

**Web-Based Comparison of Two Styles of Auditory Presentation: All TTS versus
Rapidly Mixed TTS and Recordings**

TR 29.3684
September 26, 2003

James R. Lewis
Patrick M. Commarford
Cheryl A. Kotan

IBM User Centered Design
Pervasive Computing Division
Boca Raton, Florida

Abstract

A current controversy in the interactive voice response (IVR) community is whether and under which conditions designers should use recorded audio when portions of the interface must be generated by text-to-speech (TTS). The purpose of this study was to examine user preferences for a very extreme case—a prompt that incorporates multiple units of dynamic information in a single sentence. Two groups of IBM employees listened to and compared two auditory styles of information presentation (all information given by a single TTS voice and alternating recorded audio and the TTS voice.) The groups listened to both presentation styles in counterbalanced order and then indicated their preference and degree of preference. The number of respondents indicating a preference for the all TTS style was significantly greater than the number indicating a preference for the mixture of recorded and TTS.

ITIRC Keywords

Artificial speech

Interactive voice response

IVR

Synthetic speech

Text-to-speech

TTS

Contents

INTRODUCTION1

 PARTICIPANTS.....3

 STIMULL.....3

 PROCEDURE.....3

RESULTS5

 PREFERENCE RATINGS5

 STRENGTH OF PREFERENCE5

 PARTICIPANT COMMENTS.....5

DISCUSSION.....7

REFERENCES9

APPENDIX A. FULL SURVEY 11

APPENDIX B. COMMENT DATA..... 13

Introduction

There is some debate in the interactive voice response (IVR) community as to when designers should and when they should not mix recorded audio and text-to-speech (TTS). Researchers (e.g., Schumacher, Hardzinski, & Swartz, 1995; Balentine & Morgan, 2001) recommend using a professional voice talent when possible and generally recommend against switching between professional voices, unless there is a clear and purposeful reason for doing so (e.g., signaling different modes; a help tutor; presentation in different languages). This, of course, suggests that designers also should not switch between a professional voice and a TTS voice without a clear purpose and that a professional voice is preferred to TTS. However, in many situations, it is either very expensive or altogether infeasible to play all information by recorded audio files. For example, an IVR that reads users their email messages must read these messages in a TTS voice. Also, many applications have the need to play a wide range of dynamic data, including dates, digit strings, and times. It can be extremely expensive and require a great deal of effort to record and appropriately concatenate audio files to play whenever this information is required. Therefore, it is often necessary to play some information in TTS. In these circumstances designers must decide when recorded audio should also play (or whether recorded audio should play at all).

When, how, where, and to what degree designers should mix recorded audio and TTS is a topic that has received surprisingly little investigation. Gong, Nass, Simard, & Takhteyev (2001) presented participants with a set of seven experimental sentences, for which one group listened to in all TTS and the other group listened to in mixed output (recorded audio and TTS). In the mixed condition, all sentences began with recorded audio and ended with TTS. Participants rated the all TTS version significantly more positively in terms of the three dependent measures (liking, trust, and perceived competence). Gong & Lai (2001) conducted a similar investigation, using a Wizard of Oz methodology and a voice access to email application. Participants completed eight tasks using this system and, depending on the condition, were exposed to all TTS or a mixture of TTS and recorded audio. The authors explained that, in the mixed condition, the dynamic content (e.g., the email header and body) played in TTS. However, it is unclear whether mixing occurred within sentences or only between sentences. Independent raters judged the all TTS group to outperform the mixed group on the task scenarios, however, the all TTS group had to listen to messages and calendar listings significantly more times to complete the tasks. Further, participants in the mixed condition judged their own performance to be better and judged the system to be easier to use than those in the TTS condition. In sum, the empirical research findings still are not able to provide clear guidance as to when and how designers should mix recorded audio and TTS.

The practice of mixing recorded audio and TTS between sentences in a single application is widely accepted and, in fact, necessary if one wishes to use recorded audio at all in an application that mandates some TTS. In other words, if designers were to never mix, then all applications incorporating unpredictable information would necessarily play in all TTS. Therefore, designers often use TTS “when they have to.” For example, consider the following from a voice access to email application:

You have one new urgent message. Hi Joe, please make sure to get me that TPS report as soon as possible.

There is a great deal of agreement in the IVR community that, although the italicized text above must be played in TTS, the initial sentence (normal font) should be recorded. However, opinions differ when IVR practitioners consider whether it is best to keep the mixing at the sentence level or to play recorded audio for all stable data. For example, time and budget may mandate that the following sentences use TTS for the italicized font:

1. We have accepted your payment and have charged your credit card a total of *fifty-two dollars and thirty-four cents*.
2. Select: *Office Space*, *Raising Arizona*, or *Rat Race*.
3. You are now confirmed on Flight *1179* on *August 29th*, departing from *Miami* at *3:40 PM* and arriving in *New York* at *7:00 PM*.

The present study is a first step toward understanding user preferences when the use of TTS is necessary in an interface. We examine user preferences for a very extreme case, similar to sample 3 above, in which the rendering of the information fluctuates rapidly between static and dynamic information. If users prefer to listen to extreme cases like this in mixed format, we can conclude that, as a best practice, designers should play recorded audio for all static information. However, if the all TTS presentation is preferred to mixing for this situation, the results will necessitate further investigation to determine under which conditions, if any, recorded audio should play when TTS is necessary in the interface.

Method

Participants

We invited a total of 600 IBM¹ employees to participate. We sent 300 invitations directing participants to a website that instructed them to listen to the mixed sample first, then the all TTS sample. The other 300 invitations directed participants to a site that instructed them to listen to the all TTS sample first, then the mixed sample. A total of 72 participants, 36 from each group, listened to both samples and responded to the survey questions that followed.

Stimuli

We created two audio files, each speaking the same detailed breakdown of rates, fees, and taxes for a fictional car rental. We counterbalanced the order of presentation to control for order effects. Each audio sample played the following text at a presentation rate of approximately 100 words per minute:

Ok, for a *mid-sized non-smoking car*, the rate is guaranteed *for one week* at \$39.99 per day minus ten percent through AAA, which lowers the rate *to \$35.99* per day, with no drop charge. There is a 7 percent tax and a *\$1.95 per day* vehicle licensing fee, for an approximate daily total of \$39, with unlimited mileage.

We used the most recent version of the IBM concatenative male voice (IBM WebSphere² Voice Server for Multiplatforms V3.1.1 Third-Generation Concatenative Text-to-Speech) to produce the entire sample for the all TTS condition. For the mixed condition, the TTS voice generated the dynamic information (indicated by italicized text) and the remaining information consisted of recorded human audio from a male speaker. (The speaker was not a professional voice talent, but his voice was generally perceived as clear and pleasant.)

Procedure

We sent email messages inviting employees to participate in the study and directing them to a web page with instructions, links to both samples, and a set of survey questions. After accessing the web page, participants clicked on the first link, allowing them to listen to the first sample file. The file played on whichever audio player application the participant's desktop computer system had set as the default. Participants repeated this process for the second sample, and then answered the survey questions.

After listening to the samples, participants indicated which sample they preferred and the strength of their preference on a 7-point scale. Participants then provided comments about what they specifically liked or disliked about each sample (see Appendix A for the full survey).

¹ IBM is a registered trademark of International Business Machines Corp.

² WebSphere is a registered trademark of International Business Machines Corp.

Results

Preference Ratings

A total of 72 participants responded to the survey. Fifty-three participants (73.6%) indicated that they preferred the all TTS presentation style and 19 (26.4%) indicated a preference for the mixed presentation style. A one sample t -test indicated that this difference was statistically significant [$t(71) = 14.07$; $p < .0005$]. Further, an independent samples t -test indicated a significant main effect of order of presentation [$t(70) = 3.09$; $p < .0005$] such that participants were more likely to prefer the second presentation style to which they were exposed (see Table 1). Those who listened to the mixed style first were more likely to prefer all TTS (89% preferred all TTS) than those who listened to all TTS first (58% preferred all TTS). Table 1 provides the observed proportion of respondents who preferred all TTS and the proportion that preferred the second style presented, as well as the 90% binomial confidence intervals.

Table 1

Proportion who preferred all TTS and proportion who preferred the second presentation style

	Observed	90% Lower Bound	90% Upper Bound
Preferred All TTS	73.6%	64%	82%
Preferred 2nd Presentation	65.3%	55%	75%

Strength of Preference

Overall, respondents indicated a fairly high preference for one style or the other (mean preference strength = 4.85 on a 7-point scale), but there was no significant difference in the strength of preference for each style [$M_{diff} = 0.72$; $t(70) = 1.61$; $p = .11$].

Participant Comments

Sixty-nine participants commented on each audio sample and some of the participants commented on more than one aspect of the sample. In total, the mixed sample received 78 comments, and the all-TTS sample received 90 (see Appendix B).

Consistent with the preference ratings, the all-TTS sample received more than twice as many positive comments as the mixed sample. Participants gave positive comments about the consistency of using just one voice (22.2% of the total number of all-TTS comments), and the smooth flow of the information (12.2%). Survey respondents indicated that they did not like the all-TTS sample's mechanical sound (25.6%) or lack of inflection (14.4%).

The most commonly commented upon negative trait of the mixed sample was the way the disjointed transitions between the voices distracted users from the information content (30.8%). Participants also complained about the large contrast between the sound of the recorded and TTS voices (18.0%) and the mechanical sound of the sample (9.0%). Positive comments about the mixed presentation style included the pleasant sound of the recorded voice (9.0%) and the way the switch from one voice to the other distinguished the prices from the other information (7.7%). Interestingly, these participants felt that the switch from recorded to TTS called attention to the content, while four times as many claimed the switch distracted them from the content (see above).

The order of presentation also appeared to affect the comments. Participants who listened to the all-TTS sample and then heard the mixed sample focused more on the mixed sample's information emphasis quality and did not comment as much on the mixed sample's disjointed transitions or the TTS sample's smooth flow. Participants who heard the mixed sample and then listened to the all-TTS sample seemed to perceive the transitions between recorded and TTS voices in the mixed sample as distracting from the content, rather than emphasizing it. They remarked on the smooth flow of the all-TTS sample and complemented the recorded portion of the mixed sample more than those participants who heard the all-TTS sample first (see Appendix B).

Regardless of the order of presentation, participants disliked the degree of contrast between the voices in the mixed sample and liked the consistency of the all-TTS sample. They did not like the lack of inflection in the TTS voice or the mechanical sound of both the mixed and all-TTS samples.

Discussion

Despite the influence of an order-of-presentation effect, the overall listener preference in this study favored an all-TTS presentation over the rapid mixing of TTS and recorded speech. The preferred presentation format did not affect participants' rated strength of preference, which was relatively high. This indicates that presenting the type of information investigated in this study with an all-TTS format will be generally better than using a mixed presentation, but will be annoying to a substantial minority.

Participants who preferred the all-TTS presentation cited the consistency of using one voice and the smooth flow of the information. Participants who preferred the mixed presentation attributed their preference to the pleasant sound of the recorded voice and the way the voice-switching helped them to distinguish the pieces of key information in the message.

Due to the significant preference observed in this study for an all-TTS presentation of this type of information, we recommend using all TTS when the alternative would require rapid switching between recorded and TTS voices in a single message.

We've placed a number of constraints on the above recommendation because user preference for styles of auditory presentation of messages might depend on a number of factors that have not yet undergone systematic investigation, such as the context in which the message appears, the rate of switching between TTS and recorded voice, and the length of a recorded portion that precedes a TTS portion. These factors suggest fertile ground for future studies. For example, would the results of the current study change if participants had heard the message embedded in a context of all recorded prompts and messages?

Also, consider the following messages (*italicized text indicates variable information that must be presented using TTS*):

Select: *x, y, or z.*

Your rate for this vehicle will be *\$34.97.*

For the first example, given that the lead-in to the menu consists of only one word ("Select"), would users prefer to hear that word as recorded speech or in the same TTS voice as the menu items? In the second example, the lead-in to the TTS portion is longer. In this case, would users prefer to hear the lead-in phrase as recorded speech or TTS? We hope to address these questions in future research.

References

Balentine, B., & Morgan, D. P. (2001). *How to build a speech recognition application: A style guide for telephony dialogs* (2nd ed.). San Ramon, CA: Enterprise Integration Group.

Gong, L., & Lai, J. (2001). Shall we mix synthetic speech and human speech? Impact on users' performance, perception, and attitude. In *Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI '01* (pp. 158-165). New York, NY: ACM.

Gong, L., Nass, C., Simard, C. & Takhteyev, Y. (2001). When non-human is better than semi-human: Consistency in speech interfaces. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability Evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality* (pp. 390-394). Mahwah, NJ: Erlbaum.

Schumacher, R. M., Jr., Hardzinski, M. L., & Schwartz, A. L. (1993). Increasing the usability of interactive voice response systems: Research and guidelines for phone-based interfaces. *Human Factors*, 37(2), 251-264.

Appendix A. Full Survey

Auditory Presentation of Rental Rate Information

We are investigating different ways to present auditory information to users in interactive voice response systems. Please click the links below and listen to the two audio files (Sample 1 and Sample 2), then answer the questions:

[Play Sample 1](#)

[Play Sample 2](#)

1) Which sample did you prefer?

- Sample 1
- Sample 2

2) Please indicate the strength of your preference.

- 1 : Very weak -- the samples are about the same
- 2
- 3
- 4
- 5
- 6
- 7: Very strong -- I really prefer the one I selected

3) What did you like/dislike about Sample 1?

4) What did you like/dislike about Sample 2?

Appendix B. Comment Data

Mixed Sample

	Mixed Sample	Number of Comments	% of Total Comments
Positive Comments			
	Info Emphasis	6	7.69%
	Recorded	7	8.97%
	Request for All Recorded	5	6.41%
	Natural	4	5.13%
	Total Positive	22	28.21%
Negative Comments			
	Disjointed Transitions	24	30.77%
	Contrast in Voices	14	17.95%
	Mechanical	7	8.97%
	Multiple Voices	3	3.85%
	Not Natural	2	2.56%
	Inconsistent	2	2.56%
	Total Negative	52	66.67%
Feeling-Based Comments			
	Not Professional	1	1.28%
	Less Personal	1	1.28%
	Softer	1	1.28%
	Longer	1	1.28%
	Total Feeling	4	5.13%
	Total Comments	78	100.00%
	Total Participants Who Commented	69	
	No Comment	4	

All-TTS Sample

	All-TTS Sample	Number of Comments	% of Total Comments
Positive Comments			
	Consistency	20	22.22%
	Smooth Flow	11	12.22%
	Natural	3	3.33%
	Understandable	3	3.33%
	Less Mechanical	1	1.11%
	Human	1	1.11%
	Request for All Recorded	1	1.11%
	Shorter	1	1.11%
	Total Positive	41	45.56%
Negative Comments			
	Mechanical	23	25.56%
	No Inflection	13	14.44%
	Poor Annunciation	3	3.33%
	Low Info Emphasis	2	2.22%
	Not Natural	1	1.11%
	Total Negative	42	46.67%
Feeling-Based Comments			
	Slower	3	3.33%
	Longer	1	1.11%
	Harsh	1	1.11%
	Uncaring/Unfeeling	1	1.11%
	"Carlos sounding"	1	1.11%
	Total Feeling	7	7.78%
	Total Comments	90	100.00%
	Total Participants Who Commented	69	
	No Comment	4	

Order Effects

Mixed Comments

	Mixed 1st	Mixed 2nd	
Positive Comment	Occurrences	Occurrences	Total
Info Emphasis	1	5	6
Recorded	5	2	7
Request for All Recorded	2	3	5
Natural	0	1	1
Negative Comment	Occurrences	Occurrences	Total
Disjointed Transitions	19	5	24
Contrast in Voices	8	6	14
Mechanical	4	3	7
Multiple Voices	3	0	3
Not Natural	2	0	2
Inconsistent	0	2	2

All-TTS Comments

	All-TTS 1st	All-TTS 2nd	
Positive Comment	Occurrences	Occurrences	Total
Consistency	9	11	20
Smooth Flow	3	8	11
Natural	2	1	3
Understandable	1	2	3
Less Mechanical	0	1	1
Human	0	1	1
Request for All Recorded	0	1	1
Shorter	0	1	1
Negative Comment	Occurrences	Occurrences	Total
Mechanical	11	12	23
No Inflection	8	5	13
Poor Annunciation	0	3	3
Low Info Emphasis	1	1	2
Not Natural	0	1	1
Slower	1	2	3