

Should Recorded and Synthesized Speech be Mixed?

Cliff Nass, Caroline Simard
Department of Communication
Stanford University
Stanford, CA

{nass,csimard}@stanford.edu

Yuri Takhteyev
Department of Computer Science
Stanford University
Stanford, CA

yuri@cs.stanford.edu

ABSTRACT

Information from voice systems often is structured with a highly-variable utterance concatenated to a fixed utterance (e.g., “Today’s date is ... June 3” or “The number one book for sale this week is ... *Harry Potter*.” When the variable segments cannot be recorded, there are two design choices: recorded speech (fixed) combined with synthesized speech (variable) or consistent synthesized speech. Experiments involving two different domains (housing information system and joke system) demonstrate that consistent synthesized speech is superior to mixed speech with respect to liking, perceived competence, and credibility. Pure recorded speech is not a significant improvement in these domains. Implications for design of voice systems are presented.

Keywords

Text-to-speech (TTS), voice systems, consistency, computers are social actors, IVRs, voice user interface.

INTRODUCTION

In recent years, speech-based technology has become increasingly popular. There is universal agreement that well-cast recorded speech is the optimal choice for presentation of information. Synthesized voices, also referred to as text-to-speech (TTS), are clearly inferior in clarity, prosody, and naturalness.

Unfortunately, recorded speech is often not practical for large databases. E-commerce sites boast of “millions of items,” navigation sites present hundreds of thousands of street names, email systems must present millions of names, and search engines scan billions of sites. The structure of information in these and many other cases involves a fixed utterance, such as “The most popular book is ...” or “Email message from ...”, followed by an item drawn from a database, such as “*Harry Potter and the Goblets of Fire*” or “Jane Jones.” This structure presents designers with two reasonable choices, given that universal recorded speech is impossible: a) present repetitive and fixed information in recorded speech and variable and unpredictable information in TTS, or b) present the entire sentence in TTS.

The mixing of recorded speech and TTS is a common choice in industry. The advantages of this approach seem clear and intuitive. First, designers maximize the presentation of the most desirable modality. Furthermore, the fundamental structure of these sentences provide a natural breaking point: One doesn’t have to be a linguist to

know that there is a qualitative difference between “the time is” and “your account balance is” versus “9:34 and 23.64 seconds” and “\$2,742.66.” Another advantage of this approach is that it provides a natural transition to pure recorded speech when parts of the database become more stable and hence recordable. Finally, the use of “whenever possible” recorded speech implicitly communicates that the system is providing the best available technology at each point in the interaction.

The arguments for the second option, pure synthesized speech, come from the psychological literature on novelty and consistency. As far as novelty, abrupt change, as in the shift from recorded speech to TTS during a sentence, can lead to auditory orienting responses [6]. These responses can inhibit processing of the content, especially in the half-sentence TTS utterances produced by typical systems, because shifts are arousing [6, 7]. Furthermore, it is possible that by the time users accommodate to the shift, the recorded voice is once again speaking.

In addition to these cognitive concerns, there are also social arguments against mixing recorded and synthesized speech. In normal interactions, voices do not change in mid-sentence, nor is it usual for one person to finish another person’s sentences (despite the jokes about married couples!). If users bring to bear the same social rules and heuristics when interacting with voice systems as they do in human-human interaction [5, 7], this shift would feel disturbing and inexplicable. Research in human-computer interaction has demonstrated that systems that exhibit inconsistencies are perceived much more negatively than systems that are consistent [2, 3, 4]. In general, consistency is a very strong human desire [1].

To determine whether there are differences between the mixed recorded/TTS system and the pure TTS system, we ran two experiments using different telephone-based interactive systems. The first experiment was performed in the context of a housing information system, a context which was relevant and familiar to our experimental participants and typical of the information-provision systems (e.g., stock quotes, sports scores, directions) that are the dominant contexts of telephone-based voice systems. In this experiment, each sentence was divided into a fixed half and a variable half. The dependent

variables addressed liking of the system, trust in the system, and perceived competence of the system, the three key determinants for choosing one system over another.

To replicate our results in a very different context, the second study presented a system that read jokes in the classic question/punch-line format; the question was viewed as the fixed part of the utterance, and the punch-line was viewed as the varying part of the utterance. Because trust and competence are not relevant in this context, we only examined liking.

Our primary focus in these studies was the comparison between the fixed part of the sentence in recorded speech and the varying part in TTS vs. a uniform TTS presentation, as these are the two options that are most commonly available to systems designers. Because some systems may be able to have all recorded speech or synthesized speech may eventually be good enough to be indistinguishable from recorded speech, we also ran the pure recorded speech condition. (We did not examine the absurd condition in which synthesized speech reads the fixed part of the utterance and recorded speech reads the varying part).

STUDY 1

Each participant interacted with a university campus housing information system for approximately two minutes (seven sentences). Participants filled out a web-based questionnaire after the session.

Method

Participants

University students ($N=36$) enrolled in a communication course participated in the study. To ensure that participants would not have trouble understanding the text-to-speech output, all participants had English as a first language. Participants were randomly assigned to condition, with gender balanced across conditions. All participants were debriefed after their participation and received class credit for their participation.

Procedure

The experiment was a between-participants design. There were three conditions: (1) recorded speech for the first part of the utterance and synthesized speech for the second half; (2) synthesized speech for both halves of the utterance; and (3) recorded speech for both halves of the utterance.

Participants were given a password, via e-mail, corresponding to the condition they were assigned to. The e-mail directed them to a website which contained instructions and the experiment questionnaire.

Participants called the information system from their home telephones, providing external validity as compared to a laboratory-based experiment. The system provided campus housing information, a topic that was particularly relevant to the participants. The topic was also chosen so that it would logically involve the two-part sentence structure (fixed information followed by varying information) that is common in mixed recorded speech/TTS systems. (The script for the housing system is presented in Appendix A).

Participants called the system using their telephone and were instructed by the voice system to enter the password that was given over e-mail on the touchtone interface of the telephone; the password determined the experimental condition of the participant. After listening to the information, participants were instructed to hang up and fill out the web-based questionnaire. Once participants completed the questionnaire, the web site instructed participants to submit their information by hitting a "Submit" button. (The website can be found at <http://www.stanford.edu/~nass/comm369>).

Manipulation

The CSLU Toolkit¹ running on an NT machine was used to run the experiment; a Dialogics board answered the phone calls. For TTS, we used the default male voice of the Toolkit. For the recorded voice, we used a male graduate student whose voice was similar to the TTS voice in pitch.

For a given participant, all sentences had the same pattern of TTS and recorded voices, depending on the condition. Each sentence presented fixed information in the first half of the sentence (e.g., "results of the housing lottery will be announced on") and varying information for the second half of the sentence (e.g., "Saturday, May 20th, 2000").

Measures

Dependent measures were based on items on the web-based questionnaire. The questionnaire asked, "For each word below, please indicate how well it describes the information system you just heard," followed by a list of adjectives. Each adjective was associated with a ten-point, radio button Likert scale anchored by "Describes Very Poorly" and "Describes Very Well."

Responses were combined to create factor scores measuring liking, trust, and perceived competence, three important and distinct aspects of any information system. Factor analysis confirmed that the factor scores were distinct and internally consistent (as determined by eigenvalues and factor loadings).

¹ The CSLU Toolkit is downloadable without cost at <http://cslu.cse.ogi.edu>.

Liking was a factor score of five items: enjoyable, entertaining, friendly, good, and likable ($\bar{x} = 2.9$; factor loadings ranged from .81 to .95).

Trust was a factor score comprised of three items: realistic, reliable, and trustworthy ($\bar{x} = 2.3$; factor loadings ranged from .83 to .93).

Perceived competence of the system was a factor score comprised of four items: clever, informative, useful, and well-designed ($\bar{x} = 2.6$; factor loadings ranged from .82 to .96).

The results were analyzed using two-tailed t-tests.

RESULTS

Consistent with the idea that users prefer consistent TTS to mixed recorded and TTS (see Figure 1), users liked the housing information system using consistent TTS more than the information system mixing recorded voice and TTS ($t(22) = 2.20, p < .05$; see Figure 1). There was no difference in liking between pure TTS and pure recorded speech ($t(22) = 0.34, p > .7$), but pure recorded speech tended to be liked more than the mixed format ($t(22) = -2.01, p < .06$).

Similarly (see Figure 2), users trusted the housing information system using the consistent TTS more than the information system mixing recorded voice and TTS ($t(22) = 2.3, p < .05$). There was no difference between the two pure systems ($t(22) = .33, p > .743$), but the pure recorded system was superior to the mixed system ($t(22) = 2.58, p < .02$).

As final confirmation of the superiority of a consistent voice to inconsistency (see Figure 3), the pure TTS system was perceived as more competent than the mixed system ($t(22) = 3.04, p < .01$). Once again, pure recorded speech was not different than pure TTS ($t(22) = .85, p > .4$), but it was perceived as more competent than the mixed system ($t(22) = 2.16, p < .04$).

Figure 2: Effects of Modality on Trust of the Housing Information System

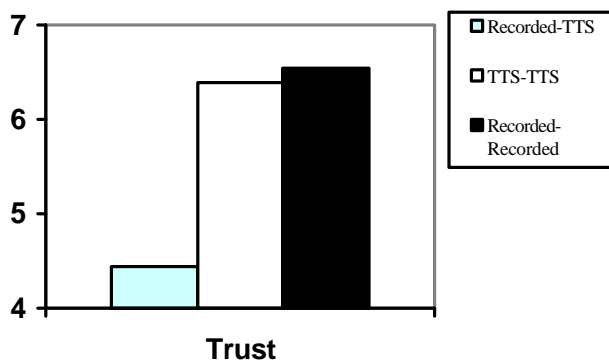
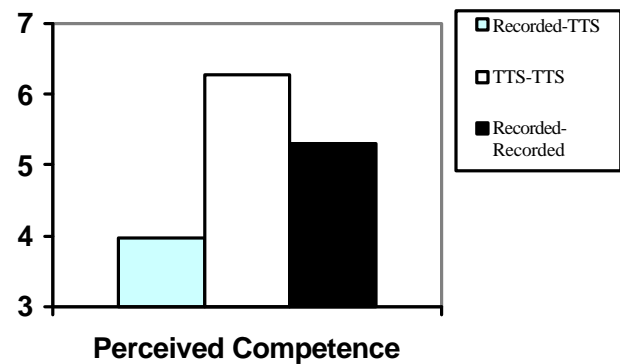


Figure 3: Effects of Modality on Perceived Competence of the Housing Information System



In summary, then, both pure systems were superior to the system with recorded speech for the fixed part of the sentence and synthesized speech for the second part.

STUDY 2

Each participant interacted with a joke telling system for approximately two minutes (five jokes). Participants filled out a web-based questionnaire after the session.

Method

Participants

University students ($N = 36$) enrolled in a communication course participated in the study. To ensure that participants would not have trouble understanding the text-to-speech output, all participants had English as a first language. Participants were randomly assigned to condition, with gender balanced across conditions. All participants were debriefed after their participation and received class credit for their participation.

Procedure

The experiment was a between-participants design: (1) recorded speech for the question and synthesized speech for the punch-line; (2) synthesized speech for the question and punch-line; and (3) recorded speech for the question and punch-line.

The study had the same format as Study 1, including a call from the participant's home and a web-based questionnaire. We chose jokes as a topic radically different from traditional information systems, but a popular form of phone-based information service. (The jokes are presented in Appendix B).

Manipulation

The CSLU Toolkit running on an NT machine was used to run the experiment; a Dialogics board answered the phone calls. For TTS, we used the default male voice of the Toolkit. For the recorded voice, we used a male graduate student whose voice was similar to the TTS voice in pitch.

For a given participant, all jokes had the same pattern of TTS and recorded voices, depending on the condition (e.g., “Did you hear about the restaurant on the moon?” “Great food, no atmosphere”).

Measures

The dependent measure was based on items on the web-based questionnaire. The questionnaire asked, “For each word below, please indicate how well it describes the information system you just heard”, followed by a list of adjectives. A separate set of questions asked about the jokes themselves. Each adjective was associated with a ten-point, radio button Likert scale anchored by “Describes Very Poorly” and “Describes Very Well.”

Responses were combined to create a factor score measuring liking, which is the key criterion for assessing an entertainment system. Although we attempted to distinguish between liking of the system and liking of the jokes, factor analysis indicated that the participants failed to distinguish between these two aspects of the system. This is not overly surprising, in that all the system did was to present jokes.

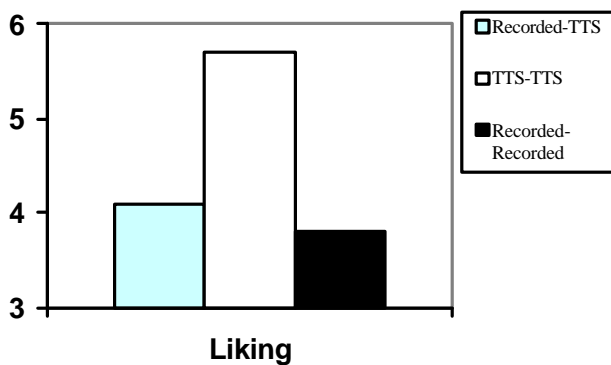
Liking was a factor score of eight items: bad joke (reverse-coded), funny joke, enjoyable system, entertaining system, good system, interesting system, likable system, and well-designed system. ($\bar{x} = 6.38$)

The results were analyzed using two-tailed t-tests.

RESULTS

As with the housing information system (see Figure 4), users liked the joke system that used pure TTS more than the mixed system ($t(22) = 2.13, p < .05$). In this case, the all-recorded system was liked less than the all-TTS condition ($t(22) = 2.51, p < .02$), while there was no difference between the recorded speech and mixed conditions ($t(22) = .22, p > .80$).

Figure 4: Effects of Modality on Liking of the Joke System



DISCUSSION

The vast majority of voice-based systems that cannot exclusively present recorded speech for each sentence use recorded speech for the fixed part of the sentence and employ TTS when they must retrieve data from large databases. The present research suggests that this strategy may be ineffective. Users’ desire for consistency seems to lead to greater liking, trust, and perceived competence as compared to the traditional approach. The results suggest that “doing the best one can at each point in the sentence” is actually inferior to providing a consistent user experience.

Of course, mixing of recorded speech and TTS is only one form of mixing. For example, a single voice can record information in both a personality-filled and emotion-filled voice and a flat, formal voice. The latter style is frequently adopted when a voice talent is reading a large database of detailed information, such as street names, dates, or numbers. The present results suggest it would be more desirable to have the entire sentence read in the flat voice rather than juxtaposing the rich and flat content.

A similar question emerges when one has different sets of data read in different voices, which can occur when multiple datasets are interspersed. This research suggests that it is worse to juxtapose two different recorded voices than to consistently hear synthesized speech. Of course, future research should address this prediction.

An important question is the level of granularity of the mixing. In the first study, the two modes were combined within a sentence. The second study juxtaposed related sentences, but they were part of a single speech act [8]. It is not clear whether mixing of voices is a positive or negative when the voices are separated by paragraphs or role.

In a multi-function system, different voices might be a means of effectively suggesting specialization [2]. For example, stock information could be read in one voice while weather could be read in a different voice. Similarly, it may be that certain functions (e.g., movie reviews) are better read in recorded speech, while other functions (e.g., driving directions) are better read in synthesized speech.

One limitation of this study was that all of the sentences presented generic information followed by unique information. While this is the most common method for presenting information in English, there are certainly sentences in which the order is reversed, for example, “Flight 272 is now arriving” or “June 27th is the last date to submit your application.” In these cases, a designer would normally be forced to transition from TTS to recorded speech. That is, the present study examined a *decline* from recorded speech to TTS; these sentences structures would allow an *improvement* in speaking style.

It is not clear whether this would mitigate or even reverse the present recommendation of pure modality.

The present research is the first study to examine the potential problems and opportunities in combining recorded speech and TTS. Future research should elaborate when and how to best leverage varying speech modalities.

APPENDIX A

Content of the Housing Information System (second half of sentence in italics).

Welcome to the XXXX University housing information system.

The number of housing slots available to undergraduate students this year is *two thousand seven hundred fifty-three*.

The most popular houses last year were *Bob, Xanadu, and Kindall*.

We predict that the most popular houses this year will be *Nardia, Zanadoo, and Story*.

If you are drawing preferred, your chances of getting your first choice should be at least *five percent*.

Results of the housing lottery will be announced on *Saturday, May 20th, 2000*.

Thank you for calling the Stanford University housing information system.

APPENDIX B

Content of the Joke System (Punch-line in italics)

Here are the top five jokes for today.

Fifth place. How many software engineers does it take to screw in a light bulb? *None. It's a hardware problem.*

Fourth place. Did you hear about the restaurant on the moon? *Great food, no atmosphere.*

Third place. Why are elephants wrinkled? *Have you ever tried to iron one?*

Second place. What do you get if you take an elephant into work? *Exclusive use of the elevator.*

And finally first place. Why are elephants large, gray, and wrinkled? *Because if they were small, white, and smooth they would be aspirin.*

We hope you enjoyed those jokes.

ACKNOWLEDGMENTS

The research was supported in part by NSF CARE and Challenge grants awarded to the Oregon Graduate Institute (R. Cole, PI) and a grant from the Center for the Study of Language and Information (CSLI) at Stanford University. The views expressed in this article do not necessarily represent the views of the National Science Foundation or CSLI.

REFERENCES

1. Fiske, S.T. & Taylor, S.E. (1991). *Social Cognition*. New York: McGraw-Hill.
2. Isbister, K. & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(1), 251-267.
3. Nass, C. & Gong, L. (1999). Maximized modality or constrained consistency? *Proceedings of the AVSP 99 Conference*, Santa Cruz, CA.
4. Nass, C. & Lee, K.M. (submitted). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*.
5. Nass, C. & Moon, Y. (2000). Machines and mindlessness. *Journal of Social Issues*, 56(1), 81-103.
6. Potter, R.F., Jr. (1999). *Effect of voice changes and production effects on orienting and memory for radio messages*. Unpublished doctoral dissertation, Indiana University.
7. Reeves, B. & Nass, C. (1996). *The media equation: How people treat computers, televisions, and new media like real people and places*. New York: Cambridge University Press.
8. Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. London: Cambridge University Press.