Advanced Technology Applications to Flight Simulation Training and Evaluation: Voice Generation and Recognition

Alfred T. Lee, Ph.D., CPE

BRI-TR-121099

October, 1999

BETA RESEARCH INC

18379 Main Blvd. Los Gatos, CA 95033 Tel: (408) 353-2665 Fax: (408) 353-6725 www.beta-research.com

Advanced Technology Applications in Flight Simulation Training and Evaluation: Voice Generation and Recognition

Abstract

Technical advancements in computer and visual display technology over the past two decades have engendered major advances in simulation technology development. Flight simulation technology has achieved sufficient levels of fidelity to allow the simulator to become an effective surrogate for the real aircraft. In some aspects of flight simulation, achievements been much less dramatic. Simulating voice communications between a trainee pilot and air traffic controllers, other crew members and ground personnel, and even the normal radio frequency traffic has proved much more difficult. As these communications task are an important aspect of training and because the fidelity of simulation may be compromised by their absence, an investigation was conducted to investigate the state of voice technology and the applications of this technology to the simulation of voice communications in flight training. This report describes the history and status of voice generation and recognition and its actual and potential application to flight simulation technology. Examples of voice technology applications in military and civilian simulators are provided to illustrate the practical utility and limits of the technology. Recommendations for use of the technology for training and evaluation are provided. The importance of identifying the training and communication task objectives and user population are as necessary steps in the process of developing and applying voice generation and recognition technology in flight simulation.

Advanced Technology Applications in Simulation Training and Evaluation: Voice Generation and Recognition

Considerable technical advancements in simulation training and evaluation technology have occurred over the past two decades in aviation, medicine, and other domains. This is particularly true in the areas of visual display and scene generation, sound generation, and motion platform technologies. Simulation or "virtual reality" technology is dependent upon enabling technologies in computer processing, memory, and software technologies which have made extraordinary advances. The advancement of enabling technologies has reached the point where some simulators, particular flight simulators, can produce training and evaluation scenarios virtually indistinguishable from their real-world counterparts. More advanced, commercial and military flight simulators are now routinely used as surrogates for the real aircraft in both pilot training and evaluation. However, this reliance on simulation technology as the primary pilot training system means that much of civil and military aviation is now heavily dependent on this technology to support its training and programs. This fact has, in turn, forced the flight simulation technology industry to provide even greater levels of physical and perceived fidelity between the virtual and the real world. This paper will examine an aspect of flight simulation technology which has been overlooked in the past, but is now being revisited as an area that needs to be addressed to support an increasingly demanding flight training and evaluation environment.

The simulation of voice communications is among those areas that flight training technology has yet to adequately address. In aircraft operations, voice communications occur between the pilot and ground-based air traffic controllers, between the pilot and other flight and cabin crew members, and between the pilot and ground maintenance and dispatch personnel. Additionally, a radio frequency may be monitored by the pilot for communications between air traffic control (ATC) and other aircraft on the same radio frequency. This frequency chatter or "partyline" is often monitored by pilots to help in tactical design-making. Voice communications in the aviation environment, particularly in air-ground communications, are notable for their brevity and speed due to the demands of flight operations and the limited capabilities of existing air-ground communications technology. The air-ground communications content is a highly abbreviated, operational language with a pace some 50% to 100% faster than normal conversational speech (Morrow, Lee, and Rodvold, 1993). Adding to the challenge is the quality of the communications channel. Air-ground propagation and other effects of VHF and HF radio communications will distort the quality of voice communications between the pilot and ground facilities as well as introduce transmission delays. These distortions and delays need to be replicated in any simulation if radio communications realism is to be achieved.

Most challenging of all the voice communication simulation requirements is the interactive nature of communications. To provide fully automated voice communications simulation, the simulator needs to be able to recognize and respond to pilot voice communications at any time during the simulation. The simulation needs to comprehend

pilot statements and inquiries and respond appropriately and quickly. This level of voice communications simulation is the ultimate goal of technologists in this area. Before this level is attained, however, more mundane voice technology challenges remain such as the intelligibility and general quality of voice simulation. The following sections describe the state of voice technology at this time and its potential for simulation applications. Conclusions and recommendations regarding the near and long term prospects of the technology for use in simulation training and evaluation follow these sections.

Voice Technology: Generation and Recognition

Several decades of research and development have been devoted to the goal of developing a high fidelity, voice generation capability which is computationally efficient, highly reliable, and which duplicates the natural prosodic characteristics of the human voice for the hundreds of extant languages and dialects. Development of this technology has had a strong commercial and military incentive for its use in personal computer applications, personal digital assistance, telecommunications, and vehicle and process control applications. The increasing demands placed on the human visual system to process visually displayed information have made the use of voice displays an attractive alternative. Additionally, voice technology is now routinely used as an assistive technology for the visually disabled.

In some instances, such as telecommunications applications, voice generation technology is the only viable means of displaying information. Telecommunications applications, specifically Interactive Voice Response (IVR) systems, have been developed to allow users to interact with a voice communications system using a touch-tone telephone. However, even before telecommunications IVRs, the military was incorporating applications of voice technology to provide alerting, warning, and aircraft state information in aircraft cockpits (Simpson, 1985). Civil aviation soon followed with voice displays such as those found in Ground Proximity Warning Systems (GPWS) and Traffic Collision and Advisory Systems (TCAS). Much of the practical knowledge of deploying voice displays originates with aviation cockpit applications. These include issues of intelligibility of the speech generated as well as user acceptance of the quality of that speech when compared to its naturally generated counterpart. Before discussing these data, however, some background is needed on voice generation technology in order to appreciate the technical complexities involved.

Digitized Speech

Voice display technology falls into two general categories, depending on the underlying voice generation principles involved. The first category, digitized speech, is not synthetic speech at all but the concatenation of large, digitized elements of human speech, such as individual words or phrases. Digitized speech is generated from digital sampling of analog human speech provided by a human donor. The quality of the resulting speech varies with the sampling rate used and the rules employed in concatenating the speech elements. At high sampling rates (e.g., at or above 8KHz), digitized speech is often indistinguishable from its analog counterpart.

Digitized speech has a number of practical short-comings that have limited its wider use. First, even at a nominal sampling rate, digitized speech requires considerable computer memory storage. This fact posed a greater obstacle early in its history than it does now, but is still a potential problem for some applications where large message sets may not always be possible due to the storage requirements of the technology. Secondly, the upgrading and maintenance of digitized speech is dependent on the availability of the original donor. Because human speech characteristics have unique qualities, modifications or additions to message sets require re-sampling of the original donor speech. If significant time has elapsed since the original recording, this may be impractical or even impossible. Thirdly, digitized speech, as with other voice technologies, introduces distortions where words or phrases are joined together. This is particularly problematic for a phenomenon termed "co-articulation". In compound words where one word is formed from two individual words, it is not uncommon to find components of one or both words which are pronounced differently or not at all as result of the union. For example, the compound word "bookcase" leaves one of the adjoining consonants unpronounced and eliminates the pause that is normally provided between the two words. Similarly, the changes in pitch of a word or syllable at the end of a sentence is different if it is used in a question as opposed to a sentence. These and many other problems of intonation induced by contextual variations in speech means that concatenation rules for digitized systems may need to be quite complex. Alternatively, message sets can be pre-recorded to accommodate wherever possible predicted speech contexts, thereby eliminating or substantially reducing the problems created by concatenation. However, such systems may involve a substantial development effort to digitized message sets. For example, to record all possible combinations of heading and airspeeds that a controller might transmit to a pilot would require at least a thousand or more unique message recordings. Yet even this solution is unlikely to meet every possible event that might transpire in the simulation. Digitized speech remains in use because the quality or naturalness of digitized speech is often so attractive as to make this technology the only alternative for many applications. For those situations where a precise replication of the human voice is needed, digitized speech systems remain superior to all existing synthetic speech systems.

Nonetheless, the problems of concatenation, legacy of donor speaker, memory storage requirements and the difficult and complex task of digitizing large word and message sets have limited the use of digitized speech. Such systems are mainly used where the message set is small, the context of the speech is highly predictable, and the need for high voice quality or naturalness is very high. Due to its ability to render high quality, natural-sounding voice characteristics, digitized speech remains the most common voice display technology for those applications where voice quality needs to be high. Most IVR systems, cockpit voice displays, and ground-based vehicle warning systems use digitized speech. However, the inherent shortcomings of digitized speech have served as a strong incentive for developing new methods of generating speech.

Synthetic Speech

Synthetic speech systems are classified based on how the constituent elements of synthesized speech are formed. In formant synthesis, as its name implies, the constituent elements of synthetic speech are formants. Formants are the resonant frequencies formed by the vocal system when speech is produced. In formant synthesis, these signals are created by modeling the vocal system that produced them. Mathematical formulae are constructed to describe the mechanics of the system in operation much the same way models are derived for the aerodynamics of airfoils, for example. By systematic variation of the models variables, it is theoretically possible to generate any speech unit (e.g., phoneme) in any language one chooses. Among the most well known of such systems is the MITalk system developed in the last decade (Allen, Hunnicutt, and Klatt, 1987). Subsequent generations of this system have been refined for various research purposes. For example, a sample model is currently being used at MIT Media Labs to develop a singing voice (Oliver, Yu, Metois, 1997). The reader can obtain more details of this model at the MIT Media Labs web site (http://www.mit.edu).

The formant approach to speech synthesis is arguably the most computationally elegant and sophisticated means to generate human speech in all of its complex variations. This computational efficiency was particularly important in the early history of computer technology when memory and processing speed were extremely limited. Despite it's computational elegance, formant synthesis has not been able to generate voice quality comparable to that possible with digitized speech. The resulting speech has a characteristic mechanical sound, albeit a generally intelligible one. In early applications, this unnatural quality was found to be objectionable to many users, particularly in consumer products. While formant synthesis can produce essentially seamless speech by smoothing the distortions caused by concatenation, it does not appear to be a viable candidate for producing high quality, synthetic speech systems.

Because of the difficulties involved in generating high quality, synthetic speech from formant analyses, alternative speech synthesis methods have been attempted. The most common method is described as concatenative synthesis (or synthesis by rule) and it dominates what is called the text-to-speech (TTS) market. The TTS systems allow speech to be generated from any text string whether the string is provided by a software application program, from a text file, or directly from an end-user. The TTS systems have the advantage over other synthesis methods in that they can rely on the grammatical rules and vocabulary of a written language. This makes conversion of the text into synthetic speech output more computationally manageable since the systems more easily predict the next component in the text string.

Underlying most of the commercially available TTS systems is their reliance on pre-recorded sound elements (e.g., diphones or triphones) from a donor speaker. The diphones are concatenated using prosodic, phonemic and other linguistic rules for the particular language to form words and sentences. As diphones already have the co-articulation effects for the particular language, co-articulation problems are eliminated. However, intonation problems can occur if a diphone is recorded from one phrase and

used in another phrase were the tonal marking is no longer appropriate. Microsoft's Whistler TTS product (Huang, et al. 1996) uses such as strategy to generate speech as does the Bell Labs (Lucent Technology) TTS product (Sproat, 1997).

A somewhat different approach is used by DecTalk, a TTS product developed by the Digital Corp (now Compaq Corp) in the early 1980's. DecTalk uses a digital formant synthesizer which takes phones as its input rather than stored formant patterns used by the MIT systems (Hallahan, 1996). Analysis of the clause structure is used to apply intonation rules and phonetic rules are used to provide co-articulation effects. Modifications to a vocal tract model are used by DecTalk to generate different voice types (e.g., male, female). The DecTalk system has become one of the more successful TTS systems on the market.

Intelligibility and Quality of Synthetic Speech

The intelligibility of synthetic speech, specifically TTS, is the measure of the degree to which individual speech segments, such as phonemes, can be recognized by the speaker of the language from which the phonemes are derived. One common measure used to measure intelligibility is the Modified Rhyme Test (MRT). MRT uses the rate of error in recognition of synthesized speech phonemes to gauge the intelligibility of a TTS relative to other synthetic systems and natural speech. Natural speech, for example, produces MRT error rates of less than one percent. Studies of the synthetic systems developed in the 1980's produced a wide range of intelligibility rates. In a test of the ten extant systems during that period, MRT rates of from 3% to 35% were found (Logan, Greene, Pisoni, 1989). Notably, the intelligibility of these synthetic speech systems improved substantially with exposure to the system (Thomas, Rosson, and Chodorow, 1984; Schwab, Nusbaum, and Pisoni, 1985).

Despite early evidence of poor intelligibility of many synthetic systems when compared to natural speech, recent studies suggest that this disparity may be minimized or eliminated by improvements to TTS systems. In a study by Paris, Gilson, and Thomas (1995), the DecTalk TTS was found to have comprehension rates comparable to natural speech[1]. This finding, however, appeared to depend on whether response latency is an important factor. This study also found that, in tasks where the speed of speech processing is important, natural speech is still superior to even the highest quality TTS systems. Additionally, a significant increase in the use of human cognitive processing resources is a recurrent finding in studies of synthetic speech system use (Delogu, Conte, Sementina, Ciro, 1998). Increased memory load and poor retention of synthetic speech compared to natural speech has also been found in previous studies of synthesized speech (Luce, Feustel, and Pisoni, 1983; Smither, 1993; Waterworth and Thomas, 1985). The poor retention of synthetic speech can, however, be eliminated if sufficient additional learning time is provided for those receiving messages in synthetic speech. (Delogu, et

[1] It is noteworthy that, in every study reviewed here where DecTalk was compared to other systems, DecTalk was rated as superior in intelligibility and comprehension.

al., 1998; Luce, Feustel, and Pisoni, 1983; McNaughton, Fallon, Tod, Weiner, 1994; Waterworth , 1983).

Synthetic speech systems that induce additional task loading beyond that of natural human speech pose a problem for training and evaluation simulations. The added workload of such systems may have the effect of increasing simulation training time. Additionally, they may elicit student adaptive strategies to handle the additional workload required of the synthetic speech system. Such task management strategies developed in response to synthetic speech in the simulator may be inappropriate when applied to real world operational situations.

Voice Recognition

In order to provide a fully automated voice communications simulation capability, voice transmissions, such as inquiries and communications readback, need to be interactive. Such a system would, as with real human communications, process the contents of a voice message from the trainee and confirm the correctness or appropriateness of the message and respond accordingly. Such a system would necessarily need to process and respond at speeds comparable to that found in real world operations.

Much of existing voice recognition technology has proceeded in step with the speech synthesis development described above. For example, the core of the Microsoft Whistler product is the same natural language processing system found in its speech recognition product ("Whisper"). This is not surprising as the analyses required to recognize speech will follow many of the same rules required to synthesize it. However, speech recognition is complicated by the fact that speech synthesis such as TTS can anticipate the next speech element (because it is derived from written language) whereas oral or vocalized speech is much less constrained.

A number of recently developed voice recognition technology products suggests that this technology is reaching the point where it may have utility in some simulation training. Two companies, IBM and Dragon Systems, have released products that allow voice recognition accuracy at a near a continuous speech rate, i.e., they allow recognition of speech without any artificial pauses. These products are designed and marketed as an alternative to conventional keyboard input for data and text entry, not as simulation technology. Both systems require intensive user voice training in order to achieve high rates of recognition accuracy[2]. The typical training of these systems involves several hours of individual entries by a given user. The training is required in order for the system to acquire an adequate sample of speech elements for a particular user. Without this training, recognition for these systems can drop dramatically to a range of perhaps 50%-70% accuracy. Such training is required of all extant recognition systems and that fact should be considered before adopting such systems in simulation training and evaluation environments.

[2] Defined as 98.5% or higher by the National Institute for Standards and Technology.

In addition to the speaker training requirement, recognition speed may also prove to be an implementation problem for these systems in more advanced simulator training applications. The target for current voice recognition systems is the normal conversational speech rate of about 180 words per sec. This is significantly slower then the typical communication rates for air-ground communications noted above. The ability of these systems to handle such high speech rates is unknown and may depend on the ability to modify the specific application software to use the higher predictability and limited vocabulary of aviation communications to enhance recognition speed. Only a research and development effort could reveal whether such a technical accommodation would be effective, however.

The technical immaturity of voice recognition technology may well limit its use in simulation training and evaluation, but it does not eliminate it as a potential tool. Simulation training has made some use of voice recognition as well voice synthesis technologies despite their inherent shortcomings.

Applications of Voice Technology in Simulation Training and Evaluation

From the discussion above, it should not be a surprise to learn that the application of voice technology to simulation training and evaluation is not very great. The technology's inherent problems of quality, extensibility and flexibility, and difficulty of implementation have restricted voice technology to simulation applications where tradeoffs among these problem areas could be made. Voice technology has been particularly useful where voice display intelligibility was important but not necessarily naturalness or quality, where the message set could be restricted but the quality of the voice kept high, and where voice recognition of high accuracy was not needed and speaker training was acceptable.

Air Traffic Control Training

The U.S. Navy and Marine air traffic control training simulators exemplify the possible use of voice technology in training simulation by virtue of their use of both voice synthesis and recognition technology . The simulators are specifically designed to support *ab initio* air traffic controllers in learning correct controller phraseology as well the fundamentals of the air traffic control task. Three ATC training simulator systems have been developed by the Naval Weapons Training Systems Center. The Tower Operator Training System (TOTS), Radar Traffic Control Facility (RATCF) and shipboard ATC facility simulator (CATTC/MTCC) all rely on the same basic voice synthesis (TTS) and recognition technology. The voice recognition systems require that both student and instructor train the system to accommodate it to their specific voice patterns. In these simulators, synthetic speech displays from pre-programmed pseudo-pilots respond as needed to trainee inputs with concurrent, appropriate changes in the controlled aircraft's heading, speed, and altitude. Studies of the effectiveness of these simulators by the Navy indicate that the systems are successful in achieving the intended training objective. However, no data are available on the specifics of voice technology usability or user acceptance.

Civil air traffic controllers in both the United Kingdom and the U.S. are taught with similar voice synthesis technology which mimics the voice of the pilots responding to ATC controller directives. For fundamental training of controllers in correct phraseology and related areas, these systems are achieving their goals. Again, the specifics of usability and user acceptance of the voice systems is not known.

Pilot Training and Evaluation

A review of the applications of voice technology in pilot training and evaluation for civil aviation application are reviewed in the Burki-Cohen, et al. (1999) study and will only be summarized here. The review found no existing pilot training simulators in either general or commercial aviation use that currently used synthetic speech or voice recognition technologies. Digital voice technology was used in many of the general aviation, personal-computer (PC) based training simulators reviewed and in the only commercial aviation simulator voice system, CAE's Ground and Air Traffic Environment System (GATES), available. None of the systems reviewed had any voice recognition capability. A few of these simulators have undergone effectiveness evaluation and those evaluations have yielded positive results. As with ATC training, no data are available on the efficacy of the voice display systems in simulating communications. The commercial aviation simulators which have incorporated The GATES system have yet to undergo formal training evaluation with the GATES system in operation so it is not known how effective GATES will be in providing the necessary level of realism. It is noteworthy that none of the existing civil aviation simulators used any form of synthetic speech technology.

No military flight simulators used for pilot training and evaluation could be found which incorporated voice technology for simulation of radio communications or other types of crew communications. Only unclassified sources available to the general public were searched, however, and it is possible that classified sources might reveal use of this technology for military training.

Conclusions and Recommendations

Instructional Utility and Simulator Realism

Before any conclusions can be drawn from the available information on voice generation and recognition technology, an important distinction needs to be drawn between simulation technology developed with the goal of providing an effective instructional support system and simulation technology developed with the goal of replicating real-world phenomena. It is well-established within the instructional technology field that considerable training value can be obtained from devices which do not replicate the real world, i.e., are not "virtual reality". For example, simple cardboard cut-outs of instrument panels can serve as useful training aids in developing instrument scanning

skills or in practicing emergency procedures. The cost-effectiveness of this approach to training simulation has been repeatedly demonstrated over the last few decades. Such an approach is heavily dependent on skills analyses to identify the appropriate level of technology needed to support the development and maintenance of those skills. This "skills-driven" approach to simulation technology would first ask the question of what types of skills aircrews are expected to develop and maintain and then whether the exiting voice technology could support the skills development process. No skills analyses of voice communications has been done within the context of simulator specifications. Only anecdotal and opinion data exist as to the form of communications simulation needed. These data appear to support the importance of communications monitoring skills for pilots (Burki-Cohen, 1999).

However, many pilots, instructors, and regulators, take the position that simulation technology success is measured solely by its ability to re-create reality. Physical fidelity requirements have, in fact, dominated much of commercial flight simulator development since its inception. Regulation of technical specifications of commercial flight simulators is much less a matter of instructional value and much more a matter of how similar the simulator is to operational aircraft (e.g., handling qualities) and their operating environment (e.g., daylight visual scenes). It might be expected then that the quality of voice generation, rather than simply its intelligibility, will be the most important requirement for communications simulation, regardless of the potential training value of lower quality voice systems.

State of Voice Generation and Recognition Technology

In assessing the state of voice technology with respect to its practical utility for commercial flight simulation training and evaluation, several criteria need to be considered. First, the intelligibility of voice displays used in simulations needs to be provided at a level comparable to that of human speech to avoid introducing additional cognitive loads to trainees during either training or evaluation. Intelligibility is a measure of the degree to which the listener can recognize the constituent speech elements. This level of intelligibility is well within the capabilities of higher quality speech systems currently available (e.g., DecTalk). It is possible that some lower quality systems could be employed which would rely on the context of the speech to enhance the comprehensibility of message components, but such a strategy is probably not worth the risk given the availability of better voice display systems.

Secondly, the quality or naturalness of the voice display needs to be considered. The fullness or richness of the speech generated and the absence of the mechanical sounds associated with machine-generated speech need to be considered. As no standard exists for objectively measuring voice quality, subjective assessments by a sample of end-users needs to be conducted on any candidate voice display system. Currently, only digitized speech has been found to meet the quality criterion. Progress in addressing the quality aspect of speech generation has been steady, but slow. It is not expected that major breakthroughs in this area will occur within the next five years, though the pace of evolution in computer technology is often difficult to predict. For this reason, synthetic

voice technology, with its many technical advantages should be not be ruled out for future simulator applications.

Related to the issue of display quality is the level of user acceptance of voice displays as surrogates for the real human speaker. User acceptance of voice generation technology has varied widely depending on the user population, application context, and type of technology. In the case of synthetic speech systems, the very mechanical quality found objectionable by some can be desirable, particularly if there is a reason for users to be able to identify the particular voice as machine-based rather than from a human speaker (Simpson, 1985). In general, it is anticipated that the commercial aviation user population (pilot, instructor, regulator) will not accept the current quality of synthetic speech. The training regimen may influence this acceptance to some extent, however. For *ab initio* or familiarization flight training, the higher quality synthetic speech systems may be acceptable. Certainly, they are deserving of some research and development attention by the aviation industry for their potential in other than line-oriented simulation. Line-oriented simulations, on the other hand, whether conducted for training or evaluation, are unlikely to use systems where the quality of the speech display is noticeably different from real life operations. Current quality levels of synthetic speech will not be adequate to meet this demanding criterion and are not likely to be so in the near future.

Voice recognition technology has made major advances in the last few years. Not only has voice recognition achieved enough reliability to be routinely used in simulation training (e.g., air traffic controller training), it has been significantly improved in commercial applications as well. Voice recognition systems can now reach levels near that of normal, conversational speech (e.g., about 180 words per min.). No system has yet achieved the much higher speech rates often found in aviation operations (e.g., more than 250 words per min.). This suggests that simulator applications of this technology may have to await further refinements. Additionally, the speaker training required of these systems to achieve very high levels of recognition accuracy may prove problematic for training management. However, this field is developing rapidly and it is conceivable that these limitations may overcome within the next five years. For this reason, the application of voice recognition technology in communications simulation, necessary for a fully automated communications simulation, should be considered as a long-term, viable enabling technology for future development in communications simulation.

References

Allen, J., Hunnicutt, S., and Klatt, D. (1987). *From Text to Speech: The MITalk System*. MIT Press. Cambridge, MA.

Burki-Cohen, et al. (1999). Realistic Radiocommunications Simulation. *SAE World Aviation Congress*, San Francisco, October.

Delogu, Cristina; Conte, Stella; Sementina, Ciro. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, May, 24, 153-168.

Hallahan, W.I. (1996). DecTalk Software: Text-to Speech Technology Implementation. *Digital Technical Journal*, April

Huang, X., Acero, A., Adcock, J., Hon, H-W, Goldsmith, J., Liu, J., and Plumpe, M. (1996). Whistler: A trainable text-to-speech system. *International Conference of Spoken Language Processing*, Philadelphia,.

Lauretta, D.J., Redman, R.D., and Antin, J.F. (1994). The effects of familiarization on the comprehension of synthetic speech in telephone communications. *Proceedings of the Human Factors Society, 34th Annual Meeting*, 1, 189-193.

Logan, J., Greene, B., and Pisoni, D. (1989). Segmental intelligibility of synthetic speech produced by rule. Journal of the Acoustical Society of America, 86, 566-581.

Luce, P., Feustel, T., and Pisoni, D. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25, 17-32.

McNaughton, D., Fallon, K., Tod, J., Weiner, F. (1994). Effect of repeated listening experiences on the intelligibility of synthesized speech. *AAC: Augmentative & Alternative Communication*, Sep, 10, 161-168.

Morrow, D., Lee, A.T., Rodvold, M. (1993). Analyzing problems in routine pilot-controller communication. *International Journal of Aviation Psychology*. 3, 285-302.

Oliver, W., Yu, J.C., Metois, E. (1997). The singing tree: Design of an interactive musical interface. *Design of Interactive Systems, DIS Conference*, Amsterdam.

Paris, C.R., Gilson, R.D., and Thomas, M.H., Silver, N.C. (1995). Effect of synthetic voice intelligibility on speech comprehension. *Human Factors*, 37, 335-340.

Schwab, E., Nusbaum, H., and Pisoni, D. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.

Simpson, C.A., et. al. (1985). System design for speech recognition and generation, *Human Factors*, 27, 115-142.

Smither, J. A. (1993). Short term memory demands in processing synthetic speech by old and young adults. *Behaviour & Information Technology*, 12, 330-335.

Sproat, R. (Ed.). (1997). *Multi-lingual Test-to-Speech: The Bell Labs Approach*. Kluwer Academic Publishers.

Thomas, J., Rosson, M., Chodorow, M. (1984). Human factors and synthetic speech. *Proceedings of the Human Factors Society 28th Annual Meeting. Santa Monica, CA: Human Factors Society*, 763-767.

Waterworth,, J. (1983). Effect of intonation form and pause duration of automatic telephone number announcements on subjective preference and memory performance. *Applied Ergonomics*, 14, 39-42.