Speech vs. Transcript: Does It Matter for Human Annotators in Speech Summarization?

Anonymous ACL submission

Abstract

Reference summaries for abstractive speech summarization require human annotation, which can be performed by listening to an audio recording or by reading textual transcripts of the recording. In this paper, we examine whether summaries based on annotators listening to the recordings differ from those based on annotators reading transcripts. Using existing intrinsic evaluation based on human evaluation, automatic metrics, LLM-based evaluation, and a retrieval-based reference-free method. We find that summaries are indeed different based on the source modality, and that speech-based summaries are more factually consistent and information-selective than transcript-based summaries. Meanwhile, transcript-based summaries are impacted by recognition errors in the source, and expert-written summaries are more informative and reliable. We make all the collected data and analysis code public to facilitate the reproduction of our work and advance research in this area.

1 Introduction

Speech summarization (Hori et al., 2002; Rezazadegan et al., 2020) is the task of analyzing long speech recordings and generating a short textual summary that concisely conveys this information. Abstractive speech summarization (Sharma et al., 2022; Palaskar et al., 2021) aims to produce summaries that a human would write (Mani, 1999). Prior work in abstractive speech summarization (Palaskar et al., 2019; Kano et al., 2023; Sharma and Raj, 2022; Sharma et al., 2023b) has focused on developing automatic methods.

However, to the best of our knowledge, there has been little work examining how humans perform speech summarization. Such research is particularly important because the goal of abstractive summarization is to produce human-like summaries, and having high-quality human summaries is crucial for learning how to automatically summarize, i.e., for model training and evaluation. Also, information about how humans summarize may facilitate improved automatic approaches, and hence it is important to explore the annotation process. In this paper, we delve deeper into the process of human annotation for abstractive speech summarization.

Fig. 1 presents an overview of our paper. Humans can summarize speech when it is presented by (a) listening to the speech directly and then constructing the summary based on this, or (b) reading a textual transcript of the content spoken in the recording, and using that to construct the summary. It is unclear whether summaries based on listening to audio differ from summaries based on reading transcripts. For transcript-based summarization, the method used to obtain the transcription, either Automatic Speech Recognition (ASR) or manual transcription, may also have a significant effect. Additionally, the quality of a summary, regardless of the source modality, is likely to be determined by the level of expertise of the summarizer. To assess the impact of source modality, transcription, and expertise, we formulate the following research questions:

(RQ1) Are summaries based on annotators listening to speech recordings different from summaries based on annotators reading textual transcripts of speech? If so, how?

(RQ2) If ASR transcripts are used instead of manual transcripts, how do recognition errors impact human summaries?

(RQ3) Do summaries written by non-experts summaries differ from those written by experts, and if so, do the differences vary based on the source modality?

Past work in automatic speech summarization suggests that there are acoustic characteristics in speech beyond the textual context that are useful for speech summarization (Kano et al., 2023).

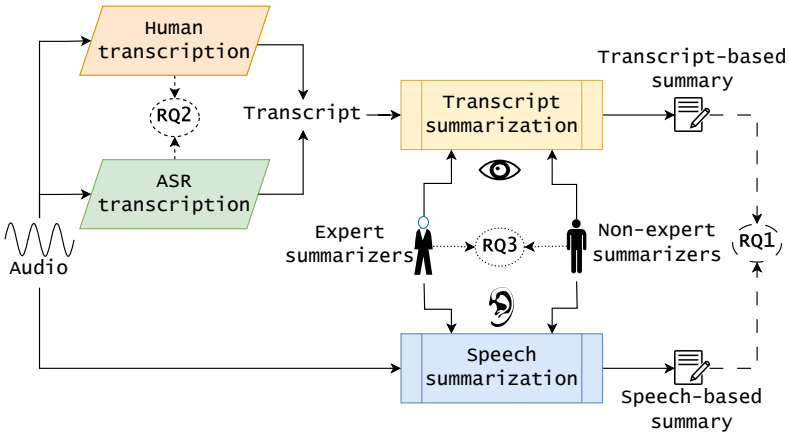

Figure 1: Overview of collection process and research questions

This implies that both transcript and raw speech carry complementary information that is useful for summarization. The fact that end-to-end models (Sharma et al., 2022; Kano et al., 2023; Matsuura et al., 2023b,a; Sharma et al., 2023b) that use speech features as input perform differently from cascade models (Kano et al., 2021; Palaskar et al., 2021; Yu et al., 2021; Palaskar et al., 2019) that use speech transcripts (Sharma et al., 2023a) suggests that there are differences in automatic summaries resulting from textual or speech input. In this paper, we study whether the same holds for human annotators.

However, any observations on the impact of reading versus listening for annotation are difficult to make from existing corpora like How2 (Sanabria et al., 2018), SLUE-TED (Shon et al., 2023), Interview (Sharma et al., 2023a), and AMI (McCowan et al., 2005), since it is unclear how human annotations were obtained from speech. Therefore, we select 1002 recordings from the *Interview* corpus test set and create a new evaluation set. We collect two expert human annotations for transcript-based summarization and two for speech-based summarization per recording and assess how the resulting summaries differ.

Comparing two given summaries is challenging. Broadly, we conduct three types of evaluation: (a) comparing the summary to its source transcript (*source-based evaluation*) to establish how extractive and factual the summary is, (b) evaluating the summary independently for its structure and grammar (*structure evaluation*), and (c) by comparing different summaries (*summary comparisons*) to assess relevance and which represents the input better using semantic similarity, LLM, and human evaluation. Experiments show that speech-based summaries are more information-selective and factually consistent while transcript-based summaries are preferred by human and LLM evaluators.

In real-world settings, human-annotated manual transcripts are not available, and, therefore, Automatic Speech Recognition (ASR) models may be used to generate such transcriptions. ASR transcripts may contain errors, and we investigate the impact of these errors on transcript-based summarization. To do this, we consider 100 recordings from the Interview test set, obtain speech transcriptions using Whisper (Radford et al., 2022), and perform human summarization using 2 expert annotators for each modality. Experiments show that ASR errors in the transcript degrade the quality of human summaries, resulting in lower coherence, fluency, and factual consistency with the source.

Prior work in human evaluation of abstractive text summarization highlights the importance of expertise for summary evaluations (Fabbri et al., 2021b; Gillick and Liu, 2010). However, to the best of our knowledge, there is no prior work that examines the need for expertise in human speech summarization. Therefore, in this work, we examine whether crowd-sourced non-expert summaries are different from expert summaries. To address this question, we obtain 2 non-expert annotations using Amazon Mechanical Turk (AMT) for each of the 1002 recordings from the *Interview* test set. From these summaries, we analyze (a) differences between expert and non-expert annotations, and (b) whether any differences can be attributed to the approach taken for summary annotation. Analysis shows that non-expert annotations are less fluent and coherent compared to expert summaries but as

factually consistent as expert summaries.

In summary, through this work, we find that transcript-based annotation is valuable if errors are minimal and longer, more informative summaries are desired, while speech-based annotation is desired for higher information selectivity, factual consistency, and resilience to transcription errors.

2 Related Work

For this study, we describe related work in three areas: (a) metrics for comparing summaries, (b) other work investigating the impact of source modality on human performance, and (c) investigations of ASR error impact on Spoken Language Understanding.

Comparing Summaries - Metrics: Numerous metrics have been proposed to evaluate the quality of summarization (Fabbri et al., 2021a). Earlier works use metrics like ROUGE-H, ROUGE-L (Zhou et al., 2006), ROUGE-WE (Ng and Abrecht, 2015), BLEU (Papineni et al., 2002), sentence recall/precision is commonly used for simple sentence based extraction summaries (Hirohata et al., 2005). Some studies have found that ROUGE and BLEU-based metrics correlate well with human judgments on summarization. However such scores are not always the best metric for summarization, especially when the input is longer like in meeting summarization and summarization of scientific articles. There have been scores of other metrics which are either variants or combinations of the earlier mentioned metrics like BertScore (Zhang et al., 2019), Sentence Mover's Similarity (SMS) (Clark et al., 2019), SummaQA (Scialom et al., 2019), SUPERT (Gao et al., 2020), CHRF (Popović, 2015), METEOR (Banerjee and Lavie), CIDEr (Vedantam et al., 2015), or model generated ones like s3 (Peyrard et al., 2017), BLANC (Vasilyev et al., 2020), summACCY (Hori et al., 2003), which are trained to predict the quality of an input summary.

LLMs have been used as an alternative method for summary evaluation. While some work expresses doubt that LLMs have reached a human level of summary evaluation capabilities (Shen et al., 2023), other methods have yielded more optimistic results (Wu et al., 2023). LLM evaluation methods usually take the form of automatic replacements for human evaluations such as Likert scale scoring, pairwise comparison, and binary factuality comparison (Gao et al., 2023). LLMs can also be used to answer qualitative questions about the factual consistency of a summary and explain the reasoning behind the provided answers (Luo et al., 2023).

Source Modality - Transcript vs. Speech: We are also interested in studying whether input modalities have an impact on task performance. Cognitive psychology literature has studied this question under various settings. (Khan et al., 2022) studies cognitive performance and information retention, when note-taking is performed using either voice or text. The paper finds that when the notes are taken using voice versus when they are written down by hand, voice leads to a higher conceptual understanding of the topic. (Stollnberger et al., 2013) studies how input modality effects learning of the English language, specifically whether learning from lecture audio vs reading lecture notes impacts retention. Other papers examine the interaction between input modality (text vs speech) and what users prefer when performing a task. Studies have reached various conclusions regarding the most useful/important modality based on the task at hand, what we do find is that the difference exists and is significant.

Impact of ASR on Downstream Tasks: ASR errors are known to have an impact on downstream task performance, like emotion recognition (Schuller et al., 2009; Feng et al., 2020), spoken language understanding (Shon et al., 2023), and Natural language understanding and dialog management (Serdyuk et al., 2018; Shivakumar et al., 2019). Therefore, it is important to assess their impact on human summarization.

3 Data Collection

Datasets that are used for speech summarization today do not have information on which of the two approaches, i.e., reading a textual transcript of the recording or listening to the recording to summarize was employed to obtain annotation. Further, they only have one annotation per recording, which may inhibit analysis due to annotator-specific biases or tendencies. Therefore, to conduct our experiments, we curate a multi-reference dataset from an existing dataset.

We considered a subset from the test partition of the *Interview* (Sharma et al., 2023a; Majumder et al., 2020; Zhu et al., 2021) corpus for speech summarization. *Interview* is the largest open-domain corpus with around 5-minute-long speech recordings containing spontaneous speech along with real audio noises representing music, and

events inside and outside the studio corresponding to discussion topics[1]. More details about the Interview corpus can be found in Appendix C.

Ensuring high-quality summary annotations is crucial to making reliable observations. Therefore, we took the necessary steps to ensure a fair and reliable data collection pipeline. Initially, we consulted with an expert from the in-house data collection team and conducted small pilot studies to make design decisions.

Audio Length: We conducted in-house pilot studies comprising six questions by trying to summarize audio recordings that were 2 minutes and 5 minutes long. As expected, summarizing 5-minute recordings consumed more time and effort, and was harder due to the high compression ratio. We concluded that limiting the audio length to 2 minutes would provide sufficient content to summarize without increasing the difficulty level of summarization significantly. Using forced alignment, we obtained audio reference transcripts up to the first 2 minutes.

Summary Length: A hard limit was not set on the length of the summary, but it was suggested that summaries should contain at least 2 sentences to convey the essence of the conversation. It was recommended that summaries contain between 50 to 80 words for the 2-minute long recordings.

Data Selection: To obtain reasonable summaries, it is important that audio recordings and textual transcripts are understandable, and have a sufficient amount of information to summarize. In cases where background noise or music is present, or when a large portion of the recording does not contain speech, these criteria are violated, leading to sub-optimal recordings for summarization. Therefore, audio recordings were transcribed using `Whisper-medium` (Radford et al., 2022), and the resulting Word Error Rate (WER) was used to remove unsuitable audio recordings from consideration. Based on manual inspection, a threshold of 25% WER was set, which eliminated recordings with only music or large amounts of noise.

Expert and Non-expert: Our data comprises 1,002 recordings in total, with each recording having a speech recording and its corresponding reference transcript. For each recording, we collected two expert summaries and two non-expert summaries by reading the text or listening to the audio, which gave us a total of 4 summaries based on the text and 4 summaries based on the audio on the same recording. Additionally, we solicited summary annotations from speech recognition transcripts for 100 of these 1002 recordings, sampled at random so the Word Error Rate with `Whisper-medium` is non-zero, i.e., the transcript has ASR errors.

Annotator Overlap: It is worth noting that each annotator was assigned to work on either the transcript or audio recording for a particular recording, but not both. This approach ensured that no annotator saw audio or transcription for a recording they had already annotated.

Expert Summary Annotation: Expert summary annotations were obtained through a third-party vendor that passed in-house qualitative assessments of summary outputs from a pilot annotation. The third-party vendor had proven their quality of annotation by consistently delivering similar summary annotations for other projects of the team. Detailed annotation guidelines are in Appendix A.

Non-Expert Summary Annotation Non-expert summary annotations were obtained through the Amazon Mechanical Turk (MTurk) platform. Each HIT required the annotator to accept the terms of a consent form that explained the task and requested consent to share responses for reproducible research. Each HIT included four questions, where annotators were asked to write 2 summaries by listening to audio recordings, and two summaries by reading transcript passages. More details are provided in Appendix B.

4 Evaluating Differences across Summaries

Summary evaluation is an inherently challenging problem for general settings since there is often no clear definition of what constitutes a good summary. Within this paper, we deem a summary good if it is factually faithful to and representative of the input source while being abstractive, coherent, and fluent. Therefore, in this section, we outline three types of evaluations.

4.1 Source-based Evaluation

Summary Length and Compression Ratio: Summary length is measured in words and the compression ratio (Mani, 1999) is defined as the number of words in the reference transcript divided by the number of words in the summary. The greater the

[1]The authors acquired permission from NPR to use NPR data for this research

summary length, typically, the higher the amount of information contained in the summary, and the lower the compression ratio.

Novel Words (%): The percentage of words in the summary that are not present within the source transcript, which could be considered to measure the extent of paraphrasing within the summary.

Extractiveness: It is a measure of how extractive the summary is and can be approximated by computing lexical overlap using ROUGE-L (Lin, 2004) between the source transcript and the summary.

Informativeness-Entities: We propose to compare named entities predicted from the source text and summaries as a proxy for the amount of information contained within the summary. Named entities are extracted using the Entity Recognizer[2] from Spacy (Honnibal and Montani, 2017).

Informativeness-Semantic Similarity: Also, we compare semantic similarity between the source and summaries using metrics like BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021).

Retrieval-based informativeness: A summary can be considered representative of or informative about the source when it can discriminate the correct source from a selection of all sources. We describe a retrieval-based measure where we compute the semantic embedding similarity (using BERTScore) between the hypothesis summary and all possible source texts. A discriminative summary should produce the highest similarity between the summary and the *correct* source text. This is measured using retrieval accuracy (RAcc). We also report the Mean Reciprocal Rank (MRR) as an indicator of retrieval performance.

Factual Consistency: We use UniEval to estimate the factual consistency of summaries with respect to the source transcript.

4.2 Structure Evaluation

To evaluate coherence and fluency automatically, we leverage UniEval (Zhong et al., 2022).

4.3 Summary Comparisons

Pairwise Similarity: We compute relevance and consistency using UniEval (Zhong et al., 2022), and use BARTScore as a measure of semantic similarity. Pairwise scoring is illustrated in Fig. 2, where s_i is a speech-based summary, t_j is a transcript-based summary, and $S(a, b)$ calculates the score

[2]https://spacy.io/api/entityrecognizer

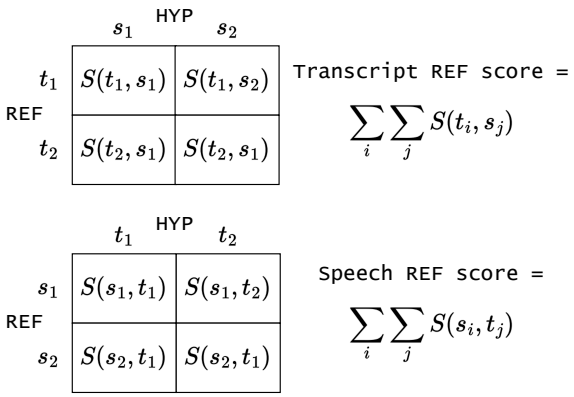

Figure 2: Pairwise Score Computation

between reference a and hypothesis b (e.g., BART score). The idea behind pairwise scoring is to compare two types of summaries, where one type is a reference and the other a hypothesis. This process results in two scores, a speech reference score, and a transcript reference score. By comparing these two scores, we can identify cases when the information in one type of summary is present within the other type of summary.

Pairwise Inter Annotator Agreement (IAA): For text generation tasks, the inter-annotator agreement is challenging to quantify. We use a combination of lexical and semantic similarity metrics like ROUGE-L (Lin, 2004), and BARTScore (Yuan et al., 2021) to automatically represent inter-annotator agreement.

Factual Consistency and **Source Representation**: We employ LLM-based evaluation and human evaluation to estimate the factual consistency of each summary, and how well each summary represents the source. These evaluation methods are described in detail below.

4.3.1 LLM-based Evaluation

Following the LLM summary evaluation methods listed in Section 2, we use ChatGPT (`gpt-3.5-turbo-0125`) as a qualitative evaluation method for the summaries collected for this work. Specifically, we designed a multiple choice-style AB test to compare summaries in terms of either factual consistency or accurate representation of the source transcript. The two questions posed to the LLM are displayed in Fig. 9 (see Appendix D.1), where the text in the brackets represents the summaries being evaluated and the associated source text. Each question was presented in a separate session to avoid bias from previous questions.

Each question is asked twice for each AB pair-

ing, but with the order in which the summaries are presented switched to avoid biasing the results. For example, if we were comparing a transcript-based summary and a speech-based summary, the question would be asked one time with the transcript summary as Summary 1, and then another time with the speech summary as Summary 1. Additionally, since all summaries are collected in pairs, we compare each summary to both summaries from the other group, exhausting all possible combinations. For example, if we were comparing a pair of transcript-based summaries to a pair of speech-based summaries, the first transcript summary would be compared to both speech summaries, and then the second would be compared to both speech summaries, resulting in four total comparisons. Considering both the order of presentation and the comparison combinations, eight questions are posed to the LLM per source sample.

4.3.2 Human Evaluation

To validate the expert summary comparison, we also conducted an AB test for 30 recordings with 15 annotators. We present 2 summaries based on different modalities and their source transcript to a user. We designed 2 multiple-choice questions that ask human annotators to compare summaries based on their factual consistency and how accurately they represent the source transcript. The setup of questions is similar to that in LLM evaluation.

5 Research Methodology

5.1 RQ1: Speech versus Text Inputs

We apply the methods and metrics described in Section 4 to explore differences between summaries derived from speech recordings and summaries extracted from text transcripts. We use the expert annotations obtained for 1002 recordings for this analysis since expert summaries are more reliable.

5.2 RQ2: Impact of Transcript Errors on Human Summarization

We measure the impact of ASR error propagation on the output summary by comparing a subset of 96 expert summaries given ASR-generated transcripts and ground-truth transcripts. The summaries resulting from the Whisper-generated transcripts are compared to the summaries based on ground-truth transcripts using the methods and metrics outlined in Section 4.

We analyze the effect of ASR errors on the quality of the summaries based on ASR transcripts by plotting informativeness measures like Entity-F1 and WER (See Figure 10 in Appendix D.2), and find that as WER increases, Entity-F1 decreases.

5.3 RQ3: Expert versus Non-Expert Annotation

We use all methods and metrics from Section 4 to explore the effect of expertise on summary quality. Similar to RQ1, we provide an overall comparison across speech-based summaries as opposed to transcript-based summaries arising from non-expert annotations and conclude on the reliability of expert and non-expert summaries.

6 Experimental Results and Discussion

6.1 RQ1: Speech versus Transcript Inputs

Evaluation across the expert summaries, including a comparison between transcript-based and speech-based summaries, is shown in the first two columns of Table 3.

It is immediately apparent that speech-based summaries are significantly shorter and more compressed than transcript-based summaries, indicating that speech-based summaries are highly selective. This is likely because the summarizers extract only the main points when listening to a recording, while a transcript serves as an easier reference for smaller details. This claim is supported by the low level of extractiveness in the speech-based summaries, indicated by a high percent of novel words, low ROUGE-L scores, and low number of named entities. Additionally, coherence for speech-based summaries is relatively low, likely because they are not long or detailed enough to form a very structured summary.

These results also indicate that speech-based summaries are less informative than transcript-based summaries. Again, this is likely because summarizers can extract details by referencing transcripts much more easily than by searching through the recording. Results that support this claim are high BERT- and BARTScores, high retrieval accuracy, factual consistency, and human and LLM ratings. The higher pairwise BARTScore for speech references indicates that more of the information within the speech summary is likely present in the transcript-based summary, making speech-based summaries more selective in relevant information. IAA scores computed based on ROUGE-L and BARTScore indicate that speech-based summaries promote higher consensus between raters.

Metric	Expert Transcript	Expert Speech	Non-expert Transcript	Non-expert Speech
Summary Length	78.58 ± 27.72	56.95 ± 18.30	75.35 ± 17.70	77.42 ± 33.85
Compression Ratio	3.76 ± 2.20	5.20 ± 3.06	3.69 ± 1.95	3.66± 1.86
Word Vocabulary Size	20,408	16,211	21511	21676
Novel Words %	24.63 ± 13.57	25.66 ± 13.22	37.47 ± 13.38	38.45± 14.21
Extractiveness↓	20.70 ± 10.53	**18.57 ± 10.88**	**18.37± 7.26**	18.38 ±11.29
Entity F1↑	**33.63 ± 29.83**	26.21 ± 26.90	**34.51 ± 17.28**	30.09± 17.72
BERTScore ↑	**85.66 ±2.45**	85.32 ±2.42	**85.28± 2.32**	85.06 ±2.76
BARTScore ↑	**-3.37 ±0.47**	-3.46 ± 0.45	**-3.41± 0.44**	-3.43± 0.53
Retrieval Accuracy ↑	**68.86 + ± 46.31**	59.04 ± 49.75	**71.82 ± 44.99**	65.78 ± 47.44
MRR ↑	70.23 ± 16.08	**75.74 ± 16.29**	**76.56 ± 71.01**	71.01 ± 41.57
Factual Consistency (UniEval) ↑	82.68 ± 16.38	**84.58± 15.72**	**82.08 ± 16.90**	78.57± 19.79
Fluency (UniEval) ↑	**94.02± 4.54**	94.01± 5.26	**92.90± 8.21**	92.65 ±9.35
Coherence (UniEval)↑	**91.82± 12.97**	89.34± 15.63	**90.88± 15.19**	87.95± 19.57
Pairwise Rep. (LLM) ↑	**66.67**	28.52	-	-
Pairwise Factualness (LLM) ↑	**60.16**	33.07	-	-
Pairwise Rep. (Human) ↑	**48.57**	21.43	-	-
Pairwise Factualness (Human) ↑	**39.52**	22.62	-	-
IAA - ROUGE-L ↑	24.92 ± 7.82	**28.98 ± 9.29**	**32.67 ± 12.56**	28.62 ± 12.52
IAA-BARTScore ↑	**-2.71 ± 0.53**	-2.55 ± 0.55	**-2.72 ± 0.61**	-2.93 ± 0.73
Pairwise BARTScore ↑	-2.81±0.51	**-2.56± 0.52**	**-2.85± 0.63**	-2.88± 0.70

Table 1: RQ1 Evaluation : Speech versus Transcript-based Summaries by expert annotators

The only result that did not exhibit statistically significant influence from the source modality was fluency. This is reasonable, as the fluency of the model is determined more by the language proficiency of the individual writing the summary than the content.

For **RQ1**, we find that *speech-based summaries seem to be more selective, factually consistent, and perhaps more abstractive than transcript-based summaries, while transcript-based summaries contain more information extracted directly from the source than their speech-based counterparts*.

6.2 RQ2: Impact of Transcript Errors on Human Summarization

Table 2 compares the summaries based on the subset of 100 recordings. It is interesting to compare the summaries based on ASR transcriptions to the summaries based on manual transcriptions. Nearly all indicators of informativeness are significantly higher for the manual transcriptions than the ASR transcriptions. This is to be expected, as propagated recognition errors can directly affect measures like Entity F1, factual consistency, and semantic similarity. Coherence is similar for the two types of transcripts, which is consistent with the idea that the ability to reference text allows the summarizer to capture the details and structure of the source.

Comparing ASR-transcript-based summaries to speech-based summaries, we see similar trends to those in RQ1. Speech-based summaries are still shorter, more selective, less informative, more factually consistent, and likely more abstractive than ASR-transcript-based summaries. In general, LLM evaluation indicates that ChatGPT prefers transcript-based summaries over speech-based summaries, even if the source transcripts contain errors. Additional analysis of the effect of WER on Entity F1 can be found in Appendix D.2.

For **RQ2**, we find that *transcription errors decrease the informativeness and coherence of transcript-based summaries, but do not significantly impact fluency*.

6.3 RQ3: Expert Annotation versus Non-Expert Annotation

From Table 3, we observe that the speech-based summary length from expert annotation is significantly lower than its non-expert counterpart, while the expert speech-based and transcript-based summary lengths are very different. The word vocabulary size and the percentage of novel words are lower in the expert summaries (both speech and transcript source) compared to the non-expert summaries.

Extractiveness is lower for non-expert speech-based and transcript-based summaries compared to the expert ones, referring to the fact that the non-experts do not tend to use phrases directly from the source whereas the experts do. Entity

Metric	Transcript Overall	Speech Overall	ASR Overall
Summary Length	**85.29±40.58**	56.94±36.89	73.67±36.89
Compression Ratio	0.40± 0.73	**0.52± 0.81**	0.27± 0.85
Extractiveness (ROUGE-L with Transcript)	21.94±10.95	**19.17±11.60**	19.25±9.33
Entity F1	**35.17 ±18.93**	26.28 ± 15.65	30.24±17.67
BERTScore with Transcript	**85.92 ± 2.55**	85.56±2.55	85.49±2.45
BARTScore with Transcript	**-3.33 ± 0.48**	-3.44±0.47	-3.41 ±0.45
Retrieval Accuracy	**85.42± 35.29**	71.88± 44.96	77.60± 41.68
MRR	**89.09± 27.55**	79.44 ±34.27	83.49±32.11
Factual Consistency (UniEval)	83.56 ± 14.45	**86.16±13.83**	83.48±14.25
Fluency (UniEval)	94.31 ± 3.62	94.23 ± 3.90	**94.36 ± 3.72**
Coherence (UniEval)	**91.51 ± 14.90**	88.64 ± 16.04	90.81 ± 13.34
Pairwise Rep. (LLM)	**40.06**	21.57	33.81
Pairwise Factualness (LLM)	**37.76**	22.74	32.81
Pairwise BARTScore	-2.81± 0.48	**-2.59 ± 0.51**	-2.79 ± 0.51

Table 2: Impact of Transcript Errors on Human Summarization

F1 score is higher for non-expert speech-based and transcript-based summaries, which is because non-expert summaries are on average longer than the expert ones and capture more of the entities in the source. BERTscore, BARTscore, retrieval accuracy, and MRR are higher for the expert summaries than the non-expert summaries (for both speech and transcript sources) referring to the fact that non-expert summaries are less informative than the expert summaries. Factual consistency is lower for non-expert summaries pointing to the fact that expert summaries are more factually consistent. Fluency, coherence, and pairwise BARTscores are lower for non-expert summaries demonstrating the reliability of expert summaries.

We ran t-tests to conclude whether the non-expert speech-based and transcript-based summaries belong to the same distribution and observed based on BARTScore and other metrics that the p-value was high (see Table 4 in Appendix D.4. This potentially indicates that non-expert annotations based on speech and transcripts may be very similar and that some annotators may have used similar methods or tools for the two types of annotations. Based on a human examination, it appears that some annotators did use generative AI to respond to annotation requests through Amazon Mechanical Turk, however, such conclusions may not be definitive since such models can generate human-like speech. This makes non-expert summaries less reliable for analysis of speech-based versus transcript-based summaries, and hence, we decided against running the LLM and human evaluation for these non-expert summaries.

For **RQ3**, we find that *expert annotations are more informative, coherent, fluent, and less information variant across speech-based and transcript-based summaries*.

7 Conclusion

In this paper, we take the first steps towards understanding human annotation for speech summarization by proposing and assessing different approaches. Annotators can either read a transcript of spoken content or listen to the audio recording to produce an abstractive speech summary. We formulate three research questions to assess differences in summaries resulting from variations in source modality (listening to audio versus reading a transcript of spoken content), the nature of transcription (manual versus ASR-based), and the expertise of annotators. Extensive analysis using automatic and human evaluation reveals that speech-based summaries are more information-selective, factually consistent, and abstractive compared to transcript-based summaries. Analysis of the impact of ASR on transcript-based summarization shows that errors in transcript generation decrease the informativeness and coherence of transcript-based summaries. Finally, analysis of expert and non-expert annotations demonstrates higher informativeness, coherence, and fluency across expert annotations.

Through this process, we develop and release a new human-annotated dataset to enable reproduction and encourage further research in the area. We hope that this study galvanizes further research into human summarization, and powers informed model design for automatic speech summarization.

Ethics Statement

We have made a concerted effort to make a positive ethical impact during this work. In data collection, we received permission through the proper channels (IRB) and treated all participants as fairly and respectfully as we could. Through the annotation process, we have valued diversity of opinion and annotator backgrounds and compensated any workers at wages beyond the minimum wage in the US. We have also tried to remove elements of bias by using automatic and manual verifications as appropriate. In the technical work, we have been honest and transparent. We have also carefully considered the implications and potential applications of our work and findings, and we do not believe they present any significant risks to individuals or society as a whole.

Limitations

A challenge in the context of this work is that existing methods are designed to evaluate summaries of written text, and may have a bias against speech-based summaries. While some of the methods presented here may begin to address this bias (e.g., retrieval accuracy and IAA metrics), the problem persists. Another potentially significant limitation of this paper is the results regarding non-expert summaries. As stated previously, it is possible that the participants on Amazon Mechanical Turk used generative AI software to produce the summaries, so the non-expert analysis is probably not representative of real non-expert summaries.

Broader Impact

Our work represents the first of hopefully many forays into understanding human annotation for speech summarization. The notion of understanding how humans perform the task, and what would enable them to produce the best possible annotations is both important and understudied. We hope that our new dataset and the analysis we have presented in this work serve to ignite discussions around community standards for speech summarization annotation. Further, we hope that the strengths and weaknesses of summaries based on different approaches inform developers on how to design and deploy the next generation of speech summarizers and more generally speech foundation models.

References

S Banerjee and A Lavie. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Han Feng, Sei Ueno, and Tatsuya Kawahara. 2020. End-to-end speech emotion recognition combined with acoustic-to-word asr model. In *Interspeech*, pages 501–505.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like Summarization Evaluation with ChatGPT. *arXiv preprint arXiv:2304.02554*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Makoto Hirohata, Yosuke Shinnaka, Koji Iwano, and Sadaoki Furui. 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1065. IEEE.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. 2002. Automatic speech summarization applied to english broadcast news speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–9. IEEE.

Chiori Hori, Takaaki Hori, and Sadaoki Furui. 2003. Evaluation method for automatic speech summarization. In *INTERSPEECH*.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Roshan Sharma, Kohei Matsuura, and Shinji Watanabe. 2023. Speech summarization of long spoken document: Improving memory efficiency of speech/text encoders. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. 2021. Attention-based multi-hypothesis fusion for speech summarization. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 487–494.

Anam Ahmad Khan, Sadia Nawaz, Joshua Newn, Ryan M Kelly, Jason M Lodge, James Bailey, and Eduardo Velloso. 2022. To type or to speak? the effect of input modality on text understanding during note-taking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Inderjeet Mani. 1999. *Advances in automatic text summarization*. MIT press.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Takatomo Kano, Atsunori Ogawa, and Marc Delcroix. 2023a. Transfer learning from pre-trained language models improves end-to-end speech summarization.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Atsunori Ogawa, Marc Delcroix, and Ryo Masumura. 2023b. Leveraging large text corpora for end-to-end speech summarization. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, Dennis Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Shruti Palaskar, Ruslan Salakhutdinov, Alan W. Black, and Florian Metze. 2021. Multimodal Speech Summarization Through Semantic Concept Learning. In *Proc. Interspeech 2021*, pages 791–795.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Dana Rezazadegan, Shlomo Berkovsky, Juan C Quiroz, A Baki Kocaballi, Ying Wang, Liliana Laranjo, and Enrico Coiera. 2020. Automatic speech summarisation: A scoping review. *arXiv preprint arXiv:2008.11897*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2009. Emotion recognition from speech: putting asr in the loop. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4585–4588. IEEE.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Roshan Sharma, William Chen, Takatomo Kano, Ruchira Sharma, Siddhant Arora, Shinji Watanabe, Atsunori Ogawa, Marc Delcroix, Rita Singh, and Bhiksha Raj. 2023a. Espnet-summ: Introducing a novel large dataset, toolkit, and a cross-corpora evaluation of speech summarization systems. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. 2022. End-to-end speech summarization using restricted self-attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076.

Roshan Sharma and Bhiksha Raj. 2022. Xnor-former: Learning accurate approximations in long speech transformers. *arXiv preprint arXiv:2210.16643*.

Roshan Sharma, Kenneth Zheng, Siddhant Arora, Shinji Watanabe, Rita Singh, and Bhiksha Raj. 2023b. Bass: Block-wise adaptation for speech summarization.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Prashanth Gurunath Shivakumar, Mu Yang, and Panayiotis Georgiou. 2019. Spoken language intent detection using confusion2vec. *arXiv preprint arXiv:1904.03576*.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Gerald Stollnberger, Astrid Weiss, and Manfred Tscheligi. 2013. Input modality and task complexity: Do they relate? In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 233–234. IEEE.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large Language Models are Diverse Role-Players for Summarization Evaluation. *arXiv preprint arXiv:2303.15078*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the human language technology conference of the NAACL, main conference*, pages 447–454.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Appendix

A Expert annotation details

A.1 Annotation guidelines

Figure 3 shows the detailed guideline we provided for expert annotation to third-party vendors.

The summaries generated for this task have fewer limitations on style and content. Generally speaking, the summaries ought to capture the main topic of the interaction, but we're otherwise not too caught up on prescribing a particular way of writing, such as maintaining a particular tense throughout.
Annotator selection - A single annotator should only interact with each segment once, either as an audio file or as a text transcript.
Summary length
- There is no hard limit on maximum summary length, but we expect at least two sentences per summary, minimum and expect 50 to 80 words.
Summary content
- The summary should convey the main topic or overarching message contained in the interaction.
- There is nothing that should never be included, but we prefer for annotators to summarize the overall message rather than attempting to incorporate direct quotes from the interaction.
- Do not worry about doing outside research to verify content. It is preferred that the summaries are written with few prior constraints and without overthinking (summaries should be fairly spontaneous).
Summary style
- There is no strict template to follow for these summaries.
- Summaries should maintain a neutral tone.
- Summaries can be as simple or complex as needed to capture the overall message of the interaction.
- While we expect the summaries to be written in complete sentences with proper grammar, no particular tense needs to be maintained.
- Names and places mentioned in the conversations (including the names of conversation participants) can be used, and accurate spellings for those do not matter.

Figure 3: Annotation guideline

A.2 Annotator Compensation and Statistics

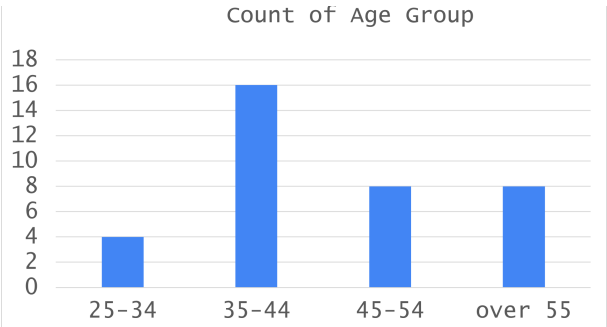

Figure 4: Age distribution for the Expert Annotators

In total, there are 36 expert annotators. The expert annotators are all American national and native English speakers, proficient in the language, therefore they don't have any difficulty performing the summarization task. Among the 36, 34 are female, one identifies as non-binary and one prefers not to say. There exists some variation in their age group, which is shown in the figure 4.

The third-party vendor passed in-house qualitative assessments of summary outputs from a pilot annotation and contracted to the company's primary third-party vendor for summary annotation. The statement of work has agreements between the company and the vendor, and the company paid 9,341.76 US dollars for a total of 4,208 expert summary annotations. This is a competitive compensation for such expertise, and well beyond the minimum wage within the United States of America, where the annotation was carried out. Annotators consented to the process through the work agreement and were aware of the plans to release the annotations to the public.

A.3 Example Source and Associated Summaries

Source

Towns like Bluffton say they're working to promote communities of inclusiveness, but what does that really mean, and will people of color come? Jim Hunt is president of the National League of Cities based in Washington, D.C. He heads up the group's partnership for working toward inclusive communities. Mr. Hunt joins us from West Virginia Radio in Morgantown, West Virginia..And joining us via phone is Harvard law professor Charles Ogletree. Professor Ogletree leads the law schools' Charles Hamilton Houston Institute for Race and Justice. Gentlemen, great to have you on the program. Professor, let me start with you. Give us a quick historical view, of you will, of these, quote, sundown towns..Well, thanks, Ed. As you know, when I was doing research on the Tulsa race riots from 1921, I was astounded to learn of the literally hundreds of towns - now we've discovered thousands of towns around America - in the South, but not exclusively in the South, where blacks were told leave before sundown. There were sirens, there were notices, and the consequences of staying in those towns were death or other serious bodily injury..And the good news is that that is largely historical, but it's a frightening sense that people could not live in the community. They could work there, they could visit there, but they couldn't' be there after dark because of the strongly held feelings about segregation. And sundown towns were just that: if you're black, get out of town before the sun goes down..Mm hmm. Jim Hunt, while those sirens may have gone away, there are certainly still areas in this country where when the sun goes down, minorities know that this is an area - a region you should not be in. That continues today..Right, Ed. And I think when we look around - and, obviously, the inclusive agenda is a broader agenda than just race, but clearly race plays a very significant part in the program..Mr. Hunt, why are we seeing this sense of inclusiveness? Is it totally altruistic? Are there financial benefits to these cities? Why are we seeing this rush now to try to move to include minorities?.Well, and I think when we look across the board of what's happening across the country - when we look at Dr. Richard Florida's work on the rise of the creative class, and some of the economic development needs of cities throughout America - we recognize that if we're not going to be inclusive, we're going to suffer from an economic development perspective.

Expert Speech Summary

Some towns in America used to be known as sundown towns because black people were told to leave town before sundown or they could be harmed. Some cities today have programs that focus on being inclusive to minorities.

Expert Text Summary

Jim Hunt, President of the National League of Cities, and Harvard Law Professor Charles Olgetree discuss sundown towns and why some towns are working to be more inclusive of minorities. A sundown town was a town, mostly in the south, where blacks were allowed work and visit, but were not allowed to be there after sundown. If they were, the could face serious injury or even death. While these town have mostly gone away, there are still areas in the country that minorities know they should not be in when the sun goes down. Many towns are now working to be more inclusive of minorities and part of the reason has to do with the financial benefits they'll receive.

Non-expert Speech Summary

Jim Hunt, President of the National League of Cities, and Harvard law professor Charles Ogletree discuss the historical and contemporary implications of "sundown towns," where African Americans were forced to leave before sunset under threat of violence. While this practice is largely historical, there are still areas where minorities feel unsafe after dark. They emphasize the importance of inclusivity for economic development and societal well-being.

Non-expert Text Summary

The National League of Cities is working on inclusive communities, addressing historical issues like "sundown towns," where minorities were told to leave before dark under threats of violence. Harvard Law Professor Charles Ogletree discusses the historical prevalence of such towns, emphasizing the progress made. National League of Cities President Jim Hunt highlights the economic benefits of inclusivity in cities, linking it to overall economic development. While acknowledging past discrimination, the focus is on building inclusive communities for the future.

Figure 5: Example source and summaries.

B Non-Expert annotation details

B.1 IRB Review and Informed Consent

Non-expert annotations were carried out using Amazon Mechanical Turk. The entire plan for this research paper was presented to an Institutional Review Board (IRB) and received assent before starting data collection. A consent form (shown next) was prepared and presented to prospective annotators, and conditional on their acceptance, they were allowed to take on the annotation task.

[redacted]

Online Consent Form

This task is part of a research study conducted by [redacted]

The task involves typing short summaries of news excerpts. Each summary will be based on either an audio recording of a news broadcast or a text transcript of a news broadcast. Each set of prompts provided to a participant includes two audio recordings and two transcripts, so four summaries will be written for each set. Each summary should take approximately four minutes to complete, so each set should take about 16 minutes.

Purpose
The purpose of the research is to understand whether human annotators write different summaries depending on whether they listen to a recording or read the transcription.

Procedures
We will ask participants to listen to recordings and read passages, and then write brief textual summaries based on what they heard or read. We will assign two (2) random audio and two (2) passages for every participant to summarize within 60-80 words.

Participant Requirements
Participation in this study is limited to individuals ages 18 and older residing in the United States. Participants should be proficient in English reading, writing, and listening. Participants should not have any hearing impairments that significantly inhibit their ability to listen to audio recordings.

Risks
The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities. Participants should be aware of the potential for a security breach at Amazon Mechanical Turk, which may result in unwanted access to personal information.

Benefits
Participants will learn more about writing and develop their summarization skills. They will have learnt about topics discussed in radio interviews, and practiced summary writing based on input knowledge.

Compensation & Costs
Compensation will be provided via Amazon Mechanical Turk at a rate of $2.67 per question set (approximately $10 an hour). Each question set should take approximately 16 minutes to complete.

Future Use of Information

In the future, once we have removed all identifiable information from your provided information, we may use the data for our future research studies, or we may distribute the data to other researchers for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality
The only identifiable personal information that will be collected for this task will be your Amazon Mechanical Turk information.

By participating in this research, you understand and agree that may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your consent form will be stored in a secure location on property and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by and published and/or disclosed by o others outside of However, your name, address, contact information and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by Note that per regulation all research data must be kept for a minimum of 3 years.

Right to Ask Questions & Contact Information
If you have any questions about this study, you should feel free to ask them by contacting

If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by mail, phone or e-mail in accordance with the contact information listed above.

If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the

Voluntary Participation
Your participation in this research is voluntary. You may discontinue participation at any time during the research activity. You may print a copy of this consent form for your records.

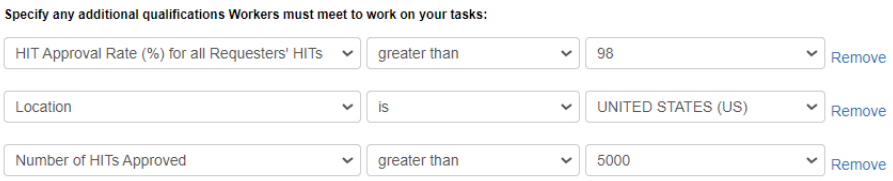

Figure 6: Filters to select workers.

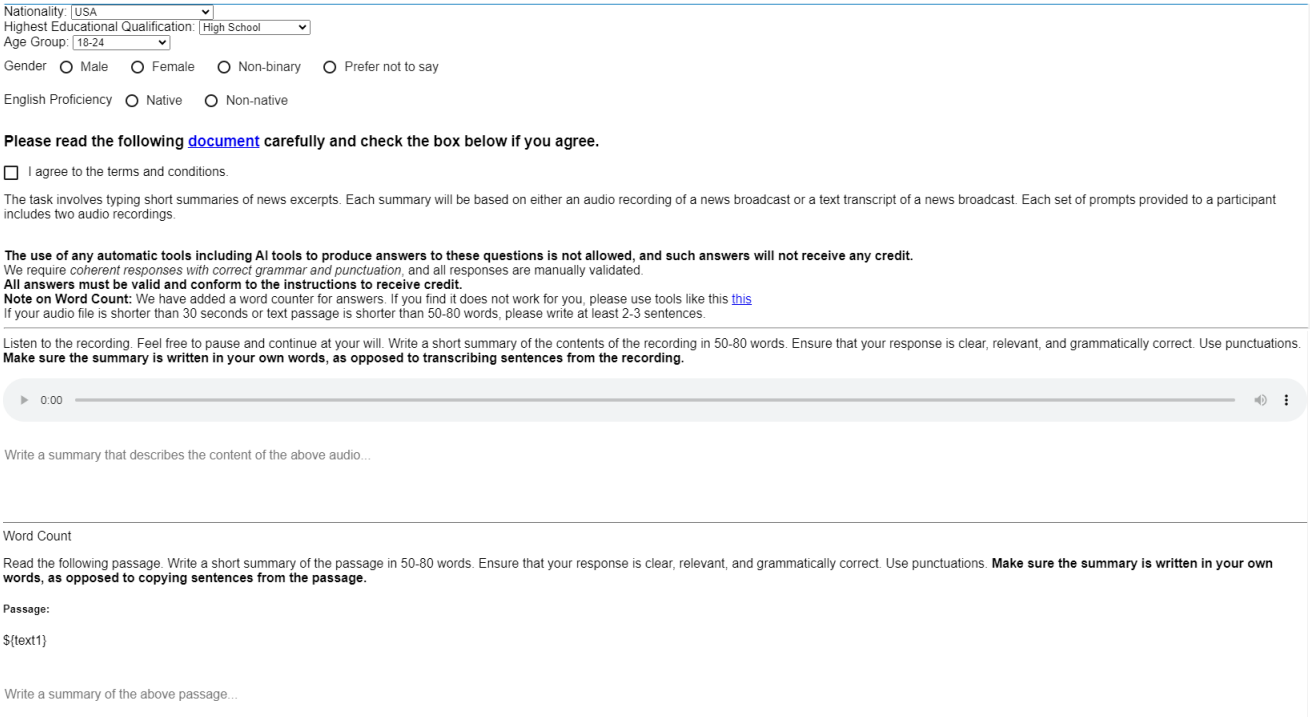

Figure 7: AMT design. Top figure shows the initial instructions presented to the workers. Bottom image shows how the audio and text source texts are presented.

B.2 Compensation

Workers were compensated at a rate of $10/hour for their work, which is above the minimum wage in the United States of America, where the workers reside.

B.3 Worker Selection

Only the workers who agree to the consent form are allowed to perform the task on AMT. Furthermore, workers who reside in the US, with an AMT HIT approval rate of greater than 98% and total HITS approved greater than 5000 are selected for the task. Figure 6 shows the criterion used.

B.4 HIT Design and Guidelines

Figure 7 shows the layout of the AMT design, how the instructions are presented and how source audio and text are provided. The worker then proceeds to write the summaries in the text box below the source.

As part of the instructions, the use of generative AI was prohibited to answer the questions.

Annotators were asked to write summaries based on the provided inputs while ensuring that the response was coherent and grammatical, and written in their own words. All responses are validated manually for conformity to the instructions, and invalid responses are rejected and re-assigned to a new annotator.

B.5 Demographics of the AMT Workers

Figure 8 shows the demographics of the AMT workers. The workers are mostly bachelor graduates, from the 25-34 age bracket, more female than other genders. 952 of the workers are native English speakers

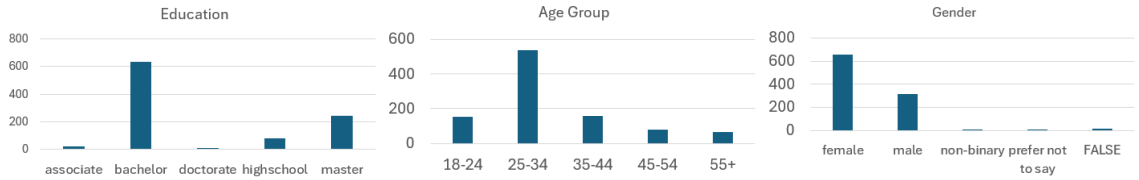

Figure 8: The demographics of the AMT workers are shown in the figure above. The variation in the education level, age groups and gender are shown as histograms.

and 41 are not. 1000 of the workers are American nationals, whereas 3 are from elsewhere. We do not collect any information from the workers that can uniquely identify them.

B.6 Data Release

We plan to release the summaries annotations collected from experts and non-experts for both modalities under the Apache 2.0 license, subject to the terms of the NPR license. Access to the data will be in accordance with the usage policy of NPR and the data is meant to be used for research purposes only.

C Interview corpus

Interview is the largest open-domain corpus for speech summarization to date, with a total duration of 4,354 hours. Speech recordings contain spontaneous speech along with real audio noises that represent music, and events inside and outside the studio corresponding to discussion topics. The dataset has 49.4k recordings that span multiple topics and domains, with an average recording length of 5.28 minutes. The transcription on average contains 886.4 words and the summary contains 40.1 words on average, representing a compression ratio of 22.09 between the transcript and summary. Around 48% of words in the summary are not present in the transcription, making this a relatively challenging task. On average, each recording has 24 turns and 4 speakers.

D Results

D.1 LLM Evaluation questions

```
The following are two summaries of the given source text:
Summary 1: [Summary 1]
Summary 2: [Summary 2]
Source: [Source]
```

Factual Consistency Question

```
Answer the following multiple-choice question:
What is the most appropriate statement about the summaries?
(a) Summary 1 is more factually consistent with the source text than Summary 2.
(b) Summary 2 is more factually consistent with the source text than Summary 1.
(c) Both Summary 1 and Summary 2 are equally factually consistent with the source text.
Give your answer as "Answer: a", "Answer: b", or "Answer: c". Then, start a new line and explain your reasoning in a
single paragraph following your answer.
```

Source Representation Question

```
Answer the following multiple choice question:
What is the most appropriate statement about the summaries?
(a) Summary 1 is more representative of the source text than Summary 2.
(b) Summary 2 is more representative of the source text than Summary 1.
(c) Both Summary 1 and Summary 2 are equally representative of the source text.
Give your answer as "Answer: a", "Answer: b", or "Answer: c". Then, start a new line and explain your reasoning in a
single paragraph following your answer.
```

Figure 9: Template for questions presented to ChatGPT. Similar questions are used to solicit human scores.

D.2 Impact of ASR Errors on Summary Quality

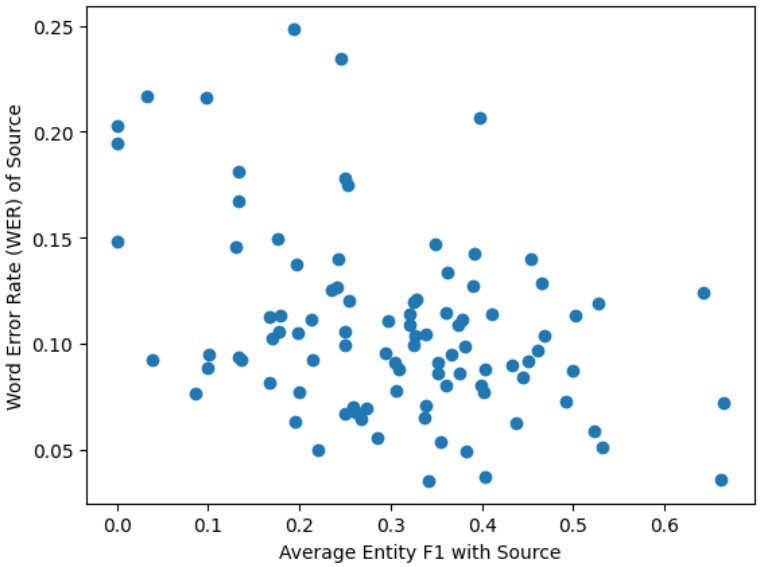

Figure 10: Scatterplot showing variance of entity F1 of summary compared to source reference transcript with Word Error Rate (WER) of audio transcription

D.3 Expert Annotations

Metric	Transcript Overall	Speech Overall	p-value
Summary Length	**78.58** ± **27.72**	56.95 ± 18.30	
Compression Ratio	3.76 ± 2.20	**5.20** ± **3.06**	8.83E-86
Word Vocabulary Size	20,408	16,211	-
Novel Words %	24.63 ± 13.57	**25.66** ± **13.22**	5.80E-02
Extractiveness↓	20.70 ± 10.53	**18.57** ± **10.88**	1.09E-08
Entity F1↑	**33.63** ± **29.83**	26.21 ± 26.90	1.34E-30
BERTScore ↑	**85.66** ±**2.45**	85.32 ±2.42	1.54E-05
BARTScore ↑	-3.37 ±0.47	-3.46 ± 0.45	7.96E-11
Retrieval Accuracy ↑	**68.86 +** ± **46.31**	59.04 ± 49.75	1.36E-19
MRR ↑	70.23 ± 16.08	**75.74** ± **16.29**	8.48E-52
Factual Consistency (UniEval) ↑	0.83 ± 0.16	**0.85**± **0.16**	1.90E-04
Pairwise Rep. (LLM) ↑	**66.67**	28.52	-
Pairwise Factualness (LLM) ↑	**60.16**	33.07	-
Pairwise Rep. (Human) ↑	**48.57**	21.43	-
Pairwise Factualness (Human) ↑	**39.52**	22.62	-
Fluency (UniEval) ↑	**94.02**± **4.54**	94.01± 5.26	9.83E-01
Coherence (UniEval)↑	**91.82**± **12.97**	89.34± 15.63	5.42E-08
Pairwise BARTScore ↑	-2.81±0.51	**-2.56**± **0.52**	1.67E-101

Table 3: RQ1 Evaluation : Speech versus Transcript-based Summaries by expert annotators

D.4 Non-Expert Annotations

Metric	Text Overall	Audio Overall	p-value
Summary Length	75.35 ± 17.70	77.42 ± 33.85	0.02
Compression Ratio	3.69 ± 1.95	3.66± 1.86	0.78
Word Vocabulary Size	21511	21676	
Novel Words %	37.47 ± 13.38	38.45± 14.21	0.09
Extractiveness (ROUGE-L with Transcript)	**18.37± 7.26**	18.38 ±11.29	0.98
Entity F1	**34.51 ± 17.28**	30.09± 17.72	7.37E-14
BERTScore with Transcript	**85.28± 2.32**	85.06 ±2.76	0.01
BARTScore with Transcript	**-3.41± 0.44**	-3.43± 0.53	0.30
Retrieval Accuracy	**71.82 ± 44.99**	65.78 ± 47.44	1.04E-04
MRR	**76.56 ± 71.01**	71.01 ± 41.57	5.80E-01
Factual Consistency (UniEval)	**82.08 ± 16.90**	78.57± 19.79	1.49E-08
Fluency (UniEval)	**92.90± 8.21**	92.65 ±9.35	0.39
Coherence (UniEval)	**90.88± 15.19**	87.95± 19.57	6.83E-07
Pairwise BARTScore	**-2.85± 0.63**	-2.88± 0.70	0.04

Table 4: Non-Expert Annotation