IEICE TRANS. INF. & SYST., VOL.E101–D, NO.3 MARCH 2018

PAPER

Effects of Automated Transcripts on Non-Native Speakers' Listening Comprehension

Xun CAO[†a)], **Naomi YAMASHITA**[††], ***Nonmembers***, ***and*** **Toru ISHIDA**[†], ***Fellow***

SUMMARY Previous research has shown that transcripts generated by automatic speech recognition (ASR) technologies can improve the listening comprehension of non-native speakers (NNSs). However, we still lack a detailed understanding of how ASR transcripts affect listening comprehension of NNSs. To explore this issue, we conducted two studies. The first study examined how the current presentation of ASR transcripts impacted NNSs' listening comprehension. 20 NNSs engaged in two listening tasks, each in different conditions: C1) audio only and C2) audio+ASR transcripts. The participants pressed a button whenever they encountered a comprehension problem, and explained each problem in the subsequent interviews. From our data analysis, we found that NNSs adopted different strategies when using the ASR transcripts; some followed the transcripts throughout the listening; some only checked them when necessary. NNSs also appeared to face difficulties following imperfect and slightly delayed transcripts while listening to speech - many reported difficulties concentrating on listening/reading or shifting between the two. The second study explored how different display methods of ASR transcripts affected NNSs' listening experiences. We focused on two display methods: 1) accuracy-oriented display which shows transcripts only after the completion of speech input analysis, and 2) speed-oriented display which shows the interim analysis results of speech input. We conducted a laboratory experiment with 22 NNSs who engaged in two listening tasks with ASR transcripts presented via the two display methods. We found that the more the NNSs paid attention to listening to the audio, the more they tended to prefer the speed-oriented transcripts, and vice versa. Mismatched transcripts were found to have negative effects on NNSs' listening comprehension. Our findings have implications for improving the presentation methods of ASR transcripts to more effectively support NNSs.

key words: *listening comprehension problems, automatic speech recognition (ASR) transcripts, non-native speakers (NNSs), eye gaze*

1. Introduction

Listening to the speech of native speakers (NSs) is a challenging task for non-native speakers (NNSs) [2], [3]. In such real-time settings as audio conferences (as a listener) or lectures, NNSs often do not have the chance to pause or repeat the speech to solve their comprehension problems. Such difficulties or confusions can easily accumulate and lead to speech misunderstandings.

Real-time transcripts generated by automatic speech recognition (ASR) technologies hold the potential to help NNSs improve their listening comprehension by providing them supplemental information to understand the speech [4], [5]. If such a technology was installed into portable devices of NNSs, they could view the ASR transcripts on the screen while listening to the speech. However, previous research has shown that ASR transcripts often place an extra burden on them [6]. Since NNSs are already burdened by processing audio information (i.e., NS speech), providing them with textual information (i.e., ASR transcripts) may further overwhelm them with excessive information [5].

Our goal is to design an ASR-based interface, which can support NNSs' listening comprehension more effectively. To reach our goal, we conducted two studies. In Study 1, we examined how the current presentation of ASR transcripts impacted NNSs' listening comprehension. We conducted a laboratory experiment with 20 NNSs who engaged in two listening tasks in different conditions: C1) audio only and C2) audio+ASR transcripts. Participants pressed a button whenever they encountered anything about which they were unclear or did not understand (i.e., comprehension problems), and described each problem in the subsequent interviews. To better understand how the NNSs used the ASR transcripts under the audio+ASR transcript condition, we recorded their eye movements using an eye-tracker. Through an exploratory analysis of the experiment data, we found that ASR transcripts helped the NNSs solve certain problems (e.g., "do not recognize words they know"), but imperfect ASR transcripts (e.g., errors and no punctuation) sometimes confused the NNSs and even generated new problems. We also found that NNSs adopted different strategies when using the ASR transcripts; some followed the transcripts throughout the listening; some only checked them when necessary. Post-task interviews and gaze analysis of the participants revealed that they did not have enough time or cognitive resources to fully exploit the transcripts. For example, they had difficulty concentrating on listening/reading or shifting between the two.

Based on our findings from Study 1, the second study explored how different display methods affect NNSs' listening experiences and their use of transcripts. We focused on two display methods: 1) accuracy-oriented display which shows the transcripts only after the completion of speech input analysis - transcripts often appear as chunks after some delay, and 2) speed-oriented display which shows interim analysis results of speech input which are often corrected by the time speech analysis is complete - transcripts normally appear word-by-word immediately upon speech input but the early appearing text often contains errors. They are often replaced with other words a few times during the ana-

Manuscript received August 7, 2017.
Manuscript publicized November 24, 2017.
[†]The authors are with the Department of Social Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with NTT Communication Science Laboratories, NTT Corporation, Kyoto, 619–0237 Japan.
a) E-mail: xun@ai.soc.i.kyoto-u.ac.jp
DOI: 10.1587/transinf.2017EDP7255

lytic process, and finally converge to the same transcript as in the accuracy-oriented display. The former method prioritizes accuracy over speed while the latter prioritizes speed over accuracy. We conducted a laboratory experiment with 22 NNSs. In the experiment, the NNSs engaged in two listening tasks with ASR transcripts presented via the two display methods. Again, NNSs' eye movements were recorded using an eye-tracker. Through analysis of the experiment data, we found that the more the NNSs paid attention to listening to the audio, the more they tended to prefer the speed-oriented transcripts, and vice versa. Mismatched transcripts were found to have negative effects on NNSs' listening comprehension.

In the remainder of this paper, we first review previous studies and discuss how our study extends them. We then describe our studies about how ASR transcripts impact NNSs' listening experiences. We conclude with a discussion of the implications of our findings for improving the presentation methods of ASR transcripts so that they can support NNSs more effectively.

2. Background

2.1 Real-Time Listening Comprehension Problems of NNSs

NNSs often face comprehension difficulties when listening to the speech of NSs. Researchers have examined the problems faced by NNSs from different perspectives. Most of the previous studies explored the factors that influence second language listening [2], [7]. A representative work by Rubin extensively reviewed the research on second language listening comprehension and attributed the factors that affect listening comprehension into five characteristics: text characteristics, interlocutor characteristics, task characteristics, listener characteristics, and process characteristics [7]. On the other hand, Goh investigated NNSs' listening comprehension from a different perspective. She classified the listening comprehension problems faced by NNSs into ten categories (Table 1). In her study, 40 non-native students wrote weekly diaries and explained the listening comprehension problems they faced during lectures [8].

2.2 Supporting NNSs' Listening Comprehension with ASR Transcripts

According to previous research, real-time transcripts generated by ASR technologies hold the potential to facilitate the listening comprehension of NNSs [4]. ASR transcripts provide textual information that can complement audio speech and improve the comprehension of NNSs [5], [9]. Pan et al. investigated how the quality of ASR transcripts impact comprehension and found that a 20% word-error-rate (WER) was the critical point for transcripts to be acceptable, and at a 10% WER, comprehension performance significantly improved compared to a no-transcript condition [4].

Table 1 Listening comprehension problems identified in Goh's work [8]

Problems
1. Do not recognize words they know
2. Unable to form a mental representation from words heard
3. Cannot chunk streams of speech
4. Neglect the next part when thinking about meaning
5. Do not understand subsequent parts of input because of earlier problems
6. Concentrate too hard or unable to concentrate
7. Understand words but not the intended message
8. Confused about the key ideas in the message
9. Miss the beginning of texts
10. Quickly forget what is heard

Yao et al. compared the NNS comprehension performance among three conditions (no-transcript, perfect transcripts with a 2-second delay, and transcripts with a 10% WER and a 2-second delay). The comprehension performance in the latter two conditions was significantly better than that in the no-transcript condition [5].

Despite the positive effects of introducing ASR transcripts, previous research also reported that NNSs sometimes get overwhelmed when they simultaneously listen to speech and read ASR transcripts that contain errors and delays [5], [6]. In addition, errors and delays negatively impacted how NNSs perceived the value of the ASR transcripts [4], [5].

Overall, the previous studies identified the usefulness of ASR transcripts for supporting NNS listening comprehension and the risk of placing an extra burden on NNSs. However, we still lack a detailed understanding of how NNSs benefit from ASR transcripts (e.g., what types of listening comprehension problems could be solved) and what are the difficulties of using them (e.g., the factors that hinder them from solving their problems).

3. Study 1

According to Goh, NNSs encounter various types of comprehension problems when listening to native speech. Among them, we expect that certain types of problems such as "don't recognize words they know" will be solved using ASR transcripts, but others will not. In addition, since NNSs are often overburdened by processing speech input, ASR transcripts may not always help them solve their problems. Therefore, we pose the following research questions:

RQ1: What types of listening comprehension problems can be solved by reading ASR transcripts? What hindered NNSs from solving these problems?

While previous work has explored the feasibility of supporting NNS listening comprehension in real time with ASR transcripts, little work has scrutinized how NNSs used transcripts and the difficulties they faced. We believe such knowledge is important for improving the presentation methods of ASR transcripts so that they can support NNSs more effectively. Thus, we pose the following re-

search questions:

RQ2: How do NNSs use ASR transcripts, and what are the difficulties?

3.1 Method

We recruited twenty non-native English speakers for our study, including ten females and ten males. Their mean age was 25.9 (SD = 2.41). All spoke Chinese as their first language. Their Test of English for International Communication (TOEIC) scores ranged from 690 to 950 (M = 823, SD = 95.05). Participants engaged in two listening tasks in different conditions:

- Without-transcript: only audio was presented
- With-transcript: both audio and ASR transcripts were presented

The experiment used a within-subject design. The conditions were counterbalanced across participants to minimize the order effects. In each condition during the listening task, the participants pressed a button to indicate when they encountered a comprehension problem. Pressing the button marked (in the lecture transcripts) specific places that were visited later to explain the details of the problems. We chose this "pressing a button" method because it has low-overhead, as suggested by previous work [10]. In addition, the method allows us to record the problems faced by NNSs in real time and simultaneously keep the task close to actual listening experiences.

Four audio clips from the Test of English as a Foreign Language (TOEFL) exam were chosen as task materials. Two clips were conversations and the other two were lectures, both from academic settings. The length of the clips varied from two to five minutes. On average, each clip contained about 600 words. Two clips (one conversation and one lecture) were randomly chosen for each experiment condition.

To remove extraneous factors which might affect the appearance of the transcripts, we recorded the videos of real-time transcripts for each audio clip and used them throughout the experiment. Under the with-transcript condition, participants watched the recorded videos. The transcripts were generated by Google speech recognition API. The word error rate (WER) of the ASR transcripts was about 10% on average. This WER was considered suitable for our study because previous work suggested that at a 10% WER, comprehension performance significantly improved compared to a no-transcript condition [4]. On average, 62% of the errors were substitutions, 30% were deletions, and 8% were insertions.

The transcripts were displayed in a streaming mode, i.e., the words appeared as the speech stream flowed forward. Two lines of transcripts were shown: one line for ongoing speech, and another for previous speech that provided participants with an optional review opportunity. The text was presented in an Arial font at 32 pt.

To understand how ASR transcripts were used during listening, participants' eye movements were recorded using a Tobii TX300 eye-tracker, which is composed of an eye-tracker unit and a 23-inch, 1920 × 1080 widescreen monitor (Fig. 1). The participants were positioned at a viewing distance of 65 cm from the monitor. The eye-tracker collects gaze data at 300 Hz and allows large head movements. The gaze data were logged by Tobii Studio. Before starting the tasks, we performed a 9-point calibration of the eye-tracker for each participant using Tobii Studio.

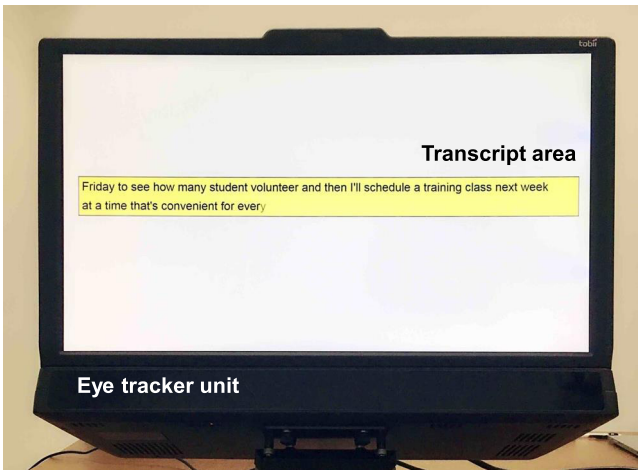


Fig. 1 Screen-based eye-tracker

The experimental procedure was as follows.

Step 1 (real-time listening). The participants listened to the audio (with/without transcripts) and pressed a button whenever they encountered a comprehension problem.

Step 2 (retrospective listening). The participants listened to the same audio again (with/without transcripts). The computer automatically stopped at the places where they pressed the button during Step 1. At this point, the participants briefly explained the type of problem they faced. This step helped them re-experience Step 1 and recall their comprehension problems. Under the with-transcript condition, their eye movements were shown on top of the ASR transcripts. The participants were asked to explain their eye movements. They were also asked if they tried to solve their problems using the ASR transcripts and if the ASR transcripts were helpful.

Step 3 (interview). The participants further explained each problem after being handed perfect transcripts of the audio clip. This step was designed to get more detailed information about the comprehension problems mentioned in Step 2. Under the with-transcript condition, they were also asked about their strategies for using the transcripts

3.2 Results

Our results are presented as follows. First, we report the types of listening comprehension problems that were generally solved by viewing the ASR transcripts. Then we describe how the NNSs used the ASR transcripts to solve problems as well as the difficulties they faced.

Table 2 Average number of problem occurrences identified by each NNS per minute under without- and with-transcript conditions

Problem	Example interview excerpt	Without-transcript		With-transcript		p-value
		Mean	SD	Mean	SD	
1.Lack of vocabulary	I didn't know this word: "archaeology." I think it's a vocabulary problem. (NNS 2)	1.529	1.106	1.640	1.158	0.567
2.Do not recognize words they know	"Tackle" I knew, but I couldn't recognize it. If I had read it, I would've understood it. (NNS 1)	1.059	0.743	0.127	0.230	**0.000***
3.Unable to form a mental representation from words heard	I knew all of the words. But when combining them, I didn't understand them. (NNS 7)	0.823	0.831	0.651	0.562	0.400
4.Cannot chunk streams of speech	I couldn't catch "Joyce in a book called Dubliners." I couldn't divide that chunk into separate words. The words linked together. (NNS 6)	0.575	0.553	0.255	0.360	**0.005***
5.Understand words but not the intended message	Even though I knew the literal meaning, I couldn't understand it in this context. (NNS 19)	0.197	0.205	0.138	0.218	0.448
6.Concentrate too hard or unable to concentrate	The whole lecture was too long. At the end, I just couldn't concentrate. (NNS 9)	0.186	0.208	0.101	0.200	**0.135**
7.Neglect the next part when thinking about meaning	I was still thinking about the meaning of "beavers," and so I missed the subsequent words. (NNS 13)	0.161	0.252	0.153	0.269	0.897
8.Confused about unexpected word appearance	They were talking about "birds." Then suddenly "mouse" came out. I got confused. (NNS 9)	0.171	0.227	0.041	0.103	**0.034***
9.Unsure about the meaning of words	"Credit" could mean academic "credit" or financial related "credit." I wasn't sure. (NNS 7)	0.155	0.296	0.110	0.157	0.498
10.Do not understand subsequent parts of input because of earlier problems	I couldn't understand the meaning of "forage". Due to that, I was unable to understand the subsequent parts. (NNS 13)	0.130	0.306	0.183	0.332	0.349
11.Confused about the key ideas in the message	The lecturer explained and explained. I could understand the literal meaning. But I was confused about the key ideas. I didn't know what she wanted to say. (NNS 1)	0.055	0.147	0.100	0.205	0.044
12.Quickly forget what is heard	When the lecturer started talking about "another critical issue," I wondered what was the previous issue? But I'd already forgotten what it was. (NNS 3)	0.026	0.081	0.032	0.079	0.745
13.Miss the beginning of texts	The audio came too abruptly, and I missed the beginning. (NNS 16)	0.022	0.067	0.000	0.000	0.163
14.Confusion caused by ASR errors	I felt what I had heard was "mainly because," but the transcripts show "maybe cuz." The error hindered my understanding. (NNS 2)	0.000	0.000	0.130	0.184	**0.005***

3.2.1 Listening Comprehension Problems Generally Solved by ASR Transcripts

RQ1 asked what types of listening comprehension problems can be solved by reading ASR transcripts. To answer this question, we first identified the listening comprehension problems faced by NNSs in each condition and investigated the types of problems that significantly decreased when ASR transcripts were provided.

To identify each type of listening comprehension problem faced by participants during the listening task, we first transcribed the interview data and classified each problem based on the problem categories suggested by two previous works [8], [11]. We used them as a base because they also focus on the listening comprehension problems of NNSs that occur during their cognitive processing of speech input. Note that we added a new category "lack of vocabulary" to the previous categories [8], [11] because it can be solved by adding a dictionary function to the ASR transcripts [12]. All the interview data were coded independently by two coders, and all discrepancies were discussed until an agreement was reached.

We counted the number of times problems occurred based on the markups (times they pressed the button). In a few cases when participants described two problems for one markup, we counted it as two. Under the without-transcript condition, 372 problem occurrences were identified; under the with-transcript condition, 267 were identified, including ten problems caused by ASR errors.

Table 2 provides an overview of the problems faced by NNSs under without- and with-transcript conditions. To compare how the ASR transcripts changed the distribution of the problem occurrences, we first counted the problem occurrences of each participant. Next, we conducted a paired t-test (two-tailed) to see whether the average number of problem occurrences per minute changed between the two conditions. Results showed that the NNSs faced significantly

Table 3 Factors that hindered NNSs from solving their problems

Factor	Example interview excerpt
ASR errors (61.3%)	I couldn't understand, so I checked the transcripts. After seeing errors in them, I became even more confused. (NNS 1)
Lack of time (25.8%)	The sentence (I had a problem with) was a bit too long. Although I checked the transcripts, I didn't have enough time to think. (NNS 10)
No punctuation (6.5%)	There was no period between "yet" and "the" in the transcripts. I thought they belonged to one sentence, but actually they belonged to two sentences, so I didn't understand. (NNS 4)
Others (6.5%)	

fewer problems in the with-transcript conditions for three types of problems: "do not recognize words they know" ($p < .01$), "cannot chunk streams of speech" ($p < .01$), and "confused about unexpected word appearance" ($p < .05$).

One common element to these three problems is that they occur in the early stage of speech comprehension. In other words, they all occur during the cognitive processing phases of perception in language comprehension, which deals with the encoding of acoustic messages [13]. ASR transcripts benefit NNSs during such perceptual processing by transforming acoustic information into textual information.

Although most of the three types of problems were solved by showing the ASR transcripts, in some cases they weren't. To identify why, we analyzed the explanations of the NNSs to the interview question, "why didn't the ASR transcripts help you solve your problem?" and attributed three main factors that hindered the NNSs from solving them: i) ASR transcript errors, ii) lack of time to identify the relevant parts of the transcripts or to consider the meaning of the transcripts, and iii) confusion caused by no punctuation of the transcripts.

Table 3 summarizes the three factors and shows some excerpts from the interviews. Since these factors hindered our participants from solving their problems, removing influence of them would improve NNS comprehension.

3.2.2 How NNSs Use ASR Transcripts

RQ2 asked how the NNSs used the ASR transcripts and what were the difficulties with them. To answer this issue, we analyzed the post-task interviews and the gaze movement data of our participants. We found that they adopted different strategies when using the ASR transcripts and identified that the main difficulties they faced were related to their strategies.

Based on post-task interviews and eye movement data, we identified two strategies for using ASR transcripts; some participants generally followed the transcripts throughout the listening while others only looked at them when needed.

For those who followed the transcripts throughout the listening, the ASR transcripts seemed to increase their confidence in what they were hearing. For example, one NNS mentioned:

While listening, I read the transcripts to check if what I heard was correct. I felt relieved. (NNS 7)

Some participants only checked the transcripts when necessary. For example, they checked them when they encountered a problem or wanted to confirm what they had heard. One participant explained why he adopted such a strategy:

I felt the transcripts were a little distracting. So I focused on listening. If I encountered something I didn't understand, I read the transcripts. After reading, I went back to the listening mode. (NNS 5)

Regardless of NNSs' attempts and efforts to use the ASR transcripts, NNSs often faced difficulties exploiting them. Post-task interviews identified two main reasons that NNSs were unable to fully use the ASR transcripts.

Lack of time and cognitive resources to simultaneously handle multimodal contents. Our participants reported the difficulty of simultaneously processing both speech and textual information, especially when there was delay in the transcript appearance. For example, participants who simultaneously dealt with two modalities reported the following:

I wanted to listen and I also wanted to read. I felt dizzy. (NNS 1)

I think the transcripts are useful, but they require too much effort. (NNS 10)

Difficulty shifting between multimodal contents. For the NNSs who only followed the transcripts when they faced comprehension problems, it took them time to search through the transcripts to solve their problems. Some participants mentioned this in their interviews.

Sometimes I didn't know where the word I had a problem with was on the screen. I needed to search for it, and that was time-consuming. (NNS 5)

4. Study 2

In study 1, we found that NNSs adopted different strategies when using ASR transcripts. Some participants generally followed the transcripts, while others only looked at them when needed. We wondered if NNSs would change their strategies when transcripts were displayed using different methods. Therefore, we pose the following research question:

RQ3: Do the display methods of transcripts affect NNSs' strategies of using transcripts? If yes, how?

We were also interested in finding out if there is a relationship between how NNSs use the transcripts and their preferred display method. Thus, we ask the following research question:

RQ4: Is there a relationship between NNSs' strategies of using transcripts and their preferences for the method used to display the transcripts?

4.1 Method

To investigate the above research questions, we conducted a laboratory experiment with 22 non-native English speakers: 12 Chinese and 10 Japanese. In the experiment, the participants engaged in two listening tasks with ASR transcripts presented in two display conditions:

Condition 1: accuracy-oriented display which shows transcripts only after the completion of speech input analysis.

Condition 2: speed-oriented display which shows interim analysis results of speech input.

The experiment used a within-subject design. Conditions were counterbalanced across subjects to minimize order effects. Two lectures from the TOEFL exam were chosen as task materials, which varied from four to five minutes. On average, each clip contained about 670 words. Real-time transcripts of each audio clip were generated by the Google speech recognition API. The WER of the ASR transcripts was about 10% on average. 65% of the errors were substitutions, 27% were deletions, and 8% were insertions. Our system showed the original text data from the Google API in condition 2. For the transcripts in condition 1, we tweaked the original google transcripts and masked most of the interim results.

As in study 1, the participants sat in front of the Tobii TX300 eye-tracker and completed a 9-point calibration for it. After the calibration, they engaged in two listening tasks, each under a different condition. Each task was divided into three steps.

Step 1 (real-time listening). The participants listened to the audio clip. One type of ASR transcripts was provided.

Step 2 (retrospective listening). The participants listened to the same audio clip again. Their eye movements in Step 1 were shown at the top of the ASR transcripts. Participants were asked to explain their own eye movements.

Step 3 (Survey). The participants answered a survey about their experiences and strategies of using the transcripts.

After the two listening tasks, they were further asked questions such as "which transcript display method do you prefer, why", "what was your strategy of using the transcripts", "did you change your strategy across the conditions"?

4.2 Results

4.2.1 How Display Methods Affect NNSs' Strategies of Using ASR Transcripts

RQ3 asked if the display methods of transcripts affected the NNSs' strategies of using transcripts. To answer this question, we analyzed the survey and interview data of our participants.

In the survey, NNSs were asked to rate the amount of attention they paid to listening to the audio and reading the transcripts (in total 100%). We calculated the difference in attention allocation for the two display methods. For example, if a NNS allocated 50% of his/her attention to listening when speed-oriented transcripts were presented, and 60% when accuracy-oriented transcripts were presented, the difference would be 10%.

Table 4 Difference in attention allocation between two display methods

Difference in attention allocation	Number of NNSs (Percentage)
diff. <=10%	14 (64%)
10< diff. <=20%	5 (23%)
20%< diff.	3 (14%)

We found that for 64% of the participants, the difference never exceeded 10%; For 87% of the participants, the difference was equal to or less than 20% (See Table 4).

The results suggest that the display method does not significantly determine NNS strategy of using transcripts. It is consistent with the NNS interview responses. For example, one participant stated:

My strategy for using transcripts (under the two display methods) are the same. I listened to the audio first, and used the transcripts to confirm my understanding. It's just that the transcripts I disliked created a bigger burden. (NNS7)

4.2.2 NNSs' Strategies of Using ASR Transcripts and Their Preferences

RQ4 asked whether a relationship existed between the NNSs' strategy of using the transcripts and their preferred transcript display method. Among the 22 NNSs, 12 preferred the accuracy-oriented display method, eight preferred the speed-oriented display method, and two had no preference.

A logistic regression was performed to assess the relationships between the amount of attention they paid to listening to the audio and their preferred display method (0 for speed-oriented display and 1 for accuracy-oriented display).

Results show that the more the NNSs paid attention to the audio, the more they tended to prefer speed-oriented transcripts. A significant association was found both when the accuracy-oriented transcripts were presented ($P < 0.05$, odds ratio (OR) = 0.94, 95% confidence interval (CI) 0.88–1.00) and when the speed-oriented transcripts were presented ($P < 0.05$, odds ratio (OR) = 0.94, 95% confidence interval (CI) 0.89–1.00).

NNSs who preferred accuracy-oriented transcripts

Some NNSs listened to the audio and read the transcripts simultaneously. Those participants normally preferred the accuracy-oriented transcripts because they were easier to read.

The speed-oriented display method was indeed faster, but it contained too many errors. They upset my comprehension. Maybe the errors were corrected later on, but I didn't have time to look back. I need to follow the ongoing speech. (NNS11)

The "flashing" was quite confusing. It made me won-

der a lot of things, like "what word was changed?", "why did it change to this word?" Then I lost track of listening. (NNS4)

Compared to the flashing caused by continuous word replacements, they felt the delay was not a big problem. Some mentioned that as long as the delay is within a certain range, there is not much difference. Below are two NNS statements.

What I am concerned most about are the errors. As for delay, it would be fine if it's within a certain range, for example, no more than 2 or 3 seconds. I can feel the transcripts came faster under real-time condition, but it has no effect. (NNS2)

Both conditions contained some delay. I didn't feel much difference in the delays between the two conditions. I mentioned that delay would exacerbate distractions. But for me, whether there is delay matters, not the amount of delay. (NNS1)

NNSs who preferred speed-oriented transcripts

In the interview, we asked our participants to explain the reasons for their preference. Some NNSs listened to the audio and read the transcripts only when necessary (e.g., when they encountered a comprehension problem). Such NNSs tended to prefer the speed-oriented transcripts. This was because that when understanding difficulties arose, they wanted to check the transcripts quickly to resolve the issue and then go back to listening. Delay in transcript display interfered with their listening. Some also reported that waiting for the transcripts made them anxious.

If I could understand what was being said, then I wouldn't read the transcripts. I took a look at some words or a short sentence only when I did not hear clearly. Therefore, I like the real-time one because I don't need to wait. (NNS3)

Waiting for the delayed transcripts made me anxious. One sentence had already appeared in my mind, but it didn't show on the screen. I felt uncomfortable. (NNS20)

Compared to delay, they thought the errors and flashing caused by continuous word replacements didn't affect them much.

Sometimes the transcripts were wrong and I could spot where it was wrong, so it wouldn't affect me that much. (NNS3)

However, two NNSs preferred the accuracy-oriented transcripts even though they mainly focused on listening and read the transcripts only occasionally. According to the participants, the delay didn't matter to them because they didn't check the latest transcripts.

Probably it's because of my strategy of using the transcripts. The delay doesn't matter to me. Basically, I focused on listening, and when I felt I got something wrong, I would go and check the transcripts. I didn't check the latest transcripts. Most of the time, I used the transcripts to check the previous sentence. (NNS5)

We also found two NNSs thought either type of method was fine. They were not sensitive to word replacements or delay and could accept both display methods well.

5. Design Implications

Our findings from the two studies suggest several ways of enhancing ASR transcripts to better facilitate NNS comprehension.

5.1 Designing More Effective ASR Transcripts

ASR errors not only hindered NNSs from solving their problems but also increased their confusion. We suggest exploiting the word recognition confidence scores when presenting ASR transcripts [15]. The lower the recognition confidence score, the greater is the likelihood of ASR error. Therefore, when presenting ASR transcripts, we could deemphasize words with low confidence scores (e.g., shown in gray) and emphasize words with high confidence scores (shown in bold).

We found that many NNSs had difficulties simultaneously handling multiple contents, which placed an additional burden on them when utilizing the transcripts. Previous studies found that some NNSs benefit more when only keywords are presented as captions rather than entire sentences [16]. This strategy may also be beneficial when presenting ASR transcripts to NNSs because the keywords could help them quickly understand the key points of the conversations/lectures without attracting excessive attention. Another possible way to reduce NNS workload is to show transcripts only when necessary, for example, showing a line of transcripts only when a NNS presses the button.

In addition, some NNSs reported that they had to search through the transcripts to spot the relevant place when they faced some problems. We suggest helping NNSs locate where they had problems in the transcripts. For example, when a NNS encounters a problem and presses the button, the system could automatically mark that place on the transcripts.

5.2 Adaptive ASR Transcript Displays to Each NNS

In our study, we found that NNSs had different preferences regarding the display method of ASR transcripts. Mismatched transcript display had negative effects on their listening comprehension. We analyzed the post-task interviews and the gaze movement data of our participants and found that NNSs' eye movements could provide some information about their display method preferences.

Some NNSs were quite sensitive to errors or flashing caused by word replacements in the speed-oriented display. As noted earlier, these NNSs tended to prefer accuracy-oriented transcripts. Figure 2 shows the gaze plots of one such NNS. As shown in the figure, when speed-oriented transcripts were presented, the eye gaze lagged slightly behind the front-end of the transcripts. This is because he tried to avoid reading the transcripts that were still being corrected. When accuracy-oriented transcripts were presented, he could follow the leading-end of the transcripts.

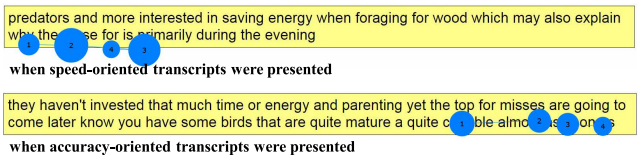


Fig. 2 Gaze plot of a NNS who preferred accuracy-oriented transcripts

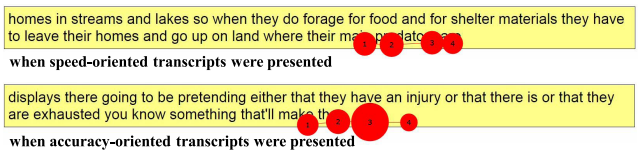


Fig. 3 Gaze plot of a NNS who preferred speed-oriented transcripts

Some NNSs were quite sensitive to delay. As noted earlier, these NNSs tended to prefer speed-oriented transcripts. Figure 3 shows the gaze plots of one such NNS. As shown in the figure, when accuracy-oriented transcripts were presented, his eye gaze sometimes preceded the transcripts, indicating that he was waiting for the transcripts to arrive. On the other hand, when speed-oriented transcripts were presented, he could follow the latest transcripts to confirm his listening.

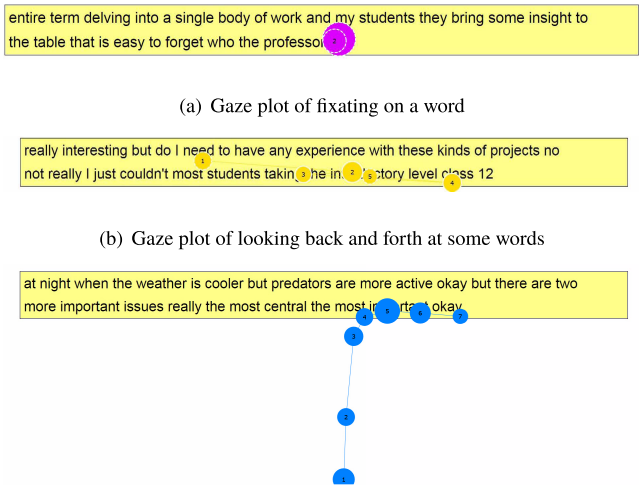
Based on the findings, we suggest providing adaptive ASR transcript displays for different NNSs to better support their listening comprehension. If the system detects via eye-tracking that the NNS generally follows transcripts, it would be better to provide more accurate transcripts to try to avoid the confusion caused by word replacements. On the other hand, if the NNS only looks at the transcripts occasionally, it may be better to provide transcripts quickly to reduce the wait time.

In addition, if the system detects certain NNS eye movements such as "waiting for the transcripts to come", providing speed-oriented transcripts may be helpful. On the other hand, if certain eye movements such as "avoids reading the interim transcripts" are detected, providing accuracy-oriented transcripts may benefit the NNSs.

5.3 Introducing Other Technologies to Supplement ASR Transcripts

For problems that were difficult or impossible to solve by ASR transcripts, we suggest introducing other technologies to supplement ASR transcripts [17]. For example, the "lack of vocabulary" problem greatly hampered the NNSs. For such problems, based on ASR transcripts where NNSs pressed a button, automatically providing dictionary definitions and images might be helpful.

Previous work suggested that eye tracking is not only useful for analyzing user behavior, but it can also be used as an input mechanism and a means of interacting with a program, a game, or some other technology [18], [19]. In our study, we observed some typical eye movements of NNSs when encountering comprehension problems: 1) fixating on a word or phrase (Fig. 4 (a)); 2) looking back and forth at words or phrases (Fig. 4 (b)); 3) shifting from no-transcript to transcript areas (Fig. 4 (c)).



(a) Gaze plot of fixating on a word

(b) Gaze plot of looking back and forth at some words

(c) Gaze plot of shifting from no-transcript to transcript area

Fig. 4 Typical eye movements of NNSs when encountering comprehension problems

These gaze patterns could be useful for detecting the types of problems experienced by NNSs. If a system could detect them in real time, it may provide a suitable support for NNSs to solve the problems without extra burdens. For example, if a NNS's gaze is fixated on a word, the system could automatically provide a translation or an image of it to support comprehension.

6. Limitations and Future Directions

There are several limitations to our studies. We identified certain gaze patterns of NNSs when they used ASR transcripts with different display methods and when they encountered comprehension problems. However, the findings were based on post-task interviews of NNSs. In future work, we need to quantitatively investigate them.

We reported that ASR errors hindered NNSs from solving their problems and increased their confusion. However, we didn't go into details about how their listening comprehension are affected by ASR errors. In future work, we want to explore 1) what types of ASR errors are critical or negligible to NNSs and 2) how the gaze patterns of NNSs might be affected by ASR errors.

7. Conclusions

In this paper, we reported two laboratory studies. The first investigated the impact of ASR transcripts on the listening comprehension of NNSs. We found that the ASR transcripts helped the NNSs solve certain problems (e.g., "do not recognize words they know"), but imperfect ASR transcripts (e.g., errors and no punctuation) hindered NNSs from solving their problems. In addition, post-task interviews revealed that the NNSs did not have enough time to fully exploit the

transcripts. For example, they had difficulty concentrating on listening/reading or shifting between the two. We also found that NNSs adopted different strategies when using the ASR transcripts. Based on the first study, the second study explored how different display methods affected NNSs' listening experience and their use of transcripts. Analysis results showed that the more the NNSs paid attention to listening to the audio, the more they tended to prefer speed-oriented transcripts, and vice versa. Mismatched transcripts were found to have negative effects on NNSs' listening comprehension. Our findings have implications for improving the presentation methods of ASR transcripts to more effectively support NNSs.

Acknowledgments

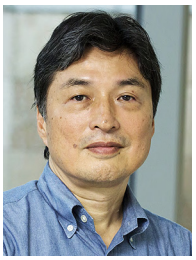
This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020) from Japan Society for the Promotion of Science (JSPS).

References

[1] X. Cao, N. Yamashita, and T. Ishida, "Investigating the impact of automated transcripts on non-native speakers' listening comprehension," Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp.121–128, ACM, 2016.
[2] A. Bloomfield, S.C. Wayland, E. Rhoades, A. Blodgett, J. Linck, and S. Ross, "What makes listening difficult? factors affecting second language listening comprehension," tech. rep., DTIC Document, 2010.
[3] E. Hinkel, Handbook of research in second language teaching and learning, Routledge, 2011.
[4] Y. Pan, D. Jiang, L. Yao, M. Picheny, and Y. Qin, "Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.1725–1734, ACM, 2010.
[5] L. Yao, Y.-X. Pan, and D.-N. Jiang, "Effects of automated transcription delay on non-native speakers' comprehension in real-time computer-mediated communication," IFIP Conference on Human-Computer Interaction, vol.6946, pp.207–214, Springer, 2011.
[6] G. Gao, N. Yamashita, A.M. Hautasaari, A. Echenique, and S.R. Fussell, "Effects of public vs. private automated transcripts on multiparty communication between native and non-native english speakers," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.843–852, ACM, 2014.
[7] J. Rubin, "A review of second language listening comprehension research," The modern language journal, vol.78, no.2, pp.199–221, 1994.
[8] C.C.M. Goh, "A cognitive perspective on language learners' listening comprehension problems," System, vol.28, no.1, pp.55–75, 2000.
[9] A. Hautasaari and N. Yamashita, "Do automated transcripts help non-native speakers catch up on missed conversation in audio conferences?," Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology, pp.65–72, ACM, 2014.
[10] V. Kalnikaitė, P. Ehlen, and S. Whittaker, "Markup as you talk: establishing effective memory cues while still contributing to a meeting," Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp.349–358, ACM, 2012.
[11] X. Cao, N. Yamashita, and T. Ishida, "How non-native speakers perceive listening comprehension problems: Implications for adaptive support technologies," International Conference on Collaboration Technologies, vol.647, pp.89–104, Springer, 2016.
[12] G. Gao, N. Yamashita, A.M.J. Hautasaari, and S.R. Fussell, "Improving multilingual collaboration by displaying how non-native speakers use automated transcripts and bilingual dictionaries," Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp.3463–3472, ACM, 2015.
[13] J.R. Anderson, Cognitive psychology and its implications., WH Freeman/Times Books/Henry Holt & Co, 1990.
[14] P. Winke, S. Gass, and T. Sydorenko, "Factors influencing the use of captions by foreign language learners: An eye-tracking study," The Modern Language Journal, vol.97, no.1, pp.254–275, 2013.
[15] J.C. Lai and J.G. Vergo, "Speech recognition confidence level display," Dec. 21 1999. US Patent 6,006, 183.
[16] H.G. Guillory, "The effects of keyword captions to authentic french video on learner comprehension," Calico Journal, vol.15, no.1-3, pp.89–108, 1998.
[17] T. Ishida, The language grid: Service-oriented collective intelligence for language resource interoperability, Springer Science & Business Media, 2011.
[18] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan, "Eye tracking and online search: Lessons learned and challenges ahead," Journal of the American Society for Information Science and Technology, vol.59, no.7, pp.1041–1052, 2008.
[19] I. Umata, S. Yamamoto, K. Ijuin, and M. Nishida, "Effects of language proficiency on eye-gaze in second language conversations: toward supporting second language collaboration," Proceedings of the 15th ACM on International conference on multimodal interaction, pp.413–420, ACM, 2013.

Xun Cao received her PhD in Informatics from Kyoto University in 2017. Her research interests include multilingual communication, natural language interfaces and intelligent systems.

Naomi Yamashita is a primary researcher of NTT Communication Science Laboratories. She received B.Eng. and M.Eng. degrees in applied mathematics and physics and a Ph.D. degree in Informatics from Kyoto University in 1999, 2001 and 2006, respectively. Her primary interests lie in the areas of computer-supported cooperative work and computer mediated communication. Her current projects include technological support for multilingual communication, studies on mental healthcare, and the development of video systems that facilitates group-to-group distant collaboration.



Toru Ishida has been a professor of Kyoto University since 1993. His academic background includes visiting scientist/professor positions at Columbia University, Technische Universitaet Muenchen, Université Pierre et Marie Curie, University of Maryland, Shanghai Jiao Tong University, Tsinghua University, Xinjiang University and Hong Kong Baptist University. He is a fellow of IEICE, IPSJ, and IEEE. He is a co-founder of the Department of Social Informatics, Kyoto University, and recently organized the Kyoto University Design School. His research interest lies with Autonomous Agents and Multi-Agent Systems and modeling collaboration within human societies. He contributed to create AAMAS/ICMAS/PRIMA conferences on Autonomous Agents and Multi-Agent Systems. His projects include Community Computing, Digital City Kyoto, Intercultural Collaboration Experiments, and the Language Grid.