2009

Speech Synthesis for Data Link: a Study of Overall Quality and Comprehension Effort

Martine Godfroy

Durand R. Begault

Elizabeth M. Wenzel

SPEECH SYNTHESIS FOR DATA LINK: A STUDY OF OVERALL QUALITY AND COMPREHENSION EFFORT

Martine Godfroy, San Jose State University Foundation
Durand R. Begault, Human Systems Integration Division (ARC-TH)
Elizabeth M. Wenzel, Human Systems Integration Division (ARC-TH)
NASA Ames Research Center
Moffett Field, CA 94035
Contact: martine.godfroy-1@nasa.gov

This study investigated subjective preference for synthesized "spoken data link" messages to provide initial design guidance for communication displays in the context of NextGen (Next Generation Air Transport System) operations. Ratings of Overall Quality and Comprehension Effort were obtained as a function of voice type, synthesized speech rate, and sentence prosody. Rank-order data analyses showed that both Overall Quality and Comprehension Effort were affected by speech rate: under the "Fast Rate" condition (vs. "Default Rate"), Overall Quality decreases and Comprehension Effort increases. However, the introduction of "Prosodic Emphasis" (pitch and level changes for specific phrases) in Fast Rate sentences produced a relative improvement in both comprehension and quality ratings. For both speaking rates, the introduction of "Prosodic Emphasis" resulted in higher quality ratings and lower comprehension effort ratings. The data suggest that faster speaking rates, which may improve message throughput in a display, may be viable when combined with prosodic emphasis.

The overall objective of our work is to investigate human performance costs and benefits of voice communications interfaces in NextGen flight decks, with the particular goal of improving shared situational awareness and minimizing human error and workload. This study investigates perceived Overall Quality and Comprehension Effort of synthesized speech systems in an aviation context of "spoken data link" and "party line" messages. The increased operational autonomy of fight crews in the NextGen will potentially result in higher overall workload and greater dependence on the visual modality to interact with flight instrumentation and data. To alleviate these problems, future systems may be designed to advantage by using synthetic speech displays to convey, e.g., shared situation awareness regarding flight status and trajectory between cockpit to cockpit and cockpit to ground. Such displays may be combined with other design strategies such as 3-D audio presentation to achieve more intelligible, discriminable, and identifiable communications (Begault, 1999).

While human performance analyses predict clear benefits to both workload and capacity in a data link-dominated airspace system (Leiden, et al., 2003), previous studies (Smith, et al., 2001) suggest that voice communications are preferred for interactive tasks involving emergency situations, position reports, and some clearance requests. A study conducted for the Federal Aviation Administration (FAA) examined the use of pre-recorded speech versus text-based data link messages, and whether or not speech displays reduced head-down time and improved response time (Rehmann, 1996). The results of that study indicated advantages of speech for reducing head-down time, but the slow cadence of the speakers was found to be slower than reading (in that system, actual voices were recorded digitally and played from a computer.) Speech was preferred in most cases over text by pilots, and workload was reportedly decreased for confirming messages.

The primary advantage of text-to-speech synthesis is that arbitrary messages can be communicated without pre-recording. A further advantage of more advanced systems is the ability to manipulate the manner of speaking in real time; in particular to change the *prosody* of the spoken utterances to convey meaning outside of the syntactic context. For example, current speech synthesis-recognition systems, known as "interactive-voice-response systems," used in telephonic commerce are able to establish specific personas that can express characteristics like urgency or compassion that enhance semantic content. The current study used the "Rvoice" (Rhetorical Systems, Ltd) commercial text-to-speech synthesis system, which has high speech quality combined with the ability to parametrically vary a number of speech characteristics related to prosody and speaking rate. By manipulating prosody, messages can be formed to add emphasis to important words, such as call signs, and to elevate the perceived urgency of a message, such as when speaking in a "raised" versus normal voice.

A speech synthesizer can also change the rate of spoken words without the loss of intelligibility that an actual speaker would have. Brungart, Simpson, and Nandini (2007) investigated the use of "time-compressed" speech for military aviation applications; they concluded "it is often possible to accelerate a speech signal by a factor of two or more without much sacrifice in intelligibility." The advantage of increasing the spoken word rate is that information transfer can occur more quickly, increasing the amount of information that can be received per unit time.

Given the potential advantages for future flight decks of varying prosody and speaking rate in spoken data link messages, this study sought to determine if these manipulations might impact the subjective Comprehension Effort and Overall Quality of the communications. Subjective Comprehension Effort refers to the perceived effort necessary to understand a message. Sound Quality refers to features of a sound that contribute to the subjective impression made on a listener, with reference to the suitability of the sound for a particular set of design goals, and is meant particularly to account for aspects of communication systems that are not quantifiable by intelligibility measurements. It is well understood that the requirements for speech intelligibility will likely be satisfied by use of an appropriate signal-noise ratio. Recent work in the area of perceptual audio evaluation (Bech and Zacharov, 2006) has shown that subjective data can elucidate a variety of stimulus factors that contribute to high quality audio system design. Telephony research (engendered in various quality testing standards of the International Telecommunications Union-ITU) has long recognized that a particular system may produce an acceptable level of intelligibility while having poor overall quality. Our research presumes that poor audio quality may work against the system's long-term use in an everyday situation.

Experimental Method

This experiment was designed to determine preferences for a given type of synthesized voice (gender, timbre) as well as the effects of speech rate and sentence prosody on the perception of overall quality and on the degree of comprehension effort. The experiment generally followed the ITU recommendations for speech quality experiments (ITU P.800, "Methods for subjective determination of transmission quality") and for synthesized speech (ITU P.85, "A method for subjective performance assessment of the quality of speech voice output devices"). Subjective scales for evaluation per ITU P.800.1 ("Mean opinion score (MOS) terminology") were used to rate "overall quality" and "comprehension effort".

Participants

Eight undergraduate students (4 male and 4 female) from the San Jose State University, CA, aged 19 to 45, took part in the experiment. The participants were all volunteers and were naïve regarding the content and objective of the experiment. All reported that they had normal hearing.

Experimental design

A stimulus corpus of sixty different sentences was composed based on a transcript of Air Traffic Control communications sent by the tower to the crew from a study previously conducted at NASA. The following are examples:

*"**Asiana 2 1 4** turn right heading to 1 3 0 and proceed direct Molen"*
*"**Air France 0 8 4** contact tower 3 2 0 point 5 good day"*
*"**Korean 0 2 3** turn left and maintain 8000 accept approach clearance in 5 nautical miles"*

The 60 sentences were randomly assigned with a voice, a rate, and a prosody manipulation. The three possible voices were American English-speaking voices that are provided as options with the standard Rvoice text-to-speech software package. Rvoice allows manipulation of a number of speech parameters using a scripting language of XML tags imbedded in the text-to-speech string.

The prosody manipulation was accomplished by using (1) the default settings for each voice or (2) increasing the pitch and loudness of the aircraft identifier phrase contained in each sentence (bold face words in the example sentences above) to provide "prosodic emphasis." Manipulation of internal Rvoice software parameters resulted in an increase in the fundamental frequency ("pitch") of the spoken voice by approximately 40 Hz and a loudness increase of approximately 4 dB, depending on the specific phrase. The result was an audible "raising" of the voice,

both in terms of pitch and level. The speech rate manipulation was accomplished by using (1) the default settings or (2) an increase in the speaking rate of the overall phrase by a factor of about 1.2. The result was that of an articulate but faster than normal speaking voice. Note that the pitch could be manipulated independently of the rate (unlike, e.g., an analog tape recorder under fast playback). Prosodic emphasis and speech rate manipulations were applied to three different synthetic voices, two male American English speakers and one female American English speaker.

The experimental variables are summarized as follows:

- VOICES: 1 female (Rvoice ID: F019), 2 males (Rvoice IDs: M002, M009)
- PROSODY: 2 prosody levels (No Prosody: NP, Prosody: P)
- RATE: 2 speed levels (Normal: N, Fast: F)

The sentences were synthesized and then digitally mixed with a binaural recording of a 747 flight deck simulator cruise phase background noise, calibrated to a level of approximately 70 dB SPL. The signal-to-noise ratio of the sentences was +10 dB RMS relative to the background noise.

Using the CRC-SEAQ (Communications Research Centre Canada System for the Evaluation of Audio Quality) subjective test module, we designed two separate sessions of 60 trials, one for evaluation of the Overall Quality and one for Comprehension Effort. Two different 5 level scales were used for each session:

-Session 1 (Overall Quality): 1 (bad) to 5 (excellent)

-Session 2 (Comprehension Effort): 1 (excellent, no effort required) to 5 (bad, high effort required)

Sample screenshots for the Overall Quality and Comprehension Effort trials are shown in Figure 1. A trial consisted of the successive presentation of 3 different sentences, noted A, B and C on the experiment screen. No reference was presented. The task of the participants was to rate the 3 different sentences one relative to the others by adjusting a cursor associated with the 5-point scale that was always visible on the left of the screen. The listeners were not allowed to assign ratings until they had listened to each sentence at least once. However, they were allowed to listen to any of the sentences as many additional times as they wished. During a particular trial, the sentences assigned to the A, B and C stimuli from the list of 60 possible sentences were always different sentences. Nevertheless, the random assignment of speaking voice, prosody level, and speaking rate to the A, B and C stimuli was such that the listeners may have heard the same voice, prosody level, or speaking rate more than once. Over the course of all 60 trials in each session, the listeners heard each voice, prosody level, or speaking rate the same total number of times. All listeners heard the same randomized list for both the Overall Quality and Comprehension Effort sessions. Prior to each session, the listeners were given written instructions and the experimenter talked with them to make sure they understood the task.

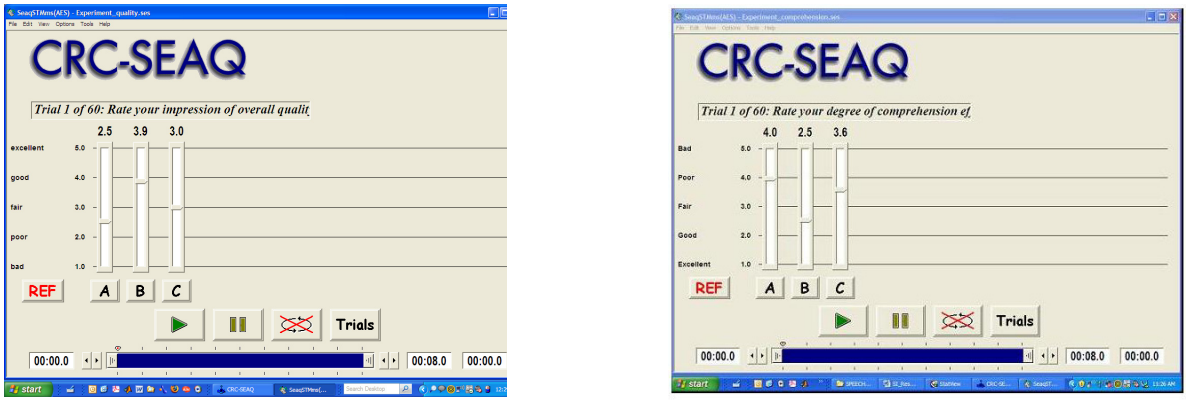

Figure 1: CRC-SEAQ interface for the Overall Quality (left) and Comprehension Effort (right) sessions. The letters A, B and C represent the three successive sentence stimuli. The score estimate is relative, i.e. each sentence stimulus is rated in relation to the two others. No reference was used in this experiment. Note that the scale for Comprehension Effort is inverted compared to the scale for Overall Quality.

Results

Descriptive analyses of the data were conducted with SPSS® to assess whether the data were normally distributed. One sample Kolmogorov-Smirnov test showed that the distributions for both Overall Quality and Comprehension Effort were normal (OQ: Z=.91, p=.37; CE: Z=.61, p=.84). The data for Overall Quality showed some asymmetry (skewness=-.43, SE=.24) and a flat distribution (kurtosis=-.47, SE=.48). Data for Comprehension Effort were closer to a normal distribution in terms of skewness (skewness=-.03, SE=.24), although the kurtosis value again indicated a flat distribution (Kurtosis =-.56, SE=.48). Because of the qualitative nature of the data, non-parametric analyses were preferred. Friedman tests were performed to determine the main effect of the condition of presentation (N=4). A posteriori Wilcoxon tests were used to further analyze the results to determine the specific effects of Speed and Prosody. The significance level was set at p=.05.

Overall Quality

A Friedman test performed on the 4 repeated-measures conditions (Non-Prosody/Normal, Non-Prosody/Fast, Prosody/Normal and Prosody/Fast) showed statistical differences between the reported ratings ($\chi^2_3 = 25.28$, p<.0001). As expected, a faster speech rate led to a decrease in perceived Overall Quality, both in the Non-Prosody and in the Prosody condition (NPN-NPF: Wilcoxon Z=-3.68, p=.0002, PN-PF: Wilcoxon Z=-2.82, p=.004). More interestingly, introduction of Prosody in the sentence did not significantly affect the perception of quality when the speech rate is normal (NPN-PN: Wilcoxon Z= -1.35, p=.17), but significantly compensated for the deleterious effect of increased speed (NPF-PF: Wilcoxon Z= -3.06, p=.002).

Table 1 summarizes the observations and provides supplementary data as a function of subject gender (SM vs. SF) and voice gender (VM vs. VF). It can be seen that the Overall Quality scores are not modified by this a posteriori categorization. The only minor difference concerns the marginal significance (p=.05) of the differences between the NPN/PN conditions for male subjects, who appear to be more sensitive than female subjects to the effect of prosody That is, male subjects produced lower ratings compared to female subjects in the Non-Prosody condition.

Table 1. *Friedman mean rank for perceived Overall Quality (scale from 1= bad to 5=excellent) as a function of the 4 conditions of presentation of the sentences*

Conditions	Total		Subject M		Subject F		Voice M		Voice F	
NPN	2.83	p=.17	2.50	p=.05	3.16	p=.87	2.93	p=.25	2.62	p=.52
PN	3.25		3.25		3.25		3.31		3.12	
NPF	1.48	p=.002	1.66	p=.03	1.29	p=.02	1.53	p=.02	1.37	p=.04
PF	2.44		2.58		2.29		2.21		2.87	
χ^2_3	p<.0001		p=.02		p=.0004		p=.0004		p=.02	

Comprehension effort

Similar to the Overall Quality evaluation, we observed a significant effect of condition on the rating of Comprehension Effort ($\chi^2_3 = 33.54$, p<.0001). An increase in speech rate increased the perceived Comprehension Effort, both under Prosody (Wilcoxon Z=-3.41, p<.001) and Non-Prosody conditions (Wilcoxon Z=-3.88, p<.0001). Once again, the role of prosody was not statistically significant under normal rate speech production (Wilcoxon Z=-1.35, p=.17), but was shown to be relevant (Comprehension Effort decreased) when sentences are produced at higher speed (Wilcoxon Z=-3.34, p<.001).

Table 2 summarizes the results for Comprehension Effort and provides supplementary data as a function of subject gender (SM vs. SF) and voice gender (VM vs. VF). Similar to the Overall Quality scores, ratings of Comprehension Effort were not modified by this a posteriori categorization. The faster speech rate always produced a greater Comprehension Effort (p at least <.01), and the introduction of prosody components in the sentence contributed to a reduction of the Comprehension Effort when the perception was altered by an accelerated presentation.

Table 2: *Perceived Comprehension Effort (scale from 1= excellent to 5=bad) as a function of the 4 conditions of presentation of the sentences.*

Conditions	Total		Subject M		Subject F		Voice M		Voice F	
NPN	2.02	p=.17	1.87	p=.96	2.16	p=.09	2.15	p=.06	1.75	p=.77
PN	1.58		1.58		1.58		1.50		1.75	
NPF	3.56	p=.001	3.62	p=.02	3.50	p=.01	3.56	p=.006	3.56	p=.06
PF	2.83		2.91		2.75		2.78		2.93	
χ^2_3	p<.0001		p=.0002		p=.002		P<.0001		p=.007	

Overall Quality and Comprehension Effort

Figure 2 plots the Friedman mean ranks for Overall Quality and Comprehension Effort as a function of prosody and rate collapsed across subject gender and voice gender since these categories had little or no impact on the results.

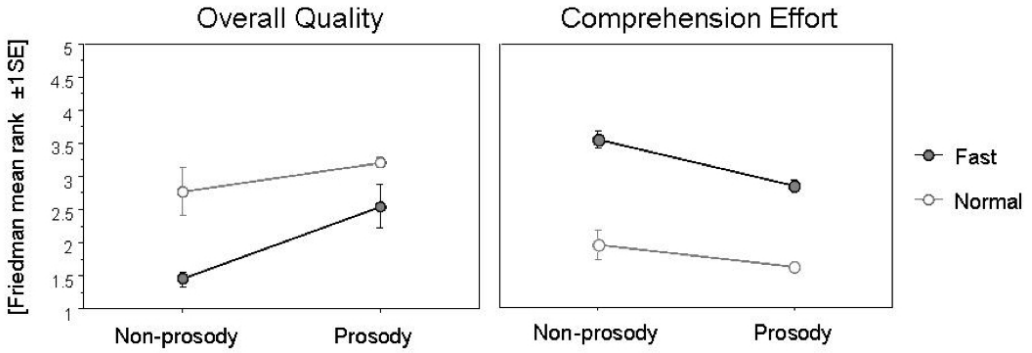

Figure 2: Left: Friedman Mean rank for Overall Quality (1 =bad to 5=excellent) as a function of prosody & rate. Right: Friedman mean rank for Comprehension Effort (1 =excellent to 5=bad) as a function of prosody & rate.

Overall, analysis of the data for Overall Quality and Comprehension Effort revealed that they were very similar (Z=-.02, p=.97). When split between the different conditions, the two measures were shown to be highly and inversely correlated as seen in Figure 3.

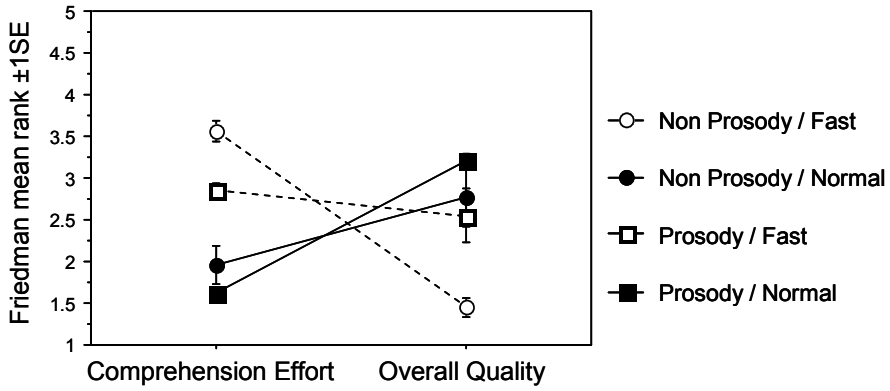

Figure 3: Overall Quality (1 =bad to 5=excellent) and Comprehension Effort (1 =excellent to 5=bad) for the synthesized voices are plotted as a function of the four conditions of presentation of the sentences (rate by prosody). Note that the inversion of the scale leads to the pattern of interaction between the two conditions.

The inverse relationship resulted from the inverse numerical order of the two rating scales. Indeed, fast presentation of the sentences was associated with the highest degree of Comprehension Effort and the lowest perception of

Overall Quality. Similarly, the addition of prosodic emphasis produced the lowest degree of Comprehension Effort combined with the highest level of Overall Quality.

Conclusions and Discussion

In general, that data showed that both the gender of the listener (OQ: Z=-.56, p=.57; CE: Z=-.31, p=.75) and the gender of the stimulus voice (OQ: Z=-.33, p=.73; CE: Z=-.16, p=.86) had no significant impact on perceived overall quality or comprehension effort. This is perhaps not surprising since previous work on audio displays (e.g., Brungart, et al, 2007) have found gender to significantly impact intelligibility in multiple talker situations as opposed to the single talker stimuli used here.

As expected, a faster speech rate led to a decrease in perceived Overall Quality and an increase in the perceived degree of Comprehension Effort, both in the Non-Prosody and the Prosody conditions. Introduction of prosodic emphasis in the messages did not significantly affect the perception of quality or comprehension effort when the speech rate was normal. While the combination of prosody and a normal speaking rate produced the highest average ratings of Overall Quality and the lowest ratings of Comprehension Effort, the data were not statistically different from ratings of the normal speaking rate without prosody. However, prosodic emphasis significantly compensated for the deleterious effect of increased speaking rate, by both increasing perceived Overall Quality and decreasing Comprehension Effort. The data indicate that the subjective acceptability of the messages when prosody was combined with the faster rate remained relatively high (above the theoretical mean). The results suggest that if faster message throughput is required in a speech display system, possible negative effects may be offset by the use of prosodic emphasis to enhance meaning in the message.

Future investigations will be concerned with the determination of ceiling effects, i.e. the determination of an acceptability threshold for speaking rate. The impact of prosody and spatial separation on Overall Quality and Comprehension Effort for synthesized speech messages in a background of competing messages will also be examined. We also plan to correlate the results of our subjective impression data with objective measures such as intelligibility.

Acknowledgements

This study was supported by the Integrated Intelligent Flight Deck project within NASA's Aviation Safety Program.

References

Bech, S. & Zacharov, N. (2006). *Perceptual Audio Evaluation: Theory, Method and Application.* West Sussex, England: John Wiley & Sons, Ltd.

Begault, D. R. (1999). Virtual acoustic displays for teleconferencing: intelligibility advantage for "telephone-grade" audio, *Journal of the Audio Engineering Society*, *47,* 824-828.

Brungart, D. S., Simpson, B. D., & Iyer, Nandini (2007). Maximizing information transfer in auditory speech displays. In *Proceedings of the 19th International Congress on Acoustics.* Madrid, Spain.

CRC SEAQ: Communications Research Centre of Canada System for the Evaluation of Audio Quality. (http://www.crc.ca/en/html/aas/home/products/products#CRC_SEAQ)

Leiden, K., Kopardekar, P. & Green, S. (2003). Controller workload analysis methodology to predict increases in airspace capacity. *AIAA Aviation Technology, Integration, and Operations Forum,* Reston, VA.

Rehmann, A. J. (1996). Airborne Data Link Study Report. DOT/FAA/CT-TN95/62, Springfield, VA: National Technical Information Service.

Smith, N., Brown, J. A., Polson, P. and Moses, J. (2001) "An assessment of flight crew experiences with FANS-1 controller-pilot data link communication in the South Pacific." *4th USA/Europe Air Traffic Management R&D Seminar,* Santa Fe, NM.