Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

‘Um, so like, is this how I speak?’: design implications for automated visual feedback systems on speech

Soonho Kwon, Hyojin Lee, Soojeong Park, Yerin Heo, Keeheon Lee & Younah Kang

To cite this article: Soonho Kwon, Hyojin Lee, Soojeong Park, Yerin Heo, Keeheon Lee & Younah Kang (2024) ‘Um, so like, is this how I speak?’: design implications for automated visual feedback systems on speech, Behaviour & Information Technology, 43:13, 3091-3110, DOI: 10.1080/0144929X.2023.2271997

To link to this article: https://doi.org/10.1080/0144929X.2023.2271997

Published online: 09 Nov 2023.

BEHAVIOUR & INFORMATION TECHNOLOGY
2024, VOL. 43, NO. 13, 3091–3110
https://doi.org/10.1080/0144929X.2023.2271997

'Um, so like, is this how I speak?': design implications for automated visual feedback systems on speech

Soonho Kwon, Hyojin Lee, Soojeong Park, Yerin Heo, Keeheon Lee and Younah Kang

Underwood International College, Yonsei University, Seoul, Republic of Korea

ABSTRACT

At a time when diverse forms of automated visual feedback systems on speech are introduced to the Human–Computer Interaction (HCI) community, our research investigates the possible design spaces and design implications to consider when creating them. We utilised a Wizard of Oz prototype that recognises fillers during speech and sends out real-time feedback in the form of a flickering screen and a post-report with graphic charts. We let 27 participants use this prototype and interviewed them about their general experience of using such systems. Based on thematic analysis, we propose eight design implications to consider when designing an automated visual feedback system on speech. Additionally, we discuss the possible future for automated visual feedback systems on speech where they collaborate with fields such as personal informatics (quantified self), self-determination theory (SDT) and motivation, and other socio-ethnic areas

ARTICLE HISTORY
Received 23 February 2023
Accepted 11 October 2023

KEYWORDS
Human-computer interaction; speech feedback system; feedback systems; self-tracking; informatics

1. Introduction

Fluent public speaking skills often lead to successful public relations and job success (Mancuso and Miltenberger 2016; Parvis 2001). There have been multiple attempts to aid one's public speaking training, from reducing anxiety (Fremouw and Breitenstein 1990) to habit reversal experiments focusing on reducing filled pauses through awareness training (Mancuso and Miltenberger 2016; Pawlik and Perrin 2020). Recognising such importance, researchers in the Human–Computer Interaction (HCI) community have explored, presented, and evaluated novel forms of speech training aids such as virtual interactive agents (Chollet et al. 2015; 2022; Hoque et al. 2013), robotic speech feedback (Trinh et al. 2017), or rehearsal support systems (Trinh, Yatani, and Edge 2014). Of them, the popular research topic is automated feedback systems on speech (Chollet et al. 2015; 2022; Kurihara et al. 2007; Schneider et al. 2015; Tanveer, Lin, and Hoque 2015; Zhao et al. 2017). The majority of these systems detect and analyse the verbal, non-verbal, and paralingual behaviour of users through computerised systems and provide various forms of feedback, such as coloured gauges, texts, visual cues, or haptic reminders. Most of the existing literature explores specific technical feasibility, and the effects of the systems on the users; however, to our knowledge, there is yet to be a study that focuses on more generalisable design considerations that could be consulted upon when creating these services. Thus, the goal of our research is to draw more general and operational design considerations for an automated visual feedback system on speech, based on empirical user research. As a design research project, we expand upon existing research efforts to explore various factors that affect the users' experience when using a computer-based automated visual feedback system on speech.

We conducted an interview study in which participants were provided with a simple feedback system on filled pauses consisting of real-time reminders and post-feedback for presentation preparation. We chose filled pauses as our target behaviour, as it was one of the most popular target behaviours for multiple feedback systems (Bubel et al. 2016; Kurihara et al. 2007; Schneider et al. 2015; Zhao et al. 2017), and some even considered the occurrence of filler as equivalent to the quality of the speech (Chollet et al. 2015). As we found it important for our system to represent the current forms of feedback systems, we decided to provide two of the most popular forms of feedback: real-time reminders and post-feedback. To see if this was a valid model that induces behavioural change, we measured the trend of the filler occurrence according to the given feedback forms. In our main study, we let users experience this system through a session in which users were asked to give a spontaneous presentation on a given topic. For each session, semi-

CONTACT Younah Kang kang.younah@gmail.com

structured interviews were conducted before and after the presentation to deepen our understanding of the users' perceptions and reactions to the system and their attitude toward unwanted speech behaviours. Additionally, through thematic analysis (Nowell et al. 2017) and a comparative analysis of existing systems, we drew a series of design implications for an automated visual feedback system on speech.

The contributions of our research are as follows:

- empirical research on the users' motive and experience in using automated visual feedback systems on speech,
- explore possible design spaces and design implications for automated visual feedback systems on speech, and
- discuss the future possibilities for research in automated feedback systems on speech and its relation to other areas.

2. Background

2.1. Improving public speaking skills

Public speaking refers to situations where an individual directly speaks to a live audience in a structured, deliberate manner to inform, influence, or entertain them (Seligman and Darley 1976). Good speech consists of various factors including fluency, using appropriate and variety of expressions, good organisation, confidence, etc. (Babaii, Taghaddomi, and Pashmforoosh 2016).

Recognising its importance, much research has touched upon the topic of improving one's speech skills. Among them, the 'habit reversal experiment' closely resembles the automated visual feedback system on speech. A habit reversal experiment (specifically, an awareness training technique) aims to eliminate nervous habits and ticks (Azrin and Nunn 1973; Mancuso and Miltenberger 2016; Pawlik and Perrin 2020). In these studies, the researchers asked the participants to identify the target behaviour (filled pauses) while watching recordings of their speech. In later sessions, they instructed the participants to raise their hands whenever they displayed one of the target behaviours. This process resulted in a successful reduction of filled pauses, and the participants found it efficient and effective. In the first sessions, participants were able to recognise their problems, which made them more mindful during the following sessions and self-recognise whenever they displayed such behaviours. The takeaway of the research is that letting users recognise their undesirable behaviour during speech and retaining that awareness was key to decreasing such habits, ultimately improving their speech performance.

2.2. Automated feedback on speech

Multiple computer-based systems were introduced to improve speech or presentation skills. Among these attempts, some systems incorporate virtual agents to interact with users, such as virtual audiences incarnated through virtual reality (VR) (Chollet et al. 2015; Hoque et al. 2013); these are not within the scope of our research as they are more interactive agents rather than direct feedback systems. The ones we do want to focus on are systems that directly let users recognise their wrongdoings during speech. The existing literature provides us with three possible design spaces to consider: target behaviour, the timing of feedback, and the form of feedback.

First, the system must have a clear target to monitor. Existing systems all focus on identifying certain undesirable behaviours, letting users know of such behaviours, and trying to retain that awareness. These behaviours include both verbal behaviours such as filler rate (Bubel et al. 2016; Kurihara et al. 2007; Schneider et al. 2015; Zhao et al. 2017), and non-verbal or paralingual behaviours including speed (Kurihara et al. 2007), voice intensity (Schneider et al. 2015), voice pitch (Bubel et al. 2016), eye contact (Kurihara et al. 2007), and body language (Kurihara et al. 2007; Schneider et al. 2015; Tanveer, Lin, and Hoque 2015).

Based on when the feedback is given, the systems can also be categorised into the following: real-time reminders, post feedback, and a mixture of both. Real-time reminders provide feedback simultaneously as users give their presentations. The examples include Presentation Trainer (Schneider et al. 2015), where they provide written feedback on a mirrored image of the presenter; Rhema, which used Google Glass to provide coloured gauges and text feedback on their verbal and non-verbal behaviours (Tanveer, Lin, and Hoque 2015); and the 'direct feedback (DF)' model by Chollet and colleagues where they also provided colour-coded gauges (Chollet et al. 2015). The second type, post-feedback, provides a summary of the performance based on the analysis of the user's speech. These include ROC speak where users can upload their speech video to receive analysis on filled pause usage, etc. (Zhao et al. 2017), and RoboCOP, where colloquial feedback is given by a robot after the presentation (Trinh et al. 2017). Several systems offer a combination of these features. For instance, Presentation Sensei provides users with real-time icon-based feedback during speech and also offers post-analysis of the speech (Kurihara et al.

2007). Another example is the public speaking training framework Cicero which offers real-time interactive feedback from a virtual audience and generates an after-action report specifically designed for public speaking training applications (Chollet, Marsella, and Scherer 2022).

Finally, feedback can take on many different forms. The form could be as simple as a colour-based gauge (Chollet et al. 2015; Tanveer, Lin, and Hoque 2015), icons, graphs (Kurihara et al. 2007), or simple texts (Schneider et al. 2015; Tanveer, Lin, and Hoque 2015), and could be as complex as colloquial feedback (Trinh et al. 2017). The researchers also emphasise the importance of exploring beyond visual forms to other innovative and interesting feedback, such as haptic ones (Schneider et al. 2015).

These identified design spaces were consulted upon to (1) design the system we utilised during our interview sessions, and (2) enrich the design implication we drew based on our interview analysis.

2.3. Filled pauses

Filled pauses, or fillers, are meaningless sounds that take place during speech (the examples include 'uh,' 'um,' or excessive usage of meaningless 'like'). Among the various factors that hinder fluency, habitual usage of fillers is reported as a distraction to the listener, hindering the quality of speech (Bell 2011; Clark and Fox Tree 2002; Henderson and Henderson 2007; Mancuso and Miltenberger 2016; Miltenberger, Fuqua, and Woods 1998; Pawlik and Perrin 2020). Other researchers have also argued that the overuse of speech disfluencies, such as filled pauses, leads to a negative social image and distracts listeners (Bell 2011; Clark and Fox Tree 2002; Henderson and Henderson 2007; Pawlik and Perrin 2020). For such reasons, filled pauses are one of the most popular targets to monitor in automated feedback systems on speech (Bubel et al. 2016; Kurihara et al. 2007; Schneider et al. 2015; Zhao et al. 2017). Chollet and colleagues even considered the occurrence of fillers as equivalent to the quality of speech itself and used them to determine how good the speech was (Chollet et al. 2015).

While filled pauses seem to be a universal behaviour, the term referring to them varied according to culture. For instance, the term 'filled pauses (휴지)' is scarcely used in Korean society and academia. Interestingly, however, some Korean linguistic studies consider discourse markers equivalent to filled pauses and have researched the reduction of discourse markers during a speech (J. Lee and Jeong 2018). Jeon identified the discourse markers in the Korean language, which consist of commonly used filled pauses such as 'uh' or 'um.' Thus, our research considered discourse markers equivalent to filled pauses in designing a feedback system on speech for the Korean language (Jeon 2017).

3. Methods

3.1. Speech feedback system design

In this section, we define our target situation, behaviour, feedback timing, and feedback format. Additional elaboration on the feasibility and execution of the system is provided, as well.

3.1.1. Target situation and target behaviour

Among a variety of public speaking situations, our system was designed for short oral presentations on a given topic. In this situation, users would have to instantly compose their responses, rather than have a chance to perfectly polish their presentation. By doing so, we expected them to exhibit more target behaviours, allowing them more chances to engage with the system and ultimately leading them to share more fruitful thoughts and insights. Specifically, we wanted to create an online presentation situation. This was not only to ensure a safer environment for the participants during the pandemic situation but also to represent the current trend where more and more public speaking and speech takes place online; even before the pandemic, more than 6 out of 10 interviews were held online (OfficeTeam n.d.).

A list of filled pause words had to be made to execute the system. We consulted the aforementioned list of Jeon (Jeon 2017), who had organised all academically identifiable Korean discourse markers from 1993 to 2016, to create a list of target filled pauses. Table 1 displays the list of words we decided to monitor, which have been reorganised from the original table of Jeon for ease of explanation. As most of these words are unique to the Korean language, they were not translated into English but were provided with brief explanations and similar English discourse markers and filled pauses.

3.1.2. Feedback design and execution

The system incorporated both real-time and post-feedback. There are two reasons for making this choice. First, as our study focuses on excavating users' general experience of using an automated visual feedback system on speech, it was important to let them experience both types of feedback.

Second, while there are a few studies that provided feedback both during and after a speech (Chollet, Marsella, and Scherer 2022; Kurihara et al. 2007),

Table 1. Target filler words.

Category	Discourse Marker	Explanation
Interjection	글쎄, 아, 어, 오. 예/네, 자, 저기(요)	Words that express emotion or surprise, similar to oh, ah, uh, etc.
Interjection / adverb	가만, 그래, 그러게, 아니, 왜, 참	Interjections that are technically adverbs, similar to really, like, etc.
Interjection / pronoun	거시기, 뭐, 어디	Interjections that are technically pronouns, similar to there, where, etc.
Adverb	그러-, 그래서, 그러니가, 그런데, 그리고, 그냥, 그만/고마, 그저, 다, 막, 아무튼/어쨌든/하여튼, 이제/인자, 정말/진짜, 좀	Adverbs that qualify phrases, similar to and, anyway, now, etc.
Pronoun / determiner	이, 그, 저	Pronouns used before a noun word to direct the noun word, similar to that, this, etc.
Determiner	무슨	A determiner that comes before a noun word to specify the noun word, similar to what, etc.
Verb / Adjective / Phrase	됐어, 말이야/말이다, 물론이다, 당연하다, 뭐냐/뭐랄까/뭐야, 있지/있잖아, 그래 가지고, 다름 아니라, 그건 그렇고, 아 근데, 아니 근데, – 어 가지고	Verbs, adjectives, and phrases, similar to you know, I mean, besides, etc.

most of these studies dealt primarily with quantitative data. For example, Chollet and colleagues conducted a similar approach to our study, but its result heavily relied on various questionnaires, such as Public Report of Confidence as a Speaker (PRCS), Big 5 Personality Inventory, Immersive Tendencies Questionnaires (ITQ), and others. The study collected quantitative results presented by a Likert scale. This highlights the need for further investigation at a qualitative level, specifically regarding how feedback given at different points in time could complement each other.

Determining the specific form of feedback was a challenge. It seemed plausible that we provide every possible type of feedback to be representative but this would cause several problems. First, providing every possible form of feedback (coloured gauges, textual guidelines, haptic alerts, colloquial feedback, etc.) to the participants all at once would be too overwhelming, leading to less engaging participants with minimal findings. Second, unlike the timing of the feedback, the type of feedback is yet to be exhaustive. Many researchers are still exploring novel forms of feedback; even if we do incorporate every type of feedback that exists until now and draw implications, it would not cover the newly created ones after our research.

We chose coloured notifications for real-time feedback. As our system simply aimed to let users know that they exhibited target behaviour, we decided that making the feedback complex would only distract users from speech. We also decided not to use the gauge interface (Chollet et al. 2015; Tanveer, Lin, and Hoque 2015) as the previous research suggested that it required constant awareness to understand the meaning, which users found obtrusive; rather, we used a flickering notification that whenever the participants exhibited one of the filled pauses in our database, the screen turned red for 0.5 s and returned to its original black state (see Figure 1) to warn users to be mindful. In the only-oral online presentation situation that we tried to create, this would be seen as the screen being 'turned off' (black screen) and only being turned on whenever the participants exhibited target behaviour.

For the post-feedback, we chose the graph chart format. Two distinctive formats stood out for the post feedback: graphs (Kurihara et al. 2007; Zhao et al. 2017) and colloquial (Trinh et al. 2017). We found prior a much more computerised version and later more of a humanoid robot giving human-like feedback. While the latter is an impressive model, we found graphs to be a more representational version for the automated visual feedback system on speech. The post-feedback (or reports) counts the number of filled pauses used during a single session of the interview and shows them in the form of bar graphs. The post-feedback on a session was

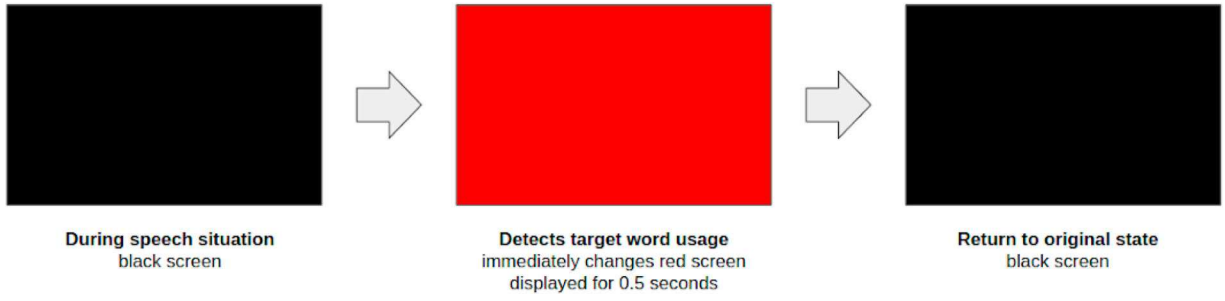

Figure 1. How the screen changes when the user exhibits one of the target filled pauses. The participant's screen would stay black and whenever the researchers manually sent the feedback, the screen would light up for 0.5 s and return to its original state.

provided during the pre-interview of the next session to be discussed during the interview. An example chart is presented in Figure 2.

The execution of the immediate feedback was implemented through Wizard of Oz prototyping. Wizard of Oz prototyping refers to a user testing method in which researchers use a human-operated device that is perceived as a working computer-based prototype by the participants (Bernsen, Dybkjær, and Dybkjær 1993; Maulsby, Greenberg, and Mander 1993). This research method is often used for the user testing of software, which requires a large budget and time to develop (Bernsen, Dybkjær, and Dybkjær 1993). While there exist many systems that recognise filled pauses, there is yet to be one for the Korean language. To create one, it requires several hundred different voice data of the spoken key phrase to properly operate. As gathering such data for more than 40 different words would delay the research process and increase our budget, we decided to utilise the Wizard of Oz prototyping to implement immediate feedback during our sessions. The researchers manually sent out real-time reminders whenever the participants used the filled pauses. While there were errors, even the commercial counterparts failed to recognise trigger words often and had delays; we concluded that minor errors would be irrelevant to drawing qualitative insights.

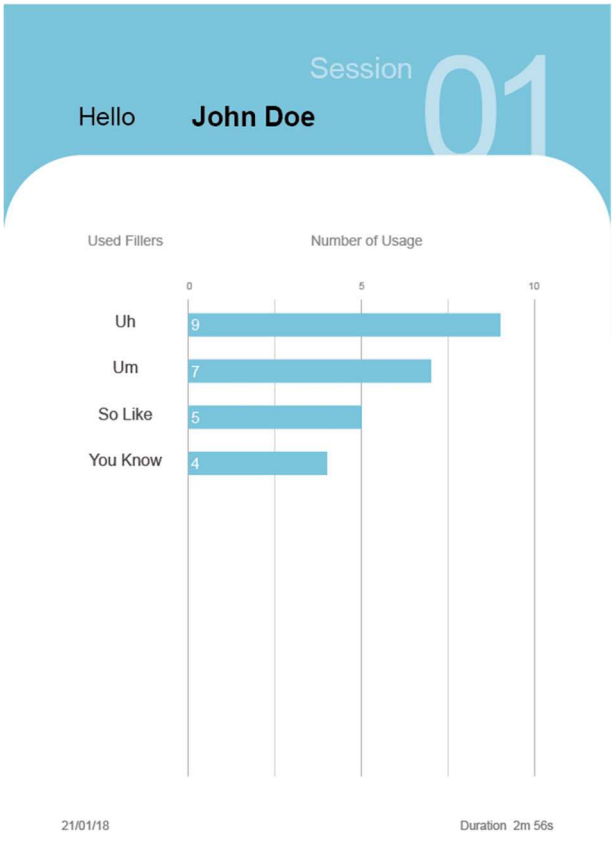

Figure 2. An example post-feedback (report) of filled pause usage showing the participant's name, session number, list of used fillers during the session and their occurrence, date of the subject session, and the duration of the speech. The participants were sent the pdf file during pre-interviews and were able to scroll through them.

To create the report, we used VREW, a speech-to-text converter, to convert the recorded speech into text files. While the created text files did not perfectly dictate the speech, especially certain filled pauses such as 'uh' or 'um,' three researchers re-listened to the recorded speech files multiple times to manually perfect the dictation. Once the dictated text file was completed, we ran computerised code on the text files to count the frequency of each filled pause during the speech and created the report.

3.2. Interview groups

We divided the interview participants into three different groups: (1) the control group (CG) was not given real-time feedback or reports, (2) Experimental Group 1 (EG1) was given reports only, and (3) Experimental Group 2 (EG2) was given both the real-time feedback and reports.

The division was done for two reasons: (1) to compare the effect of each feedback system to validate our model as automated visual feedback on speech, and (2) to comparatively analyse the interview responses to see how user experiences differ and coincide according to the forms of feedback.

For groups that received reports (EG1 & EG2), the reports were provided at the start of the next session, rather than giving them right after the speech ended. This was done to maximise the effects of feedback on the following speech. For EG2 who received both forms of feedback, the real-time feedback was provided only for the second and third sessions. This was done to provide the same number of feedbacks for both forms (If we were to give real-time feedback for the first session, EG2 would have had two sessions with reports and three sessions with real-time feedback, leading to a mismatch between treatments). Table 2 provides a visual description of the groups and provided feedbacks.

3.3. Participants and setting

27 participants (16 female, 11 male) were recruited through online communities, with the inclusion criteria as follows: (a) natives of the Korean language, (b) in their early 20s, in either 3rd or 4th year of university, (c) planning to participate in an interview for a job or higher education within two years, and (d) had no

Table 2. Interview groups and received forms of feedback.

		Control Group ($n = 9$)	Experimental Group 1 ($n = 8$)	Experimental Group 2 ($n = 9$)
Session 1	Before Speech			
	During Speech			
Session 2	Before Speech		Quantified Report on the Previous Session	
	During Speech			Immediate Feedback
Session 3	Before Speech		Quantified Report on the Previous Session	
	During Speech			Immediate Feedback

prior experience in speech coaching (such as mock-interview or speech academy). The 27 participants were randomly divided into three groups. All participants were required to participate in three individual interview sessions spread across a span of five days, allowing them to rest and have enough time to reflect on the interventions provided. Sessions were conducted online through video calls and recorded upon agreement. All participants were given 20,000 KRW (approximately 17 USD) cash as compensation for their time and participation.

3.3.1. Pre-interview

All sessions started with a semi-structured pre-interview. In the first session, participants were asked to share their motives for wanting to improve their language skills, their biggest difficulty in doing so, and their attempts to improve their speaking skills. In the second and third sessions, participants were asked to recall their previous sessions and share their general impressions. EG1 and EG2 were also provided with a report of their filled pauses usage. They were then asked to share their impressions and thoughts about the report.

3.3.2. Speech preparation

After the pre-interview, participants were given 10 min to prepare their speeches. During the preparation time, the screen showed the speech topic and the evaluation criteria that will be used during the self-evaluation after the session to inform them what factors should be considered for a good speech (Babaii, Taghaddomi, and Pashmforoosh 2016). They were asked not to write the scripts in full sentences, but to briefly write keywords and create an outline for their speech. The topics were chosen from the TOEFL (Test of English as a Foreign Language) speaking test, as it is one of the most respected standardised English proficiency tests in the world. All participants were given different topics per session, but the topics were the same for all participants. The questions were translated into Korean and reviewed by all researchers to validate the clarity of the translation. The questions were as follows:

- Session 1: Which company do you prefer to work with, a company where you work independently, or a company with a tutor to help you?
- Session 2: Do you prefer to finish things early or wait until the deadline is almost there?
- Session 3: If you get extra money, would you save the money or spend it immediately?

3.3.3. Speech

Researchers notified the users when the preparation time was over, and asked them to present their speech for three minutes. EG2 participants were given real-time feedback during the speech whenever they displayed filled pauses in the second and third sessions. During the speech, one of the researchers manually sent feedback whenever the researcher recognised the utterance of filled pauses. Participants were told that feedback was provided by the newly developed software. After all three sessions were completed, they were notified that this was a Wizard of Oz experiment.

3.3.4. Self-Evaluation

For all three sessions, after the speech, participants were asked to evaluate their speech based on self-assessment criteria ranging from 1 (poor) to 5 (excellent), chosen from the speech assessment criteria developed by Babaii and colleagues (Babaii, Taghaddomi, and Pashmforoosh 2016): (1) fluency: (without pauses, hesitation, and false starts), (2) grammar (accuracy and variety of structures), (3) vocabulary (appropriateness and variety of expressions), (4) pronunciation (stress, rhythm, and intonation), (5) communicative effectiveness (clarity of ideas and comprehensible (i.e. understandable) speech); (6) topic management (topic relevance, topic coverage, and adequacy of details and examples), (7) confidence (anxiety-free speech), and (8) organisation (initiation, development, termination, and interconnectedness of ideas). The self-evaluation scores were taken into account to understand how the participant's perception of their own speech differs according to the given form of feedback. Regardless of the actual decrease in filler usage, we hoped that these scores could provide us with better insights into the users' satisfaction and overall user experience in using automated visual feedback on speech.

3.3.5. Post-interview

A semi-structured post-interview was conducted with the participants to discuss their general experience in all three sessions. They were asked to share their main goal for the session, what bothered them most, elaborate on their self-evaluation scores, and what filled pauses they noticed during the speech. For the second and third sessions, they were also asked to share how much they thought they had improved and in what aspect. For EG1 and EG2, we asked them to share how they felt about the feedbacks provided. In all sessions, the participants' sessions were recorded and transcribed under consent.

3.4. Data analysis

For quantitative analysis, we aimed to see if our model qualifies as a system that reduces the use of fillers. We observed the change in the frequency of filled pauses and self-evaluation scores according to the forms of interventions provided over time. Data from 27 participants and 81 sessions were collected (CG [n = 9], EG1 [n = 9], EG2 [n = 9]). We decided to provide descriptive statistics for two reasons: (1) the focus of our study leans toward qualitative insights rather than discovering whether our model is statistically significant, and (2) the sample size is not big enough to conduct comprehensive statistic analysis. The ratio of filled pauses per total word count of speech (FP/W) was used to ensure proper comparison among participants. We conducted a paired t-test to our primary metric, FP/W of the first and last session to see how significantly the usage of fillers increased or decreased according to the provided feedback format. We also compared all the self-evaluation scores of the first and last sessions.

For qualitative analysis, we explored the participants' comments and reactions to our system to draw design implications. A thematic analysis (Clarke, Braun, and Hayfield 2015) was conducted based on the written manuscript of the interviews as well as the observation notes made by the researchers. We took an inductive approach to thematic analysis with no specific coding schemes determined beforehand and let themes naturally emerge during data analysis. The researchers went through the manuscript one by one to identify the themes and label significant responses with adequate tags. The final analysis was conducted by putting the themes together and relating them to create theoretically meaningful results. All processes were revised and approved by internal review boards of the relevant institutes.

4. Results

4.1. The effect of visual feedback system on speech to filled pause usage and self-evaluation

In this section, we discuss how effective the proposed system was in reducing filled pauses for participants and how their reactions varied or coincided according to the assigned groups. Due to the small sample size and short span of the monitoring period, it will be difficult to draw a definitive conclusion quantitatively; despite such limitations, a trend was observed that varied according to the experimental groups. The results suggest that both forms of intervention significantly reduce the usage of filled pauses but providing both quantified reports and immediate feedback produced higher significance.

The filler usage reduction of CG participants was the lowest with 20.75%. EG1 participants who received only quantified reports showed a slightly higher reduction of 29.98% (excluding an outlier who exhibited a remarkably low number of fillers from the first session, resulting in a striking 472.52% increase). Finally, EG2 participants who received both quantified reports and real-time feedback showed the biggest reduction of 50.48%. We conducted a paired t-test on the first and last sessions' FP/W to confirm statistical significance. The paired t-test results on the EG1 FP/W in the first session ($M = 0.0782$, $SD = 0.036$) and the last session ($M = 0.0522$, $SD = 0.033$) suggest that providing only quantified reports resulted in a significant reduction in filled pauses ($t(7) = 2.5464$, $p = .0383$). The paired t-test result on EG2 FP/W in the first session ($M = 0.0591$, $SD = 0.02$) and the last session ($M = 0.0291$, $SD = 0.017$) suggested that providing both interventions resulted in the reduction of filled pauses with greater significance ($t(8) = 4.5298$, $p = .0019$). (Table 3)

In short, while all participants exhibited a decreasing trend in filler usage after getting used to the sessions over time, the reduction rate was more dramatic when provided with feedback, especially in both forms. Thus, we considered our model validated and continued our qualitative analysis. Detailed data on each participant's filler usage can be found in Appendix 1.

As for the self-evaluation scores comparison between the first and last session, the average score of CG participants decreased by 2.07%. EG1 participants who received only quantified reports showed an increase of 8.92% (9.62% excluding the outlier), and EG2 exhibited a bigger 12.56% increase. Despite all groups exhibiting a reduction in filler usage, only the participants that received feedback exhibited an increase in self-evaluation scores. While a detailed analysis will be provided

Table 3. Paired T-Test results on the first and the last session's filled pauses per words.

Groups	Session 1 M(SD)	Session 3 M(SD)	*dF*	*t*	*p*
Filled Pauses per Words					
Control Group	0.0892 (0.024)	0.0729 (0.033)	8	2.2210	0.0571
Experimental Group 1	0.0782 (0.036)	0.0522 (0.033)	7	2.5464	0.0383*
Experimental Group 2	0.0591 (0.02)	0.0291 (0.017)	8	4.5298	0.0019**

*$p < 0.05$,**$p < 0.01$,***$p < 0.001$

afterward, our qualitative analysis suggests that such a phenomenon took place because participants have objective data to reflect upon their decreasing trend (refer to 4.3.2 & 4.4.2). Detailed data on each participant's self-evaluation score trend can be found in Appendix 2.

4.2. User perception and experience of filled pauses

4.2.1. Filled pauses are not always bad but negatively perceived in general

There were some instances when participants reported that some filled pauses were used positively. These included when a filled pause performed a certain function, such as changing the tone of speech, expanding logic, or downplaying their expressions to make their speech modest; however, the majority reported that the overuse of filled pauses gives the impression of being unprepared or being too nervous to the listeners. Such an impression leads to an image of the speaker as being unconfident.

> 'It does not feel like proper language for the speaker […] It seems like you either forgot what to say, are nervous, or want to leave this situation.' (EG2, P11, Session 1)

4.2.2. It is hard to sustain awareness, thus external help is needed

Despite the negative perception, filled pauses were considered a secondary problem to solve, compared to the content and organisation of the speech. Participants were intrigued with the idea of using fewer filled pauses; however, the effort needed was excessive compared to the benefits as it is difficult to focus on the content and be mindful of the filled pauses simultaneously. This characteristic calls for interference on the part of the feedback system, as the system can constantly pay attention to filled pauses while letting the speaker focus on the core factor, such as the content, of the speech.

> 'I focus more on the delivery of the content itself, and take notice of repeated words only at that moment. It's secondary but whenever the red light lit up, it helped me notice during those moments. But still, I think delivering what I want to say is more important.' (EG2, P15, Session 2)

4.2.3. Dissonance between the stream of thoughts and the stream of speech

Participants reported that filled pauses often occur to fill the hesitating void during speech. As they speak, there are moments when they have to pull themselves together to think of what to say next. During this void, they found 'the silence unbearable and wanted to continue the stream of speech.' (EG2, P11, Session 2) Participants also reported using more filled pauses when they felt rushed. This can also be explained through the initial interpretation; as they speak faster, their 'speaking speed' exceeds their 'thinking speed', making more voids, and ultimately more fillers.

4.3. Role of graph-based post-feedback

4.3.1. A novel form of feedback on speech

Participants found number-based feedback on language data to be extremely interesting. Generally, human-given feedback is often in a qualitative format. Using words, they describe and analyse the speech, focusing on content, topic, organisations, etc. While the feedback may point out the overusing of filled pauses (or any other form of undesirable habits), they fail to give specific numbers ('you should use fewer fillers'). On that note, a quantified report on speech behaviour was a fairly new form of feedback for the user.

> 'You think you know your speaking habits well, but when someone else judges it and gives precise data, problems can arise in areas outside of what is being evaluated.' (EG1, P5, Session 3)

> 'In the past, I blindly tried to use fewer words that I thought I overused, but this time, because I received specific feedback, I tried to use those words less.' (EG1, P3, Session 2)

Providing them with hard numbers to work with motivated them much more than simply telling them that they used excessive fillers. Participants were initially

shocked to see that their filler usage far exceeded their expectations. Later in the second and third sessions, participants reported that they found themselves noticing the use of filled pauses even during day-to-day conversations outside of the sessions.

> 'When I first received the report, I didn't know how much I used them and what words I used, so I was shocked.' (EG2, P19, Session 3)

> 'In the first and second session, it felt like everything I said was analyzed and that was pretty much it. But during the last session, I sincerely felt like I wanted to fix it.' (EG2, P17, and Session 3).

4.3.2. Tracking progress

Several participants said that they did not realise their improvement throughout our sessions until they saw all the reports together. These reports functioned as an objective point of reference for them. Based on the reports, participants decided whether they were content with the results or dissatisfied and whether they should keep on with their efforts. It was also notable that during the final session, the participants actively compared all their previous sessions and sought to understand the changes over time.

> 'I actually thought the first and the second sessions' results would not have much difference, but it seems like I did fix errors when I became aware of them. The result is completely different from my expectation since it shows that I repeated the words less frequently' (EG2, P15, Session 3).

4.4. Role of real-time feedback

4.4.1. Reminding users to stay mindful

Real-time feedback seems to have helped participants sustain their mindfulness. During the interview, many participants said that they 'wanted to fix it, but it was difficult to fix it while they were speaking' (EG2, P17, Session 2). This is where real-time feedback comes into play; users notice the red screen flicker, which immediately reminds them to stay alert. When the screen lights up, it causes the participants to instantly look back on what they have just said and polish their speech immediately.

> '(When the screen lit up) I instantly thought 'What did I just say?' By doing so, I was able to control, polish, and reorganize the words that I used.' (EG2, P11, and Session 3).

4.4.2. Gives a sense of how well they are doing but disrupts speech

The participants also utilised real-time feedback to judge whether they were doing a good job during the speech. Based on how often the screen lit up, they were able to tell how well they were doing. While some found this useful, others were intimidated by it and lost confidence. Such findings coincide with the research on habit reversal, in which they discuss that while awareness training is indeed effective, it also results in decreased confidence for some users (Pawlik and Perrin 2020). Such an idea seems to be in line with Lupton's discussion on a self-tracking culture that makes users overly focused on their health, leading to a feeling of failure, anxiety, or self-hatred (Lupton 2013)

> 'I felt like I was doing something wrong and lost confidence. It startled me too much.' (EG2, P17, and Session 2).

Some participants also reported that real-time feedback ultimately disrupted the quality of their speech because they were overly focused on using fewer filled pauses rather than focusing on other important factors such as the content of the speech.

> 'Whenever I used the words and the red light flickers, I had to get rid of that word and instantly re-create my speech. So, I felt like the second session was more difficult than the first one, where I was given feedback. The usage of the words may have decreased, but I felt like I lost something more important while chasing something small.' (EG2, P15, Session 3)

It is also notable that some users were startled at first, but eventually reported getting used to it to the point of ignoring them. The question of whether they simply got used to it and ignored them entirely or they still noticed the feedback on an unconscious level, where it helped them to stay mindful, is a worthwhile topic for future research. (Figure 3)

> 'At first, it startled me, but after the third feedback, I just ignored it. It felt like being startled negatively affected my presentation. I do not think I actively tried to ignore it, but the level of shock naturally decreased.' (EG2, P14, and Session 2).

4.5. Role of feedback within the 'timeline of speech'

Timeline of speech and limited human capacity. Based on the participants' remarks and our observations, we propose a framework 'timeline of speech,' which displays the users' behavioural status at specific time points associated with the different speech stages (see Figure 4). Before the speech, users are often nervous, anxious, and pressured to speak well; however, once they start speaking, they can go into a trance-like state where they are unable to feel their previous anxiety and the surroundings to solely focus on what they are saying.

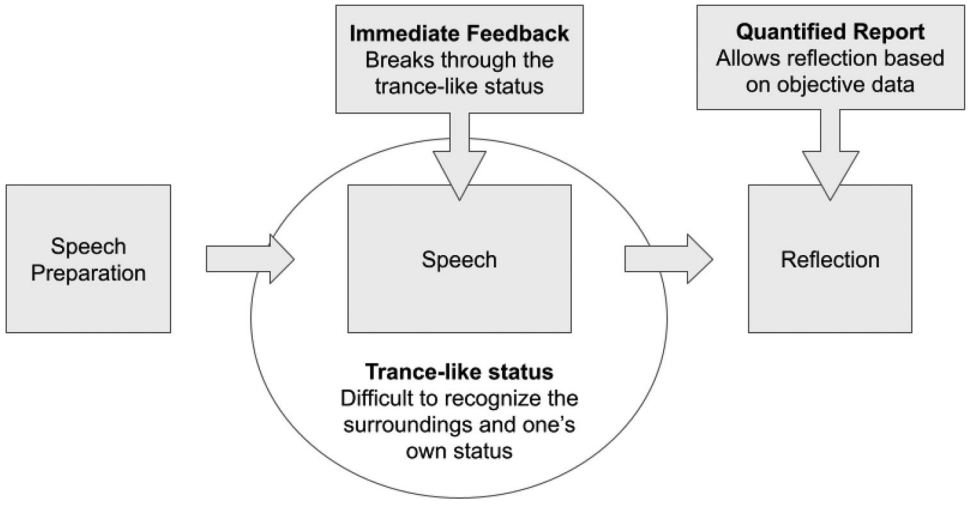

Figure 3. Timeline of speech and feedback.

> 'In the beginning, I get really nervous. But when I actually start speaking, I do not feel anything … I get nervous to the point of feeling like staring at a blank paper, but when I am speaking, I do not notice that I am nervous myself.' (EG2, P15, Session 1)

Once the speech is over, users reflect upon how well they did, but due to the limited capacity of working memory, the recollection of the trance-like status is often not complete and flawed. We suspect such a result to be in line with the cognitive load (Sweller 2011), and will expand further in our following section as to how it should be considered during the design of feedback systems. Some participants reported recording their speech to personally identify problems. These efforts are often too demanding and only provide subjective references with no hard data.

> 'After I talk, it feels like my memories evaporate. When I was self-evaluating my speech after the session, I asked myself, 'Was I really fluent?' I only remember speaking and nothing else, so I was not able to tell if other people thought I was fluent enough.' (EG1, P1, Session 2)

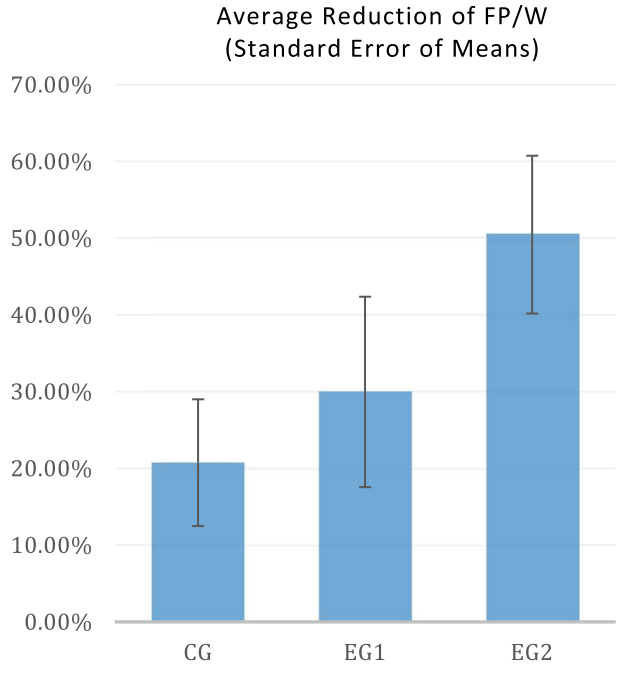

Figure 4. Mean of filled pauses per word according to groups with error bars (standard error of means).

During the Reflection phase of the timeline, an automated quantified report came as a novel, efficient, and reliable method for speech self-evaluation. Several existing studies have already discussed the power of numbers in letting users know of their problems as a more scientific and neutral approach (Lupton 2014; Wolf 2010). Users also utilised their previous data as a baseline against which they could compare their changes. This behaviour is consistent with previous research that emphasises the role of numbers in allowing comparisons between present and past data.

During the Speech stage where users go into a trance-like status and become less aware of anything other than the speech itself, real-time feedback comes into play, provoking users to stay mindful of their unwanted habits. However, when the warning is stronger than it should be, it obstructs the users' trance-like status and ultimately harms the quality of the speech. Thus, future work should dive deeper into forms of non-obstructive feedback that could sustain the user's focus on speech while letting the user know when unwanted behaviour occurs.

5. Design implications

Based on the analysis of existing research and our interview results, we present eight design implications to

consider when creating an automated visual feedback system on speech. Here, we would like to emphasise that the materials should be understood within the context of our proposed system, and hope to see them expand through the course of the following research efforts. (Table 4)

5.1. Target of the system: secondary problem for the users, still important for speech

According to our results, participants considered filled pauses a secondary problem to solve, compared to other factors such as the content and organisation of the speech (see 4.2.2). They were intrigued with the idea of using fewer filled pauses; however, the effort needed was excessive compared to the benefits, as it is difficult to focus on the content and be mindful of the filled pauses simultaneously. This characteristic calls for intervention on the part of an automated system, as it can constantly pay attention to filled pauses (or any other secondary target behaviour that designers and engineers deem worthy of attention) while letting the speaker focus on the core factors, such as content, of the speech.

5.2. Let users decide what 'undesired behaviours' are

Some participants ignored data on filled pauses that they considered positive (see 4.2.1). These participants found that certain filled pauses were essential to their speech to maintain naturalness and would be impossible to get rid of. As users' perceptions of positive language behaviour differ, the system needs to leave room for customisation. This idea is also in line with previous research that emphasised the importance of user autonomy in reiterating the results to find the optimal solution for themselves (Li, Dey, and Forlizzi 2010). There have also been research efforts on 'control features' in personal informatics, emphasising the importance of letting users feel that they are in control to choose how to interact with it (An et al. 2006; Crutzen, Cyr, and De Vries 2012; Jennett et al. 2008; McClure et al. 2013; Perski et al. 2017; Strecher et al. 2008).

When focusing on filled pauses as a service consideration, it is important to note these different types of pauses. We could go beyond the simple categorisation of 'good fillers' and 'bad fillers' to consider multiple types of filled pauses (in the case of Korean, the aforementioned categories), with each serving a different purpose during speech. While our research did not study specific data according to type, we expect future research to dive deeper into how users perceive each filled pause with access to a more detailed feedback system.

5.3. Decide what emotion you want to evoke

It was notable how our real-time feedback interfering with the trance-like status evoked the emotion of irritation (see 4.5). Existing studies suggest that sudden changes may disrupt flow (Costa et al. 2016) or impose

Table 4. List of design implications in relation to participants' remarks.

	Design Implications	Relevant Comments
1	Target of the system: secondary problem for the users, still important for speech	*'I focus more on the delivery of the content itself, and take notice of repeated words only at that moment. It's secondary but whenever the red light lit up, it helped me notice during those moments. But still, I think delivering what I want to say is more important.' (EG2, P15, Session 2)*
2	Let users decide what 'undesired behaviors' are	*'It does not feel like proper language for the speaker […] It seems like you either forgot what to say, are nervous, or want to leave this situation.' (EG2, P11, Session 1) 'To make a good speech, I think the flow is important. In doing so, having a lot of propositions or markers that show the change of flow can be helpful for the listeners to clearly understand the flow, so I think it's positive.' (EG1, P8, Session 2)*
3	Decide what emotion you want to evoke	*'I felt like I was doing something wrong and lost confidence. It startled me too much.' (EG2, P17, and Session 2). 'At first, it startled me, but after the third feedback, I just ignored it. It felt like being startled negatively affected my presentation. I do not think I actively tried to ignore it, but the level of shock naturally decreased.' (EG2, P14, and Session 2).*
4	Let users keep track of their progress	*'I actually thought the first and the second sessions' results would not have much difference, but it seems like I did fix errors when I became aware of them. The result is completely different from my expectation since it shows that I repeated the words less frequently' (EG2, P15, Session 3).*
5	Provide specific guidelines for their next step	*'I thought the reports would be more detailed, like, showing where the target word was used in my sentences or something like that. I have never received feedback on my language habits, and when I got one, I thought, 'When did I use this? How should I fix this?' It was not that it did not help, but I would have appreciated something more.' (EG1, P2, Session 3)*
6	Consider the optimal timing for your feedback	*'If you were to show me the reports right after the first session, there may not have been any improvements in the second speech. Because you gave me the report right before the second speech, it stayed in my brain and I think it reinforced my awareness.' (EG1, P1, Session 2)*
7	Consider the context of the use	*'I think I'm more nervous during presentations or speech because it's special circumstances. Because I prepared more, I become more nervous.' (EG2, P13, Session 1)*
8	Consider the form of feedback	*- (Drawn from a comprehensive meta-level review of the existing literature)*

cognitive load which can further adversely affect inhibition and performance by triggering anxiety (Eysenck et al. 2007; Miyake et al. 2000). However, our results suggest that providing immediate feedback that disrupts such flow results in significantly lower use of filled pauses (see 4.1 and 4.4.2). This is also in line with the existing research which suggested that anxiety does not always have an adverse effect on performance and normally it is unlikely to affect language behaviour in a typical public speaking setting (King and Finn 2017). Given these conflicting effects, when designing a feedback system, we must carefully consider whether disruption is a negative user experience in this context; users may achieve their desired goal of reducing filled pauses through irritation. We are not contending that annoying the users with obtrusive feedback to the point where they are forced to improve is the answer; nor are we saying that providing kind and encouraging feedback, such as colloquial feedback given by a humanoid (Trinh et al. 2017) is the best option. Our insight here is that when it comes to changing human behaviour, various emotions could or could not work based on the context and personal taste of the users. Thus, it is important for designers to clearly set the emotion that their system is trying to evoke and to conceptualise how such an emotion would bring about the desired outcome.

5.4. Let users keep track of their progress

The participants' primary use of quantified data was to track the changes in their results and compare their current status with their past selves (see 4.3.2). Regardless of the experimental groups, one of the most common phrases the participants used during the interview was 'compared to the last session.' Through quantified data and visual representations, users did not simply repeat themselves over and over but were able to reflect upon their speech more effectively and observe their improvements over time. Thus, the system needs to reflect on and emphasise the ever-changing status of users through interfaces that allow for an intuitive comparison (the point of comparison may be expanded from past selves to other users or unified criteria). Existing research has already emphasised the role of numbers in creating a baseline for users to compare present and past data (Lupton 2013).

5.5. Provide specific guidelines for their next step

The participants also wanted the system to do more than simply display the status quo and provide specific guidelines on how to improve. While users were glad to see their current status, some felt lost after seeing the report because they were not given specific guidelines on what to do next. We must also acknowledge how simple statistical reports are increasingly being considered ineffective. (Murnane et al. 2020; Pawlik and Perrin 2020).

> 'I thought the reports would be more detailed, like, showing where the target word was used in my sentences or something like that. I have never received feedback on my language habits, and when I got one, I thought, 'When did I use this? How should I fix this?' It was not that it did not help, but I would have appreciated something more.' (EG1, P2, Session 3)

Thus, the system should consider going beyond simply providing numbers to guide or nudge users to the next step. For instance, previous research discussed the system that notified the participants when and in which situations they used certain filled pauses (Pawlik and Perrin 2020). Such a guide could tell users at what point they should be most mindful during the speech. The system could even go directly to tell users what they should do (Trinh et al. 2017).

5.6. Consider the optimal timing for your feedback

Another important consideration is the timing of feedback. Along with real-time feedback and post-feedback that we discussed, we identified an additional time point for feedback: the pre-feedback. Participants in EG1 and EG2 reported that they were able to maintain high awareness of refraining from filled pauses during the second and third sessions because they saw the previous session's report right before they began their speech. Participants reported that if the report had been given right after the speech, they would have overcome the initial shock by the time of the next session.

> 'If you were to show me the reports right after the first session, there may not have been any improvements in the second speech. Because you gave me the report right before the second speech, it stayed in my brain and I think it reinforced my awareness.' (EG1, P1, Session 2)

On such a note, it is still unexplored how the previous report provided right before the next speech practice would affect the users. Although our model proved to be effective for some users, future research may focus on the relationship between the timing of providing the reports and how differently it impacts the users.

We are not suggesting that there is an 'optimal time' for providing feedback; rather, we are emphasising that these different time frames are likely to affect users differently. Thus, designers, researchers, and engineers

must carefully explore how their systems could take advantage of each time frame.

5.7. Consider the context of the use

In this study, we created a system for online presentation situations; however, there are many more diverse situations where an automated feedback system on speech is being used and could be used. A day-to-day conversation where users may be warier of using vulgar languages would require vastly different requirements, creating a much different-looking intervention than what we have been seeing so far (not to mention the technical feasibility, as well). For instance, if we were to create a system for daily conversation, designers, and engineers must consider interference from others during conversations or difficulty in speech recognition. The existing literature on personal informatics emphasises the seamless, unobtrusive integration of data collection into daily activities, necessitating further research on user scenarios and design considerations (V. R. Lee 2014; Van Der Woude, Tetteroo, and Liang 2020).

5.8. Consider the form of feedback

Throughout our results section, we discussed the positive, negative, and other interesting user experiences for visual feedback on speech; graph and instant colour flickering, to be exact. As for the graphs, users were intrigued by the quantified analysis of their speech behaviour and found them intuitive to understand. They also found it easy to track their trend over time. However, users found it difficult to identify and move on to the next step to improve their speech (see 4.3). In the case of instant colour flickering, users easily synthesized and understood that they displayed unwanted behaviour once gotten used to the system; however, some found this to be obtrusive, leading them to focus on minor issues and forgetting the bigger goal (see 4.4.)

It is our understanding that these pros and cons would exist in other forms of feedback. Existing research has explored multiple forms of feedback such as colloquial (text) (Sedig, Klawe, and Westrom 2001; Tanveer, Lin, and Hoque 2015; Trinh et al. 2017), visual (icon, colour) (Clark and Fox Tree 2002; Kurihara et al. 2007; Tanveer, Lin, and Hoque 2015), numeric (graph) (Kurihara et al. 2007), or haptic ones (Schneider et al. 2015). Yet, we can safely assume that this list is not exhaustive; multiple forms of feedback still need to be explored, such as sonic feedback, as well as different media to provide them (wearable devices, computer screens, mobile screens, etc.). Even the flickering form of feedback may produce different results based on its colours or flickering durations. Rather than trying to find the 'best' version of feedback, we would like to encourage researchers to explore these various and innovative feedback forms to identify their pros and cons, possibly providing the designers with a chance to choose the best version of feedback for their system's goal.

6. Discussions

6.1. Automated speech feedback and personal informatics (Quantified self)

During our analysis, we noticed similarities between automated feedback systems on speech and personal informatics. Personal informatics (also known as 'living by numbers,' 'quantified self,' 'self-surveillance,' 'self-tracking,' 'personal analytics,' and 'lifelogging') refers to a computer-based system that collects, integrates, and presents personal data and the history of the users to enable self-reflection and enhance self-knowledge (Daskalova et al. 2017; Hixon and Swann 1993; Li, Dey, and Forlizzi 2010; Lupton 2014; Wolf 2009). To date, personal informatics has been used to analyse diverse aspects of our lives, including geolocation, weight, sleep, diet, mood, feeling, and especially health and fitness (Bickmore, Caruso, and Clough-Gorr 2005; Consolvo et al. 2006; Gasser et al. 2006; Lupton 2014; Steinfeld et al. 2015; Toscos et al. 2006). Personal informatics provides objective, quantified data to help users become aware of their behaviours, and ultimately, improve them. With computerised systems, personal informatics outperforms often-flawed human abilities, such as the inability to monitor themselves objectively, limited memory capacity, and inability to sustain awareness (Li, Dey, and Forlizzi 2010; Malacria et al. 2013; Sedig, Klawe, and Westrom 2001; Wilson and Dunn 2004).

On such a note, feedback systems on speech that automatically collect users' verbal data and present them to improve them almost seem like a branch of personal informatics. Thus, interdisciplinary research efforts may lead to a more fruitful discussion on how computerised self-improvement systems could better aid users.

For instance, personal informatics studies suggest that users desire improved self-understanding for various reasons, but mostly for self-improvement (Doherty and Doherty 2018). Raised self-knowledge and improved self-insight were reported to allow users to recognise their behaviours (Hixon and Swann 1993; Seligman and Darley 1976), leading to behavioural change, often in the form of improving themselves by solving the problem reported by the system. Some

enthusiastic self-quantifiers also reported that, by studying themselves, they desire to be an 'optimal human being' and 'your best self' (Lupton 2014). If we are to apply such an understanding of the users when designing automated feedback systems on speech, designers, and engineers will have a much clearer understanding of the logic and ultimate goal of providing users with self-knowledge.

Studies on automated feedback systems on speech may enrich the field of personal informatics, as well. Researchers in personal informatics have already emphasised the importance of new types of behavioural information for measurement, and human speech could be a good point to expand on such topics (Li, Dey, and Forlizzi 2010; 2012). Furthermore, personal informatics has emphasised the importance of presenting objective data through visual reports (Bode and Kristensen 2015; B. Lee et al. 2020). Seeing that the field of automated feedback systems on speech has explored various visual feedback methods from coloured gauges, to simple texts, and icons, personal informatics may be able to take further inspiration in creating more diverse feedback methods. We are excited to see how the collaboration between the two fields could further expand both agendas, leading to a more enriched discussion within the HCI community focusing on human improvement through computerised systems.

6.2. Speech aid, self-determination theory, and motivation

To aid users' speech improvement successfully, the speech aid system needs to ensure that the user is continuously motivated. To this end, self-determination theory (SDT) provides an interesting perspective. SDT discusses how behaviours are anchored in either controlling motivations (such as pleasing other people or satisfying self-esteem) or autonomous motivations (such as enjoying the activity itself) (Deci and Ryan 1985; 2002). Introjected or external motivations are considered controlling, non-self-determined forms of motivation, whereas internal motivations such as intrinsic or identified motivations are considered autonomous, self-determined forms of motivation (Deci and Ryan 1990). Ferrand and colleagues suggested that autonomous motivations provide longer-lasting motivations to people compared to controlling motivations (Ferrand, Martinent, and Bonnefoy 2014). In this regard, SDT is already being used to categorise and evaluate motivating applications (Villalobos-Zúñiga and Cherubini 2020).

Our interviews revealed that the motivation for improving speaking skills is closely related to controlling motivation, as the primary goal is to improve the public image and gain the trust of the audience; however, according to SDT, this may have been the reason why the motivation and efforts to improve speech habits did not last long. As the sessions continued, several participants expressed that the results themselves were interesting to watch and that they enjoyed the experience. Thus, it seems worthwhile to examine whether and how new forms of feedback systems on speech can foster autonomous motivation.

6.3. Speech aid systems and socio-ethical issues

Speech aid systems aim to change the linguistic behaviour of users by monitoring their speech and, in doing so, pose several ethical questions concerning their methods and goals. The first point of discussion is the concept of the 'perfectible body'. Lupton discusses how personal informatics introduces the concept of 'technological bodily enhancement and techno-utopian visions of the perfect(ible) body.' Personal informatics transformed how bodies were conceptualised, touched, managed, and displayed to users (Lupton 2013). Such an idea poses an interesting question on how personal informatics of language would alter how we perceive language in terms of technology. Will our language also become another subject to 'perfect(ify)?' If so, how would it change our perspective on linguistic life? How will a 'perfect language' be defined and by whom? How would the technology evolve and what should be considered to cause less harm? It will be an exciting journey to observe and discuss as the system integrates into our daily lives.

Designers and researchers should also ponder the responsibilities to the users in creating these systems. During the experiment, we identified how users pay greater attention to the factors that the system deems important (in our case, filled pauses); however, the point of emphasis was chosen by the system designers. Thus, the system needs to identify what users truly desire to create responsible services. This point may also be consulted with one of our design implications where we emphasised the importance of letting users define what undesirable behaviours are.

As a system that quantifies user behaviour, the system must always be cautious concerning the number of effects on users. The existing literature discusses the politics of measurement. Despite the numbers' association with science, objectiveness, and neutrality, the way phenomena are quantified, interpreted, and presented is largely affected by the providers and receivers (Lupton 2014). The aforementioned participants, who

lost confidence because of the numbers provided by the system, could be a great example; if the operator wishes to hold onto the users, creating an interface that shocks the users and ceasing their confidence may encourage user dependency on the system. It is always important to study the potential negative effects of numbers on users when creating such systems.

The final point is privacy. Previous research on personal informatics has discussed how technology expanded from purely private to social institutions. In the fields of education, medicine, and health, agents in their respective fields have encouraged people to use personal informatics for enhancement in each field. Lupton warns of the possibility of self-tracking technologies becoming surveillance tools, as the data collected are no longer being used only for private reasons (Lupton 2014). As speech aid systems provide a huge opportunity for language data collection, the discussion on the potential invasion of privacy must be addressed while developing such systems.

Nafus and Sherman argue that 'self-quantifiers use their bodies and cultural resources around them to see outside the frame that is set by technological devices' (Nafus and Sherman 2014). We hope that the trend of computer-based speech aid systems can expand the frame in which users perceive language to gain valuable experience in self-monitoring to improve their speech.

6.4. Limitations

Despite our findings, there are several limitations to this study. The research sessions were spanned out over only a week, a too short period to draw meaningful longitudinal data. Additionally, as linguistic behaviour differs according to controlled conditions and daily conditions, the effects of the feedback system may vary outside of our speech session settings. The presentations were conducted with nothing at stake, but in real-life situations, the behaviours may change, as their performance may affect their grades or investment opportunities. We hope these limitations could be consulted upon for future research that hopes to discuss similar topics. Furthermore, during the Wizard of Oz experiment of immediate feedback, there were human errors whereby the researcher failed to recognise some of the target behaviours. The average error rate frequency was 3.44 (min 0, max 9). Despite errors, we decided to present our findings considering that this was qualitative research that discusses users' experience with the system and even the commercial hot-word recognising services occasionally fail to recognise a key phrase.

6.5. Future works

There were also several points to be further investigated through future research efforts. While there were other trends observed in various self-evaluation categories, we failed to discover meaningful insights within the scope of our interview analysis. We hope to see more extensive research on this topic both quantitatively and qualitatively to deepen our understanding of the relationship between various feedback formats and user satisfaction (self-evaluation). Moreover, as for the immediate feedback given in real-time during a speech, it seems important to investigate the nature of users getting used to them. Understanding whether they reach a point of ignoring them or unconsciously recognising them would allow the system to make better use of immediate feedback. Additionally, the obtrusiveness of other immediate feedback formats (such as haptic or sonic reminders) during a speech, and whether such obtrusiveness benefits or hinders the quality of speech, would also further enrich the discussion. The timing in which the delayed feedback is provided is also an important topic to discuss. Our results suggested that providing the results right before the following session reinforces the users' mindfulness, but a more concrete confirmation would enrich the system design of feedback systems. Another important avenue for further investigation is the extent to which users retain their awareness of various linguistic habits, considering factors such as feedback format, frequency, repetition, and timing. The concept of retention can be examined by employing the forgetting curve model, and its interpretation can be intertwined with the retention of users' motivation. Finally, much research needs to be conducted on different formats, languages, and speech situations to create more exhaustive and concrete design considerations. We hope our research could contribute as a valuable discussion point for operational design consideration for feedback systems on speech.

7. Conclusion

Based on empirical user interviews, our study expanded the currently existing discussion of automated visual feedback systems on speech and provided a list of design implications that should be considered when creating such systems. The guideline we provide is most likely not exhaustive and will need more evaluation and improvement. We look forward to how our efforts evolve in the future works of both automated feedback systems on speech and personal informatics. In addition to our main findings, we also identified multiple areas of discussion concerning automated visual feedback

systems on speech. We hope our research contributes to a more fruitful discussion for the HCI community and look forward to future studies on feedback systems on speech with diverse points of usage and innovative intervention forms to enhance their positive effects on the user and our society.

Acknowledgements

Hojae Lee and Hojae Han provided valuable support for creating software for the quantified reports. We are also grateful to Dong Whi Yoo and Taewook Kim for their comments on an earlier version of the manuscript. We convey our gratitude to all those who participated in our survey, interview, and experiments, as well as to the reviewers who enriched our paper with insightful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Soonho Kwon http://orcid.org/0000-0002-2783-6364
Hyojin Lee http://orcid.org/0000-0002-3383-9270
Soojeong Park http://orcid.org/0000-0002-0268-611X
Yerin Heo http://orcid.org/0000-0001-8606-6913
Keeheon Lee http://orcid.org/0000-0002-3291-5753
Younah Kang http://orcid.org/0000-0003-4981-6329

References

An, L. C., C. L. Perry, E. B. Lein, C. Klatt, D. M. Farley, R. L. Bliss, D. J. Hennrikus, et al. 2006. "Strategies for Increasing Adherence to an Online Smoking Cessation Intervention for College Students." *Nicotine and Tobacco Research* 8 (SUPPL. 1), https://doi.org/10.1080/14622200601039881.

Azrin, N. H., and R. G. Nunn. 1973. "Habit-reversal: A Method of Eliminating Nervous Habits and Tics." *Behaviour Research and Therapy* 11 (4): 619–628. https://doi.org/10.1016/0005-7967(73)90119-8.

Babaii, E., S. Taghaddomi, and R. Pashmforoosh. 2016. "Speaking Self-Assessment: Mismatches between Learners' and Teachers' Criteria." *Language Testing* 33 (3), https://doi.org/10.1177/0265532215590847.

Bell, R. L.. 2011. "Is your speech filled with um? 3 tips to eliminate filled pauses from your professional presentation." *Supervision* 72: 10–13.

Bernsen, N. O., H. Dybkjær, and L. Dybkjær. 1993. "Wizard of Oz Prototyping : When and How ?" *Cognitive Science* 94: 1–13.

Bickmore, T. W., L. Caruso, and K. Clough-Gorr. 2005. "Acceptance and Usability of a Relational Agent Interface by Urban Older Adults." *Conference on Human Factors in Computing Systems – Proceedings*. https://doi.org/10.1145/1056808.1056879.

Bode, M., and D. B. Kristensen. 2015. The Digital Doppelgänger Within: A Study on Self-Tracking and the Quantified Self Movement. In *Assembling Consumption: Researching Actors, Networks and Markets*. https://doi.org/10.4324/9781315743608.

Bubel, M., R. Jiang, C. H. Lee, W. Shi, and A. Tse. 2016. AwareMe. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 68–73. https://doi.org/10.1145/2851581.2890633.

Chollet, M., S. Marsella, and S. Scherer. 2022. "Training Public Speaking with Virtual Social Interactions: Effectiveness of Real-Time Feedback and Delayed Feedback." *Journal on Multimodal User Interfaces* 16 (1): 17–29. https://doi.org/10.1007/s12193-021-00371-1.

Chollet, M., T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer. 2015. "Exploring Feedback Strategies to Improve Public Speaking." *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1143–1154. https://doi.org/10.1145/2750858.2806060.

Clark, H. H., and J. E. Fox Tree. 2002. "Using uh and um in Spontaneous Speaking." *Cognition* 84 (1), https://doi.org/10.1016/S0010-0277(02)00017-3.

Clarke, V., V. Braun, and N. Hayfield. 2015. "Thematic Analysis." *In Qualitative Psychology: A Practical Guide to Research Methods* 3: 222–248.

Consolvo, S., K. Everitt, I. Smith, and J. A. Landay. 2006. "Design Requirements for Technologies that Encourage Physical Activity." *Conference on Human Factors in Computing Systems - Proceedings*, *1*. https://doi.org/10.1145/1124772.1124840.

Costa, J., A. T. Adams, M. F. Jung, F. Guimbretière, and T. Choudhury. 2016. EmotionCheck. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 758–769. https://doi.org/10.1145/2971648.2971752.

Crutzen, R., D. Cyr, and N. K. De Vries. 2012. "The Role of User Control in Adherence to and Knowledge Gained from a Website: Randomized Comparison between a Tunneled Version and a Freedom-of-Choice Version." *Journal of Medical Internet Research* 14 (2), https://doi.org/10.2196/jmir.1922.

Daskalova, N., K. Desingh, A. Papoutsaki, D. Schulze, H. Sha, and J. Huang. 2017. "Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (3), https://doi.org/10.1145/3130911.

Deci, E. L., and R. M. Ryan. 1985. "The General Causality Orientations Scale: Self-Determination in Personality." *Journal of Research in Personality* 19 (2), https://doi.org/10.1016/0092-6566(85)90023-6.

Deci, E. L., and R. M. Ryan. 1990. "A Motivational Approach to Self: Integration in Personality." *In Nebraska Symposium on Motivation. Nebraska Symposium on Motivation* 38.

Deci, E. L., and R. M. Ryan. 2002. "Self-determination Research: Reflections and Future Directions." In *Handbook of Self-Determination Research*, edited by Edward Deci and Richard M. Ryan, 431. Western New York, United States: University of Rochester Press.

Doherty, K., and G. Doherty. 2018. "The Construal of Experience in HCI: Understanding Self-Reports." *International Journal of Human-Computer Studies* 110: 63–74. https://doi.org/10.1016/j.ijhcs.2017.10.006.

Eysenck, M. W., N. Derakshan, R. Santos, and M. G. Calvo. 2007. "Anxiety and Cognitive Performance: Attentional Control Theory." *Emotion* 7 (2): 336–353. https://doi.org/10.1037/1528-3542.7.2.336.

Ferrand, C., G. Martinent, and M. Bonnefoy. 2014. "Exploring Motivation for Exercise and its Relationship with Health-Related Quality of Life in Adults Aged 70 Years and Older." *Ageing and Society* 34 (3), https://doi.org/10.1017/S0144686X12001092.

Fremouw, W. J., and J. L. Breitenstein. 1990. "Speech Anxiety." In *Handbook of Social and Evaluation Anxiety*, 455–474. Springer US. https://doi.org/10.1007/978-1-4899-2504-6_15.

Gasser, R., D. Brodbeck, M. Degen, J. Luthiger, R. Wyss, and S. Reichlin. 2006. "Persuasiveness of a Mobile Lifestyle Coaching Application using Social Facilitation." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3962 LNCS*. https://doi.org/10.1007/11755494_5.

Henderson, J., and R. Henderson. 2007. *There's No Such Thing as Public Speaking: Making Any Presentation Or Speech as Persuasive as a One-on-one Conversation*. New Jersey, United States: Prentice Hall Press.

Hixon, J. G., and W. B. Swann. 1993. "When Does Introspection Bear Fruit? Self-Reflection, Self-Insight, and Interpersonal Choices." *Journal of Personality and Social Psychology* 64 (1), https://doi.org/10.1037/0022-3514.64.1.35.

Hoque, M., M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. 2013. MACH. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 697–706. https://doi.org/10.1145/2493432.2493502.

Jennett, C., A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. 2008. "Measuring and Defining the Experience of Immersion in Games." *International Journal of Human Computer Studies* 66 (9), https://doi.org/10.1016/j.ijhcs.2008.04.004.

Jeon, Y. 2017. "Current Status and Tasks of Research in Korean Discourse Markers." *Korean Language Research* 45. https://doi.org/10.16876/klrc.2017..45.127.

King, P. E., and A. N. Finn. 2017. "A Test of Attention Control Theory in Public Speaking: Cognitive Load Influences the Relationship between State Anxiety and Verbal Production." *Communication Education* 66 (2): 168–182. https://doi.org/10.1080/03634523.2016.1272128.

Kurihara, K., M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi. 2007. "Presentation Sensei." *Proceedings of the 9th International Conference on Multimodal Interfaces*, 358–365. https://doi.org/10.1145/1322192.1322256.

Lee, V. R. 2014. "What's Happening in the "Quantified Self" Movement?" *ICLS 2014 Proceedings*.

Lee, B., E. K. Choe, P. Isenberg, K. Marriott, and J. Stasko. 2020. "Reaching Broader Audiences with Data Visualization." *IEEE Computer Graphics and Applications* 40 (2), https://doi.org/10.1109/MCG.2020.2968244.

Lee, J., and J. Jeong. 2018. "A Comparative Study on Oral Fluency Between Korean Native Speakers and L2 Korean Learners in Speech Discourse - With Focus on Speech Rate, Pause, and Discourse Markers." *Journal of Korean Language Education* 29 (4): 137–168. https://doi.org/10.18209/iakle.2018.29.4.137.

Li, I., A. Dey, and J. Forlizzi. 2010. "A Stage-Based Model of Personal Informatics Systems." *Conference on Human Factors in Computing Systems - Proceedings*, *1*. https://doi.org/10.1145/1753326.1753409.

Li, I., Y. Medynskiy, J. Froehlich, and J. Larsen. 2012. "Personal Informatics in Practice: Improving Quality of Life through Data." *Conference on Human Factors in Computing Systems – Proceedings*. https://doi.org/10.1145/2212776.2212724.

Lupton, D. 2013. "Quantifying the Body: Monitoring and Measuring Health in the age of MHealth Technologies." *Critical Public Health* 23 (4), https://doi.org/10.1080/09581596.2013.794931.

Lupton, D. 2014. "Self-Tracking Cultures: Towards a Sociology of Personal Informatics." *Proceedings of the 26th Australian Computer-Human Interaction Conference, OzCHI* 2014. https://doi.org/10.1145/2686612.2686623.

Malacria, S., J. Scarr, A. Cockburn, C. Gutwin, and T. Grossman. 2013. "Skillometers: Reflective Widgets that Motivate and Help Users to Improve Performance." *UIST 2013 - Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/2501988.2501996.

Mancuso, C., and R. G. Miltenberger. 2016. "Using Habit Reversal to Decrease Filled Pauses in Public Speaking." *Journal of Applied Behavior Analysis* 49 (1), https://doi.org/10.1002/jaba.267.

Maulsby, D., S. Greenberg, and R. Mander. 1993. "Prototyping an Intelligent Agent through Wizard of Oz." *Conference on Human Factors in Computing Systems – Proceedings*. https://doi.org/10.1145/169059.169215.

McClure, J. B., S. M. Shortreed, A. Bogart, H. Derry, K. Riggs, J. S. John, V. Nair, and L. An. 2013. "The Effect of Program Design on Engagement with an Internet-Based Smoking Intervention: Randomized Factorial Trial." *Journal of Medical Internet Research* 15 (3), https://doi.org/10.2196/jmir.2508.

Miltenberger, R. G., R. W. Fuqua, and D. W. Woods. 1998. "Applying Behavior Analysis to Clinical Problems: Review and Analysis of Habit Reversal." *Journal of Applied Behavior Analysis* 31 (3), https://doi.org/10.1901/jaba.1998.31-447.

Miyake, A., N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager. 2000. "The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis." *Cognitive Psychology* 41 (1): 49–100. https://doi.org/10.1006/cogp.1999.0734.

Murnane, E. L., X. Jiang, A. Kong, M. Park, W. Shi, C. Soohoo, L. Vink, et al. 2020. "Designing Ambient Narrative-Based Interfaces to Reflect and Motivate Physical Activity." *Conference on Human Factors in Computing Systems - Proceedings*, https://doi.org/10.1145/3313831.3376478.

Nafus, D., and J. Sherman. 2014. "This One Does not go up to 11: The Quantified Self Movement as an Alternative big Data Practice." *International Journal of Communication* 8 (1): 11.

Nowell, L. S., J. M. Norris, D. E. White, and N. J. Moules. 2017. "Thematic Analysis: Striving to Meet the Trustworthiness Criteria." *International Journal of Qualitative Methods* 16 (1), https://doi.org/10.1177/1609406917733847.

OfficeTeam. n.d. *Survey: Six in 10 Companies Conduct Video Job Interviews*. Accessed June 22, 2023. https://www.prnewswire.com/news-releases/survey-six-in-10-companies-conduct-video-job-interviews-167973406.html.

Parvis, L. F. 2001. "The Importance of Communication and Public-Speaking Skills." *Journal of Environmental Health* 63 (9): 44.

Pawlik, B., and C. J. Perrin. 2020. "Reducing Speech Disfluencies During Public Speaking Using Brief Habit Reversal." *Journal of Applied Behavior Analysis* 53 (2), https://doi.org/10.1002/jaba.627.

Perski, O., A. Blandford, R. West, and S. Michie. 2017. "Conceptualising Engagement with Digital Behaviour Change Interventions: A Systematic Review Using Principles from Critical Interpretive Synthesis." *Translational Behavioral Medicine* 7 (2), https://doi.org/10.1007/s13142-016-0453-1.

Schneider, J., D. Börner, P. van Rosmalen, and M. Specht. 2015. "Presentation Trainer, your Public Speaking Multimodal Coach." *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 539–546. https://doi.org/10.1145/2818346.2830603.

Sedig, K., M. Klawe, and M. Westrom. 2001. "Role of Interface Manipulation Style and Scaffolding on Cognition and Concept Learning in Learnware." *ACM Transactions on Computer-Human Interaction* 8 (1), https://doi.org/10.1145/371127.371159.

Seligman, C., and J. M. Darley. 1976. "Feedback as a Means of Decreasing Residential Energy Consumption." *Journal of Applied Psychology* 62 (4), https://doi.org/10.1037/0021-9010.62.4.363.

Steinfeld, B., J. Scott, G. Vilander, L. Marx, M. Quirk, J. Lindberg, and K. Koerner. 2015. "The Role of Lean Process Improvement in Implementation of Evidence-Based Practices in Behavioral Health Care." *The Journal of Behavioral Health Services & Research* 42 (4): 504–518. https://doi.org/10.1007/s11414-013-9386-3.

Strecher, V. J., J. McClure, G. Alexander, B. Chakraborty, V. Nair, J. Konkel, S. Greene, et al. 2008. "The Role of Engagement in a Tailored web-Based Smoking Cessation Program: Randomized Controlled Trial." *Journal of Medical Internet Research* 10 (5), https://doi.org/10.2196/jmir.1002.

Sweller, J. 2011. *Cognitive Load Theory* (pp. 37–76). https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

Tanveer, M. I., E. Lin, and M. Hoque. 2015. "Rhema." *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 286–295. https://doi.org/10.1145/2678025.2701386.

Toscos, T., A. Faber, S. An, and M. P. Gandhi. 2006. "Chick Clique: Persuasive technology to motivate teenage girls to exercise." *Conference on Human Factors in Computing Systems – Proceedings*. https://doi.org/10.1145/1125451.1125805.

Trinh, H., R. Asadi, D. Edge, and T. Bickmore. 2017. "RoboCOP." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2): 1–24. https://doi.org/10.1145/3090092.

Trinh, H., K. Yatani, and D. Edge. 2014. "PitchPerfect." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1571–1580. https://doi.org/10.1145/2556288.2557286.

Van Der Woude, W., D. Tetteroo, and R. H. Liang. 2020. "Presley: Designing Non-Obtrusive Tactile Rhythmic Wearable Devices for Improving Speech Fluency." *DIS 2020 Companion - Companion Publication of the 2020 ACM Designing Interactive Systems Conference*. https://doi.org/10.1145/3393914.3395880.

Villalobos-Zúñiga, G., and M. Cherubini. 2020. "Apps That Motivate: A Taxonomy of App Features Based on Self-Determination Theory." *International Journal of Human-Computer Studies* 140: 102449. https://doi.org/10.1016/j.ijhcs.2020.102449.

Wilson, T. D., and E. W. Dunn. 2004. "Self-knowledge: Its Limits, Value, and Potential for Improvement." *Annual Review of Psychology* 55. https://doi.org/10.1146/annurev.psych.55.090902.141954.

Wolf, G. 2009. "Know Thyself: Tracking Every Facet of Life, from Sleep to Mood to Pain, 24/7/365." *Wired Magazine*.

Wolf, G. 2010. "The Data-Driven Life - The New York Times." *New York Times* 17.

Zhao, R., V. Li, H. Barbosa, G. Ghoshal, and M. E. Hoque. 2017. "Semi-Automated 8 Collaborative Online Training Module for Improving Communication Skills." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2): 1–20. https://doi.org/10.1145/3090097.

Appendices

Appendix 1. Decrease rate of filled pauses per total words between the first and the last session for all groups.

EG1 (Only Report)				EG2 (Report + Real-time Feedback)				CG (No Feedback)			
	Filled Pauses / Total Words (FP/W)				Filled Pauses / Total Words (FP/W)				Filled Pauses / Total Words (FP/W)		
Participant Number	First Session	Last Session	Decrease Rate	Participant Number	First Session	Last Session	Decrease Rate	Participant Number	First Session	Last Session	Decrease Rate
P1	12.31%	4.52%	63.27%	P11	5.88%	4.15%	29.53%	P21	8.87%	8.04%	9.33%
P2	3.25%	3.27%	−0.41%	P12	3.65%	1.92%	47.42%	P22	10.67%	8.47%	20.55%
P3	6.30%	4.17%	33.90%	P13	8.84%	4.24%	52.11%	P23	10.31%	5.41%	47.51%
P4	1.34%	7.67%	−472.52%	P14	3.29%	0.00%	100.00%	P25	9.34%	9.84%	−5.28%
P5	13.36%	12.00%	10.19%	P15	4.79%	5.44%	−13.52%	P26	11.32%	7.11%	37.17%
P7	5.45%	7.01%	−28.57%	P16	7.41%	2.22%	70.00%	P27	8.27%	8.41%	−1.75%
P8	6.13%	3.29%	46.32%	P17	8.59%	2.98%	65.34%	P28	3.68%	2.01%	45.29%
P9	5.92%	1.21%	79.60%	P18	4.01%	1.82%	54.49%	P29	6.59%	3.40%	48.46%
P24	9.80%	6.32%	35.54%	P20	6.71%	3.42%	48.93%	P30	11.26%	12.89%	−14.49%
Average of Decrease Rate			29.98%	Average of Decrease Rate			50.48%	Average of Decrease Rate			20.75%
Standard Error of Average (Means)			12.41%	Standard Error of Average (Means)			10.29%	Standard Error of Average (Means)			8.26%

*P4's data was excluded as an outlier in calculating the average decrease rate for EG1

Appendix 2. Decrease rate of self evaluation scores between the first and the last session for all groups.

Fluency (without pauses, hesitation, and false starts)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	4	4	0.00%	P1	3	4	25.00%	P11	2	4	50.00%
P22	2	2	0.00%	P2	4	3	−33.33%	P12	4	5	20.00%
P23	2	3	33.33%	P3	3	3	0.00%	P13	2	4	50.00%
P25	2	3	33.33%	P4	4	4	0.00%	P14	4	4	0.00%
P26	1	3	66.67%	P5	3	3	0.00%	P15	3	3	0.00%
P27	4	3	−33.33%	P7	4	5	20.00%	P16	2	4	50.00%
P28	3	2	−50.00%	P8	3	2	−50.00%	P17	3	3	0.00%
P29	3	3	0.00%	P9	2	3	33.33%	P18	2	2	0.00%
P30	3	2	−50.00%	P24	3	4	25.00%	P20	3	4	25.00%
	AVERAGE		0.00%		AVERAGE		2.22%		AVERAGE		21.67%

Grammar (accuracy and variety of structures)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	3	2	−50.00%	P1	4	5	20.00%	P11	3	4	25.00%
P22	2	2	0.00%	P2	3	3	0.00%	P12	3	5	40.00%
P23	2	3	33.33%	P3	3	4	25.00%	P13	3	5	40.00%
P25	3	3	0.00%	P4	4	3	−33.33%	P14	4	4	0.00%
P26	1	3	66.67%	P5	2	3	33.33%	P15	4	4	0.00%
P27	4	3	−33.33%	P7	2	4	50.00%	P16	2	4	50.00%
P28	2	2	0.00%	P8	3	2	−50.00%	P17	4	4	0.00%
P29	2	4	50.00%	P9	2	3	33.33%	P18	4	4	0.00%
P30	3	2	−50.00%	P24	2	3	33.33%	P20	5	5	0.00%
	AVERAGE		1.85%		AVERAGE		12.41%		AVERAGE		17.22%

Vocabulary (appropriateness and variety of expressions),

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	2	3	33.33%	P1	3	4	25.00%	P11	2	4	50.00%
P22	3	2	−50.00%	P2	2	3	33.33%	P12	3	5	40.00%
P23	3	3	0.00%	P3	3	4	25.00%	P13	4	5	20.00%
P25	3	3	0.00%	P4	4	4	0.00%	P14	3	4	25.00%
P26	3	4	25.00%	P5	3	3	0.00%	P15	4	4	0.00%
P27	3	3	0.00%	P7	4	4	0.00%	P16	3	4	25.00%
P28	2	3	33.33%	P8	2	2	0.00%	P17	3	4	25.00%
P29	2	4	50.00%	P9	2	3	33.33%	P18	4	4	0.00%
P30	3	2	−50.00%	P24	3	3	0.00%	P20	4	5	20.00%
	AVERAGE		4.63%		AVERAGE		12.96%		AVERAGE		22.78%

Pronunciation (stress, rhythm, and intonation)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	2	3	33.33%	P1	4	5	20.00%	P11	2	2	0.00%
P22	3	3	0.00%	P2	4	2	−100.00%	P12	4	5	20.00%
P23	2	4	50.00%	P3	4	3	−33.33%	P13	4	4	0.00%
P25	3	3	0.00%	P4	5	4	−25.00%	P14	4	5	20.00%
P26	3	3	0.00%	P5	4	4	0.00%	P15	5	4	−25.00%
P27	4	3	−33.33%	P7	3	4	25.00%	P16	3	3	0.00%
P28	3	2	−50.00%	P8	4	4	0.00%	P17	4	4	0.00%
P29	2	2	0.00%	P9	3	4	25.00%	P18	2	2	0.00%
P30	3	2	−50.00%	P24	4	4	0.00%	P20	4	5	20.00%
AVERAGE			−5.56%	AVERAGE			−9.81%		AVERAGE		3.89%

Communicative Effectiveness (clarity of ideas and comprehensible (i.e. understandable) speech)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	4	4	0.00%	P1	3	4	25.00%	P11	3	4	25.00%
P22	3	3	0.00%	P2	3	4	25.00%	P12	4	4	0.00%
P23	4	2	−100.00%	P3	4	4	0.00%	P13	4	5	20.00%
P25	3	2	−50.00%	P4	3	3	0.00%	P14	4	5	20.00%
P26	2	4	50.00%	P5	3	3	0.00%	P15	4	4	0.00%
P27	3	3	0.00%	P7	4	5	20.00%	P16	4	3	−33.33%
P28	4	4	0.00%	P8	3	4	25.00%	P17	4	4	0.00%
P29	4	3	−33.33%	P9	4	4	0.00%	P18	2	2	0.00%
P30	3	2	−50.00%	P24	3	3	0.00%	P20	5	5	0.00%
AVERAGE			−20.37%	AVERAGE			10.56%		AVERAGE		3.52%

Topic Management (topic relevance, topic coverage, and adequacy of details and examples)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	4	3	−33.33%	P1	3	4	25.00%	P11	3	4	25.00%
P22	3	3	0.00%	P2	4	4	0.00%	P12	4	5	20.00%
P23	4	4	0.00%	P3	4	4	0.00%	P13	4	5	20.00%
P25	4	3	−33.33%	P4	4	4	0.00%	P14	5	4	−25.00%
P26	4	4	0.00%	P5	3	3	0.00%	P15	4	4	0.00%
P27	3	3	0.00%	P7	4	5	20.00%	P16	4	4	0.00%
P28	4	4	0.00%	P8	3	4	25.00%	P17	4	3	−33.33%
P29	4	4	0.00%	P9	3	3	0.00%	P18	2	4	50.00%
P30	3	2	−50.00%	P24	3	4	25.00%	P20	5	5	0.00%
AVERAGE			−12.96%	AVERAGE			10.56%	AVERAGE			6.30%

Confidence (anxiety-free speech)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	4	4	0.00%	P1	5	5	0.00%	P11	3	4	25.00%
P22	2	3	33.33%	P2	3	4	25.00%	P12	5	5	0.00%
P23	3	4	25.00%	P3	4	3	−33.33%	P13	4	5	20.00%
P25	2	4	50.00%	P4	3	4	25.00%	P14	4	5	20.00%
P26	1	3	66.67%	P5	3	4	25.00%	P15	4	5	20.00%
P27	5	4	−25.00%	P7	5	4	−25.00%	P16	4	4	0.00%
P28	4	2	−100.00%	P8	4	4	0.00%	P17	3	4	25.00%
P29	2	4	50.00%	P9	4	4	0.00%	P18	4	4	0.00%
P30	3	2	−50.00%	P24	2	4	50.00%	P20	3	5	40.00%
AVERAGE			5.56%	AVERAGE			7.41%	AVERAGE			16.67%

Organization (initiation, development, termination, and interconnectedness of ideas)

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	4	4	0.00%	P1	3	4	25.00%	P11	3	3	0.00%
P22	3	2	−50.00%	P2	3	3	0.00%	P12	4	5	20.00%
P23	4	4	0.00%	P3	3	3	0.00%	P13	5	4	−25.00%
P25	2	3	33.33%	P4	2	4	50.00%	P14	4	5	20.00%
P26	4	4	0.00%	P5	2	3	33.33%	P15	4	4	0.00%
P27	4	3	−33.33%	P7	4	5	20.00%	P16	4	4	0.00%
P28	4	4	0.00%	P8	4	4	0.00%	P17	4	4	0.00%
P29	4	4	0.00%	P9	3	2	−50.00%	P18	4	4	0.00%
P30	3	2	−50.00%	P24	4	4	0.00%	P20	5	4	−25.00%
AVERAGE			−11.11%	AVERAGE			8.70%	AVERAGE			−1.11%

Average of All 7 Scores

	CG				EG1				EG2		
	First	Last	Increase		First	Last	Increase		First	Last	Increase
P21	3.375	3.375	0.00%	P1	3.5	4.375	20.00%	P11	2.625	3.625	27.59%
P22	2.625	2.5	−5.00%	P2	3.25	3.25	0.00%	P12	3.875	4.875	20.51%
P23	3	3.375	11.11%	P3	3.5	3.5	0.00%	P13	3.75	4.625	18.92%
P25	2.75	3	8.33%	P4	3.625	3.75	3.33%	P14	4	4.5	11.11%
P26	2.375	3.5	32.14%	P5	2.875	3.25	11.54%	P15	4	4	0.00%
P27	3.75	3.125	−20.00%	P7	3.75	4.5	16.67%	P16	3.25	3.75	13.33%
P28	3.25	2.875	−13.04%	P8	3.25	3.25	0.00%	P17	3.625	3.75	3.33%
P29	2.875	3.5	17.86%	P9	2.875	3.25	11.54%	P18	3	3.25	7.69%
P30	3	2	−50.00%	P24	3	3.625	17.24%	P20	4.25	4.75	10.53%
AVERAGE			−2.07%	AVERAGE			8.92%	AVERAGE			12.56%