risk misrepresenting the credibility of the source. To mitigate these risks, future systems could allow editors to “lock” specific transcript passages such as quotations or factual claims to prevent the system from modifying these passages to ensure preserving the verbatim speech. Additionally, automated claim detection [48] can be used to flag high-risk edits where factual statements are modified, while generating fact-check explanations [46] can support editors to verify correctness before publication. To increase transparency when content is shared or published, content credentials [87] describing the edit history would help collaborators or future editors track the speech’s modification history and audio watermarking [16] would help audiences identify modified segments.

8 CONCLUSION

We introduced TalkLess, a system that flexibly integrates extractive and abstractive summarization to shorten speech while preserving the original content and style. Our approach combines a novel speech shortening pipeline, an automated audio editing mechanism, and a state-of-the-art open-source speech editing model. Our evaluations revealed that TalkLess significantly reduces speech disfluencies and incoherence errors while retaining more content compared to a previous automatic shortening method. All participants in our user study (N=12) preferred editing speech with TalkLess over a baseline interface, reporting significantly lower cognitive load and significantly higher creative flexibility. In the exploratory study (N=3), we also explored how creators edit their own speech using TalkLess. We hope our work inspires future work to balance abstraction and extraction in content-based editing tools.

REFERENCES

- [1] Sherif Abdulatif, Ruizhe Cao, and Bin Yang. 2024. Cmgan: Conformer-based metric-gan for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 2477–2493.
- [2] Adobe. 2025. Adobe Audition. <https://www.adobe.com/ca/products/audition.html> Accessed: 2025.
- [3] Alitu. 2025. Alitu. <https://alitu.com/> Accessed: 2025.
- [4] Audacity. 2025. Audacity. <https://www.audacityteam.org/> Accessed: 2025.
- [5] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–38.
- [6] Aadit Barua, Karim Benharrah, Meng Chen, Mina Huh, and Amy Pavel. 2025. Lotus: Creating Short Videos From Long Videos With Abstractive and Extractive Summarization. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 967–981. <https://doi.org/10.1145/3708359.3712090>
- [7] Chris Baume, Mark D Plumbley, Janko Čalić, and David Frohlich. 2018. A contextual study of semantic speech editing in radio production. *International Journal of Human-Computer Studies* 115 (2018), 67–80.
- [8] Karim Benharrah, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [9] Floraine Berthouzoz, Wilnot Li, and Maneesh Agrawala. 2012. Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–8.
- [10] J Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).
- [11] Scot Bruce. 2025. Modern American History Zoom Lecture 1. <https://youtu.be/QtN2UihMNi4> Accessed: 2025.
- [12] Scot Bruce. 2025. Modern American History Zoom Lecture 2. https://youtu.be/5hzV4bBXX_M Accessed: 2025.
- [13] Juan Casares, A Chris Long, Brad A Myers, Rishi Bhatnagar, Scott M Stevens, Laura Dabbish, Dan Yocum, and Albert Corbett. 2002. Simplifying video editing using metadata. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*. 157–166.
- [14] Hakan Ceylan, Rada Mihalcea, Umut Özertem, Elena Lloret, and Manuel Palomar. 2010. Quantifying the limits and success of extractive summarization systems across domains. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*. 903–911.
- [15] JK Chambers and Peter Trudgill. [n. d.]. *The Handbook of Language Variation and Change*. ([n. d.]).
- [16] Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 2023. Waymark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770* (2023).
- [17] Xiang “Anthony” Chen, Chien-Sheng Wu, Lidiya Murakhovs’ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2023. Marvista: exploring the design of a human-AI collaborative news reading tool. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–27.
- [18] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.* 21, 4, Article 21 (jun 2014), 25 pages. <https://doi.org/10.1145/2617588>
- [19] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2024. Proxona: Leveraging LLM-Driven Personas to Enhance Creators’ Understanding of Their Audience. *arXiv preprint arXiv:2408.10937* (2024).
- [20] Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-Level Planning for Text Simplification. In *Conference of the European Chapter of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:258378147>
- [21] Hai Dang, Karim Benharrah, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA) (UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. <https://doi.org/10.1145/3526113.3545672>
- [22] Descript. 2025. Descript. <https://www.descript.com/> Accessed: 2025.
- [23] Descript. 2025. Descript Underlord. <https://www.descript.com/underlord> Accessed: 2025.
- [24] ElevenLabs. 2025. ElevenLabs. <https://elevenlabs.io/> Accessed: 2025.
- [25] Frederick Erickson. 1978. Howard Giles and Peter F. Powesland. *Speech style and social evaluation*. London and New York: Academic Press, 1975. Pp. 218. *Language in Society* 7, 3 (1978), 428–433.
- [26] Lisa A Fast and David C Funder. 2008. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology* 94, 2 (2008), 334.
- [27] Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 360–368.
- [28] C Ailie Fraser, Joy O Kim, Hujung Valentina Shin, Joel Brandt, and Mira Dontcheva. 2020. Temporal segmentation of creative live streams. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [29] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.
- [30] Sadaaki Furui, Tomonori Kikuchi, Yosuke Shinnaka, and Chiori Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing* 12, 4 (2004), 401–408.
- [31] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792* (2018).
- [32] Azin Ghazimatin, Ekaterina Garmash, Gustavo Penha, Kristen Sheets, Martin Achenbach, Oguz Semerci, Remi Galvez, Marcus Tannenbergh, Sahitya Mantravadi, Divya Narayanan, et al. 2024. PODTILE: Facilitating Podcast Episode Browsing with Auto-generated Chapters. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4487–4495.
- [33] Howard Giles, Anthony Mulac, James J Bradac, and Patricia Johnson. 2012. Speech accommodation theory: The first decade and beyond. In *Communication yearbook 10*. Routledge, 13–48.
- [34] Nianlong Gu, Elliott Ash, and Richard HR Hahnloser. 2021. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. *arXiv preprint arXiv:2107.08929* (2021).
- [35] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K Kummerfeld, and Elena L Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [36] Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications* 121 (2019), 49–65.
- [37] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*. Springer, 505–520.
- [38] 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*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [39] Bernd Huber, Hujung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J Mysore. 2019. B-script: Transcript-based b-roll video editing with recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [40] huggingface. 2025. all-mpnet-base-v2. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2> Accessed: 2025.
- [41] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang “Anthony” Chen, Young-Ho Kim, and Amy Pavel. 2023. Avscript: Accessible video editing with audio-visual scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [42] Yu-Kai Hung, Yun-Chien Huang, Ting-Yu Su, Yen-Ting Lin, Lung-Pan Cheng, Bryan Wang, and Shao-Hua Sun. 2025. SimTube: Simulating Audience Feedback on Videos using Generative AI and User Personas. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1256–1271.
- [43] Haojian Jin, Yale Song, and Koji Yatani. 2017. Elasticplay: Interactive video summarization with dynamic time budgets. In *Proceedings of the 25th ACM international conference on Multimedia*. 1164–1172.
- [44] Zeyu Jin, Gautham J. Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein. 2017. VoCo: text-based insertion and replacement in audio narration. *ACM Trans. Graph.* 36, 4, Article 96 (July 2017), 13 pages. <https://doi.org/10.1145/3072959.3073702>
- [45] Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/> Online manuscript released August 20, 2024.
- [46] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and abstractive explanations for fact-checking and evaluation of news. *arXiv preprint arXiv:2104.12918* (2021).
- [47] Ambika Kirkland, Joakim Gustafson, and Eva Szekely. 2023. Pardon my disfluency: The impact of disfluency effects on the perception of speaker competence and confidence. In *Proceedings of INTERSPEECH*. 5217–5221.
- [48] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats* 2, 2, Article 14 (April 2021), 16 pages. <https://doi.org/10.1145/3412869>

- [49] NE Kriman. 2024. Measuring text summarization factuality using atomic facts entailment metrics in the context of retrieval augmented generation. *arXiv preprint arXiv:2408.15171* (2024).
- [50] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–23.
- [51] Adobe Labs. 2023. *Project Blink*. <https://labs.adobe.com/projects/blink/> AI-powered video editing tool.
- [52] Charlyn M Laserna, Yi-Tai Seih, and James W Pennebaker. 2014. Um... who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology* 33, 3 (2014), 328–338.
- [53] Mackenzie Leake and Wilmot Li. 2024. ChunkyEdit: Text-first video interview editing via chunking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [54] Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical summarization for longform spoken dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 582–597.
- [55] Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical Summarization for Longform Spoken Dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (UIST '21). Association for Computing Machinery, New York, NY, USA, 582–597. <https://doi.org/10.1145/3472749.3474771>
- [56] Chin-Yew Lin and Eduard Hovy. 2003. The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*. 73–80.
- [57] Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9815–9822.
- [58] Susan Lin, Jeremy Warner, JD Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Björn Hartmann, et al. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [59] Yang Liu. 2019. Fine-tune BERT for Extractive Summarization. *ArXiv abs/1903.10318* (2019). <https://api.semanticscholar.org/CorpusID:85500417>
- [60] lowerquality. 2025. lowerquality/gentle. <https://github.com/lowerquality/gentle> Accessed: 2025.
- [61] Shih-Hao Lu, Huyen Thi Thanh Tran, and Thanh-Sang Ngo. 2025. Are we ready for artificial intelligence voice advertising? Comparing human and artificial intelligence voices in audio advertising in a multitasking context. *Quality & Quantity* 59, Suppl 1 (2025), 1–22.
- [62] Hannah Mieczkowski, Jeffrey T Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–14.
- [63] Max Morrison, Pranav Pawar, Nathan Pruyne, Jennifer Cole, and Bryan Pardo. 2024. Crowdsourced and automatic speech prominence estimation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12281–12285.
- [64] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023* (2016).
- [65] Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval* 5, 2–3 (2011), 103–233.
- [66] OpenAI. 2025. OpenAI API. <https://platform.openai.com> Accessed: 2025.
- [67] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* (2017).
- [68] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos.. In *UIST*, Vol. 10. Citeseer, 2642918–2647400.
- [69] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759.
- [70] Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. *arXiv preprint arXiv:2403.16973* (2024).
- [71] Podcastle. 2025. Podcastle. <https://podcastle.ai/> Accessed: 2025.
- [72] Rev.AI. 2025. rev.ai. <https://www.rev.ai/> Accessed: 2025.
- [73] Emma Rodero, Robert F Potter, and Pilar Prieto. 2017. Pitch range variations improve cognitive processing of audio messages. *Human communication research* 43, 3 (2017), 397–413.
- [74] Victor Rosi, Emma Soopramanien, and Carolyn McGettigan. 2025. Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans* 4 (2025), 100143. <https://doi.org/10.1016/j.chbah.2025.100143>
- [75] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, and Maneesh Agrawala. 2015. Capture-time feedback for recording scripted narration. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 191–199.
- [76] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 113–122.
- [77] F. H. Sanford. 1942. Speech and personality. *Psychological Bulletin* 39, 10 (1942), 811–845. <https://doi.org/10.1037/h0060838> Place: US Publisher: American Psychological Association.
- [78] Edward Sapir. 1927. Speech as a Personality Trait. *Amer. J. Sociology* 32 (1927), 892 – 905. <https://api.semanticscholar.org/CorpusID:144173316>
- [79] Björn Schuller and Anton Batliner. 2013. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.
- [80] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. 2022. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065* (2022).
- [81] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre Tixier, Polykarp Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *arXiv preprint arXiv:1805.05271* (2018).
- [82] Hujung Valentina Shin, Wilmot Li, and Frédo Durand. 2016. Dynamic authoring of audio with linked scripts. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 509–516.
- [83] Mohamed R Shoaib, Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. 2023. Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In *2023 international conference on computer and applications (ICCA)*. IEEE, 1–7.
- [84] slowkow. 2025. slowkow/needleman-wunsch.py. <https://gist.github.com/slowkow/06c6dba9180d013dfd82bec217d22eb5> Accessed: 2025.
- [85] Rajka Smiljanić and Ann R Bradlow. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass* 3, 1 (2009), 236–264.
- [86] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.
- [87] Eliza Strickland. 2024. This Election Year, Look for Content Credentials: Media organizations combat deepfakes and disinformation with digital manifests. *IEEE Spectrum* 61, 01 (2024), 24–27.
- [88] Jiaqi Su, Adam Finkelstein, and Zeyu Jin. 2019. Perceptually-motivated environment-specific speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7015–7019.
- [89] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. 2021. HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 166–170.
- [90] Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. *arXiv preprint arXiv:2404.10774* (2024).
- [91] Anh Truong and Maneesh Agrawala. 2019. A Tool for Navigating and Editing 360 Video of Social Conversations into Shareable Highlights.. In *Graphics Interface*. 14–1.
- [92] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [93] Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)* 3, 1 (2007), 3–es.
- [94] u/HighReps. 2022. *How long does it take you to edit an episode?* https://www.reddit.com/r/podcasting/comments/umlxia/how_long_does_it_take_you_to_edit_an_episode/ Reddit post in r/podcasting.
- [95] Bryan Wang, Zeyu Jin, and Gautham Mysore. 2022. Record Once, Post Everywhere: Automatic Shortening of Audio Stories for Social Media. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–11.
- [96] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2024. PodReels: Human-AI Co-Creation of Video Podcast Teasers. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 958–974.
- [97] Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. *arXiv preprint arXiv:2204.04907* (2022).
- [98] Theodora Worledge, Tatsunori Hashimoto, and Carlos Guestrin. 2024. The Extractive-Abstractive Spectrum: Uncovering Verifiability Trade-offs in LLM Generations. *arXiv preprint arXiv:2411.17375* (2024).
- [99] Tal Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44,

- 3 (2010), 363–373. <https://doi.org/10.1016/j.jrp.2010.04.001>
- [100] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive Summarization via ChatGPT for Faithful Summary Generation. In *Conference on Empirical Methods in Natural Language Processing*. <https://api.semanticscholar.org/CorpusID:258048787>
- [101] Shiyue Zhang, David Wan, and Mohit Bansal. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *arXiv preprint arXiv:2209.03549* (2022).
- [102] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938* (2021).
- [103] Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9709–9716.

A APPENDIX

A.1 Prompt Templates

A.1.1 Prompt Template for Generating Shortened Candidates. We distilled principles that describe requirements for ideal shortened speech transcripts based on our design goals and include them in our prompt:

- All filler words should be removed (**G1**).
- All repetitions should be removed to avoid redundancies (**G1**).
- There should be no loss of information (**G2**).
- The original style should be preserved (**G3**).
- No new words should be added besides when merging sentences (**G3**).
- The original wording should not be changed (**G3**).
- Unique words used for emphasis should be preserved (**G3**).
- Do not change any word spelling (**G3, G5**).

We share the prompt template in Python that we used below (some formatting added for the paper). The few-shot examples also followed this template and were inserted between the system description and the final instruction, as indicated below. This format is expected by the OpenAI API (Chat Completion). *Python variables are color-coded in blue.*

```
{
  "role": "system",
  "content": [
    {
      "type": "text",
      "text": f'You are a helpful writing assistant. Your
task is to edit my transcript according to the following
guidelines:
1. **No New Words**: Do not add any new words that are
not already in the transcript, besides when merging sentences.
2. **Same Wording**: Do not change the original
wording.
3. **Include Everything**: Ensure that no piece of
information from the original transcript is left out.
4. **Remove Filler Words**: Eliminate all filler words
like "um" and "uh."
5. **Preserve Style**: Keep the original language
style intact; don't change the tone or formal/informal nature of
the language.
6. **Remove Repetitions**: Delete any repeated
information to avoid redundancies.
7. **Unique Words**: Keep unique/rare words as they
may aid in memory when listening to the transcript
8. **No Hyphens or Word Corrections**: Do not combine
words using hyphens or introduce hyphens and do not change a
single character of a word to correct it if the word itself would
be kept.
9. **Target Length**: Ensure the final output is not
more than {target_length} words.
```

```

Your response will not be formatted and will only
contain the shortened transcript.',
    ]
  ],
},
...few-shot examples...
{
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": f"{segment}"
    }
  ]
}
}
```

A.1.2 Prompt Template for Classifying Edit Types. This is the prompt template we used with the same formatting as indicated in A.1.1.

```
{
  "role": "system",
  "content": [
    {
      "type": "text",
      "text": "I will give you a sentence before and after
an edit. Your task is to describe the edit that was made. Your
response should not be formatted and only contain the type of the
edit. Pick among the following types: ['Filler Word Removal', '
Repetition Removal', 'Clarification', 'Emphasis Removal', '
Information Removal'].
```

Here are definitions of the edit types:

****Filler Word Removal****: Removing unnecessary words or phrases that do not add meaning to the sentence, such as "um," "like," "you know," or "basically." These words are often used as verbal pauses and can be omitted without changing the overall content.

Examples: ...

****Repetition Removal****: Removing repeated words, phrases, or ideas that add no new value to the content. This type of edit removes repeated words for clarity and conciseness.

Examples: ...

****Clarification****: This type of edits make an unclear or ambiguous phrase more understandable by providing additional context or rephrasing the phrase.

Examples: ...

****Emphasis Removal****: Removes words or phrases used to stress or exaggerate a point. These may include words like "really" and "very".

Examples: ...

****Information Removal****: Removing details, facts, or sections of the content that are either irrelevant to the main information or not relevant in a different context.

Examples: ..."

```
    }
  ],
},
{
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": f"
Before: {before}
After: {after}
",
    }
  ],
}
}
```


A.1.3 Prompt Template for Importance Rating. This is the prompt template we used with the same formatting as indicated in A.1.1.

```
{
  "role": "system",
  "content": [
    {
      "text": "Rate from 1-10, how crucial are the following
information points to the main text content based on the
following purpose:

      Purpose: {purpose}

      For example, when an information is important
contextually but not central to the main content of the text, it
should be rated low.

      Format your output as JSON with a list of ratings from
1-10 for each information bit.

      Example output: {"importances": [X, Y, ...]}" where X,
Y, ... are numbers between 1 and 10 for each respective
information bit in the order I gave them to you.',
      "type": "text",
    }
  ],
},
... few-shot examples
{
  "role": "user",
  "content": [
    {
      "text": f"
      Information Points: {information_points}
      Text: {text}
      "
      "type": "text",
    }
  ],
},
}
```

A.1.4 Prompt Template for Content Extraction. This is the prompt template we used with the same formatting as indicated in A.1.1.

```
{
  "role": "system",
  "content": [
    {
      "text": "Extract a list of all atomic information bits
that occur in a text I will give to you. An information bit is an
atomic essential piece of information that occurs in the text.
Align each information bit with the respective phrase from the
original text where this information bit is coming from.
Additionally, give each information bit an importance score from 1
to 10 based on the information bits importance within the overall
text. Your output response should be formatted as JSON in the
following way:

      {
        "information_bits": [
          {
            "title": X,
            "alignment": Y,
            "importance": Z
          },
          {
            "title": X,
            "alignment": Y,
            "importance": Z
          },
          ...,
          {
            "title": X,
            "alignment": Y,
            "importance": Z
          }
        ]
      }
    }
  ],
}
```

```
}
  where X is the information bit, Y is the phrase from
the original text that aligns the most with the text, and Z is the
information bit's importance (1-10) within the text.',
  "type": "text",
}
],
},
{
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": f"{text}"
    }
  ]
}
```

A.2 List of prompts used in the exploratory user study

Creators chose one topic from a list of ten to record a 2-3 minute speech recording:

- Describe your morning routine on a typical weekday.
- Talk about your favorite meal to cook or eat. Why do you love it?
- Tell a story about a time you overcame a challenge.
- What's one hobby or activity you enjoy, and how did you get into it?
- Describe your ideal vacation - where would you go and what would you do?
- If you could have any superpower, what would it be and why?
- Talk about a piece of technology you use every day and how it affects your life.
- What's something you wish you had learned earlier in life?
- Describe a place (real or fictional) that feels peaceful or inspiring to you.
- Share your thoughts on how the internet or social media has changed communication.

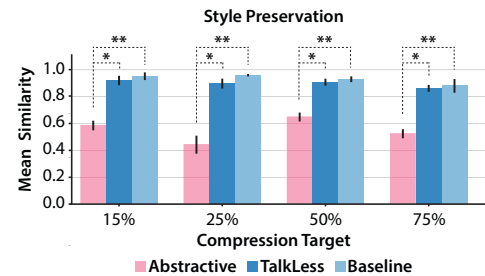


Figure 11: Cosine style similarities between abstractive, TalkLess, and ROPE transcripts. Transcript styles of TalkLess and ROPE were significantly closer to the original transcript style than the abstractive one for all compression rates but 25%.

* $p < .05$, ** $p < .01$

Target	Method	Compression Target Deviation ↓			% Words Synthesized ↓			Coverage ↑		
		μ	σ	range	μ	σ	range	μ	σ	range
15%	LLM-only	33.75%	16.12%	0.40%-73.05%	9.17%	6.55%	0.00%-34.39%	66.54%	9.15%	43.48%-92.65%
	Optimization	7.53%	2.85%	0.09%-32.38%	1.15%	0.54%	0.00%-8.12%	87.30%	2.84%	66.38%-97.40%
25%	LLM-only	32.70%	16.25%	0.23%-59.96%	8.11%	6.42%	0.00%-32.38%	66.36%	7.66%	37.00%-90.69%
	Optimization	8.42%	3.74%	0.00%-21.56%	1.19%	0.80%	0.00%-9.80%	84.41%	3.59%	65.32%-96.43%
50%	LLM-only	17.78%	9.18%	0.00%-35.40%	9.51%	7.26%	0.00%-37.27%	61.30%	6.30%	39.97%-89.44%
	Optimization	3.69%	2.41%	0.00%-18.14%	2.58%	2.05%	0.00%-13.17%	75.20%	3.33%	62.45%-88.40%
75%	LLM-only	2.76%	4.00%	0.00%-60.14%	14.22%	7.33%	0.00%-42.67%	52.04%	6.22%	33.05%-92.36%
	Optimization	1.49%	1.18%	0.00%-6.53%	6.98%	4.32%	0.00%-21.97%	60.22%	2.86%	43.97%-75.57%
Overall	LLM-only	21.75%	5.14%	0.00%-73.05%	10.25%	0.41%	0.00%-42.67%	61.56%	1.20%	33.05%-92.65%
	Optimization	5.28%	0.92%	0.00%-32.38%	2.98%	1.49%	0.00%-21.97%	76.78%	0.32%	43.97%-97.40%

Table 3: TalkLess’s optimization approach generates shortened candidates with lower deviation from the target compression, lower % of words synthesized and higher coverage compared to an ablation method where we prompted an LLM alone using our base prompt A.1.1.

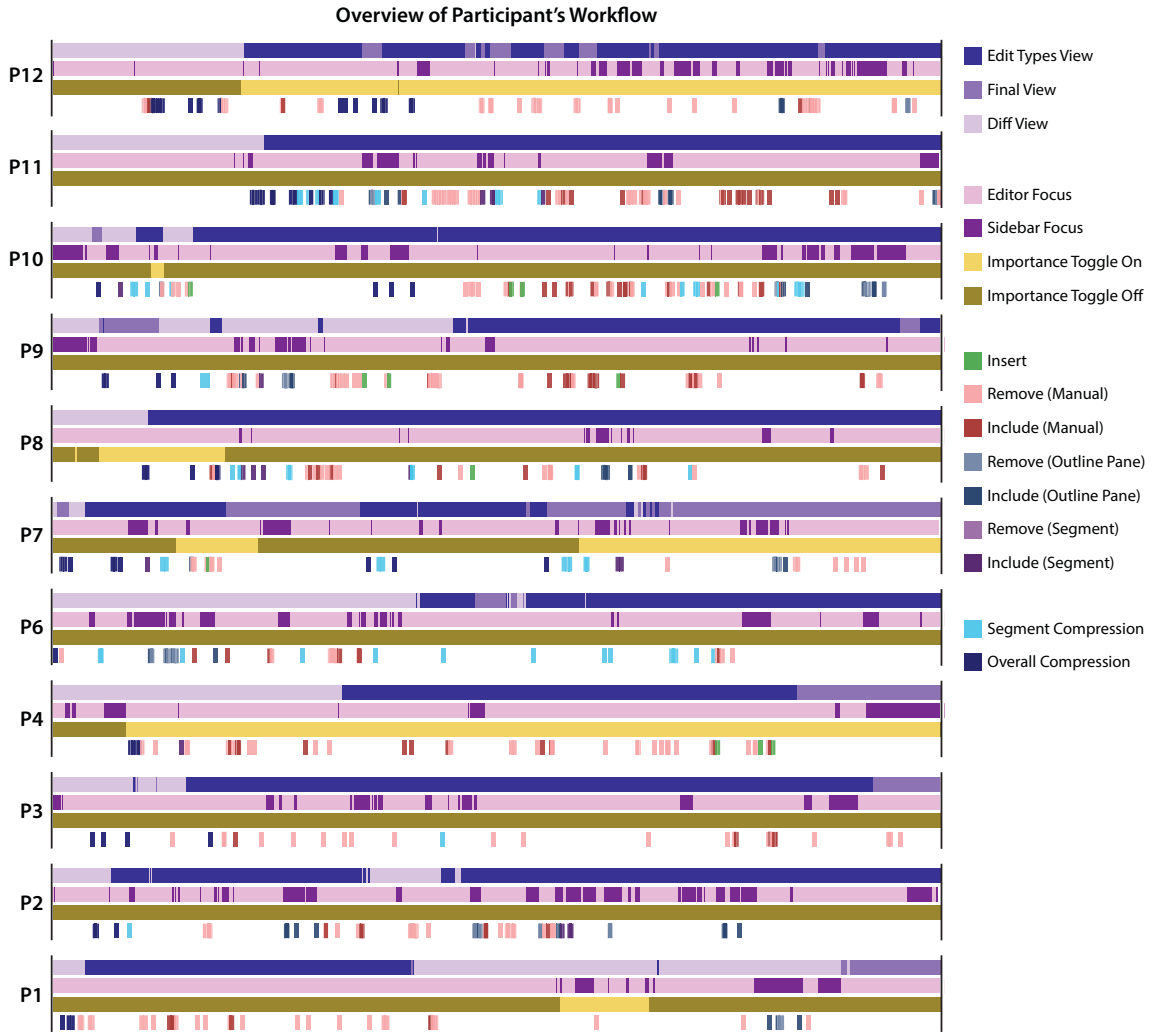


Figure 12: Interaction Behavior Overview.

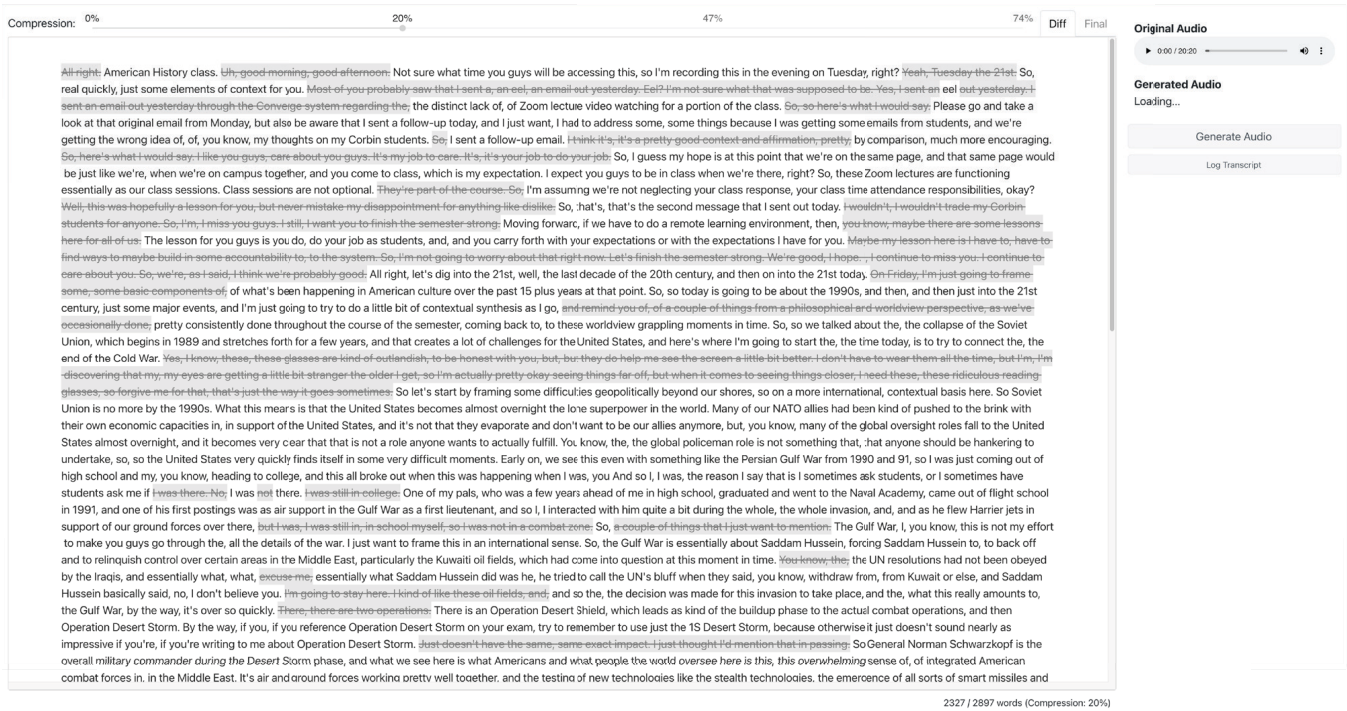


Figure 13: Baseline interface used in our user evaluation.

Measure	Baseline Interface	TalkLess	Z	p
NASA TLX				
Mental Demand	$\mu = 4.17, \sigma = 0.94$	$\mu = 3.08, \sigma = 1.16$	2.0	$p < .05$
Temporal Demand	$\mu = 3.58, \sigma = 0.90$	$\mu = 2.67, \sigma = 0.89$	2.0	$p < .05$
Effort	$\mu = 3.83, \sigma = 1.03$	$\mu = 3.08, \sigma = 0.79$	2.1	$p < .05$
Frustration	$\mu = 2.50, \sigma = 1.51$	$\mu = 1.58, \sigma = 0.67$	2.1	$p < .05$
Creativity Support Index				
Exploration	$\mu = 1.75, \sigma = 0.75$	$\mu = 4.42, \sigma = 0.51$	3.1	$p < .01$
Engagement	$\mu = 2.75, \sigma = 1.22$	$\mu = 4.33, \sigma = 0.89$	2.8	$p < .01$
Worth the Effort	$\mu = 2.67, \sigma = 1.07$	$\mu = 4.50, \sigma = 0.67$	2.7	$p < .01$
Expressiveness	$\mu = 1.42, \sigma = 0.90$	$\mu = 3.75, \sigma = 1.14$	3.0	$p < .01$
System Usability Scale				
Frequent Use	$\mu = 2.08, \sigma = 0.79$	$\mu = 4.5, \sigma = 0.67$	3.0	$p < .01$
Easy to use	$\mu = 3.33, \sigma = 0.98$	$\mu = 4.42, \sigma = 0.67$	3.1	$p < .05$
Quick to learn	$\mu = 4.00, \sigma = 0.74$	$\mu = 4.58, \sigma = 0.51$	2.2	$p < .05$

Table 4: Analysis of task load, creativity support, and usability ratings using paired Wilcoxon Sign Rank tests. We applied Greenhouse-Geisser correction when the equal-variances assumption is violated (Mauchly’s test $p < .05$).

Compression	TalkLess	ROPE	Abstractive
15%	$\mu = 0.92, \sigma = 0.06$	$\mu = 0.95, \sigma = 0.06$	$\mu = 0.58, \sigma = 0.07$
25%	$\mu = 0.9, \sigma = 0.07$	$\mu = 0.96, \sigma = 0.02$	$\mu = 0.44, \sigma = 0.13$
50%	$\mu = 0.91, \sigma = 0.05$	$\mu = 0.93, \sigma = 0.04$	$\mu = 0.65, \sigma = 0.07$
75%	$\mu = 0.86, \sigma = 0.05$	$\mu = 0.88, \sigma = 0.01$	$\mu = 0.52, \sigma = 0.07$

Table 5: Mean and standard deviation of cosine style similarities between abstractive, TalkLess, and ROPE transcripts, each compared to the original transcript.