# Limited Domain Synthesis of Expressive Military Speech for Animated Characters

*W. Lewis Johnson[1], Shrikanth Narayanan[2], Richard Whitney[1], Rajat Das[1], and Catherine LaBore[1]*

[1]Information Sciences Institute, [2]Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089
Johnson@isi.edu; shri@sipi.usc.edu; whitney@isi.edu; dasrajat@yahoo.com; labore@isi.edu

## ABSTRACT

Text-to-speech synthesis can play an important role in interactive education and training applications, as voices for animated agents. Such agents need high-quality voices capable of expressing intent and emotion. This paper presents preliminary results in an effort aimed at synthesizing expressive military speech for training applications. Such speech has acoustic and prosodic characteristics that can differ markedly from ordinary conversational speech. A limited domain synthesis approach is used employing samples of expressive speech, classified according to speaking style. The resulting synthesizer was tested both in isolation and in the context of a virtual reality training scenario with animated characters.

## 1. INTRODUCTION

Interactive training and learning environments frequently incorporate interactive animated characters, playing roles as virtual tutors, guides, companions, or participants in dramatized scenarios [11]. It is desirable for such characters to be able to communicate with users using a combination of spoken language and verbal gestures. They typically need to be able to generate a variety of spoken utterances, dependent upon the learner's previous actions and the state of the learning scenario. Therefore speech synthesis technology is potentially attractive for such applications.

Computer animation techniques are capable of creating synthetic characters that appear very lifelike [12][21]. It is important for synthetic voices for such characters to be similarly lifelike, natural sounding and emotionally expressive. Furthermore, human tutors often try to influence the motivational state of learners [13], in part through their tone of voice. Virtual tutors could be more effective if they could also use tone of voice to influence learners.

New concatenative synthesis methods, such as the unit selection approach used in AT&T's Next-Gen system [4], are able to generate higher sound quality. However it is difficult to control prosody in such systems, and to convey emotions and attitudes in particular. The role of attitude and emotion in prosody is still poorly understood in general. In the unit selection case prosodic qualities are limited by the prosodic qualities of the available speech samples. Samples collected by having speakers recite prompt texts are likely to have neutral or stilted affect, unsuitable for emotional synthesis. Waveform modification techniques such as PSOLA [20] can compensate for such limitations, but they introduce distortions and thus degrade the good sound quality that concatenative synthesis is intended to provide.

This paper describes initial efforts to create a high-quality synthetic voice for use in military training applications. It has been incorporated into the Mission Rehearsal Exercise [17], a prototype immersive training environment incorporating animated pedagogical agents being developed at the USC Institute for Creative Technologies. The voice is designed to exhibit prosodic characteristics of expressive military speech, including a portrayal of affect and attitude appropriate for military training contexts.

## 2. THE SCENARIO

The Mission Rehearsal Exercise (MRE) enables army officers to practice the kind of decision making that arises in peacekeeping missions. A virtual scene is presented using a panoramic display, a portion of which is shown in Figure 1, and a multi-channel sound system. The scene is populated with animated characters representing members of the trainee's platoon as well as local townspeople. One of the characters, shown in the foreground of the figure, acts as the platoon sergeant of the unit, both advising the trainee regarding what course of action to take and implementing the trainee's orders by issuing commands to the other members of the team. The trainee communicates with the platoon sergeant via a speech recognition module developed by Shrikanth Narayanan. A dialog management system controls interaction between the sergeant and the trainee [19].

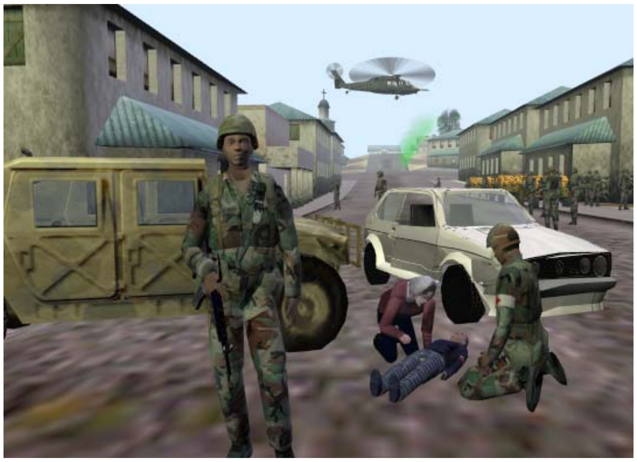

**Figure 1. View of Mission Rehearsal Exercise (MRE)**

In this particular scenario, the trainee has two conflicting goals to accomplish. He needs to help another unit across town that is trying to manage a hostile crowd. He must also deal with the situation shown in the figure, where one of his unit's vehicles has been involved in a traffic accident and a civilian boy has been hurt.

Our expressive speech synthesizer, called ESpeech, provides the voice of the platoon sergeant. The requirements of the scenario call for a synthesizer with a great deal of versatility and expressiveness, and good sound quality. The sergeant needs to speak in a tone of voice that expresses realistic emotional reactions to the situation: distress about the condition of the injured boy, a desire to act quickly and urgently, alarm if the trainee makes a decision that may jeopardize the mission, and disappointment if the goals of the mission are not met. He needs to speak with the lieutenant in a conversational tone of voice, consistent with his pedagogical role as the lieutenant's coach and advisor and his team role as subordinate officer. He also must shout orders to the troops, or shout to the trainee if the background noise becomes too loud. Since the graphical rendition of the scenario is very realistic, the sound quality of the synthesizer must also be realistic. Yet the emotionally charged story line and the rich immersive experience could also influence the way the user interprets and responds to the synthesized speech. We are evaluating the influence of each of these factors in the context of this scenario.

## 3. IMPORTANT PROPERTIES OF EXPRESSIVE MILITARY SPEECH

One area of focus of our synthesis work to date has been the generation of realistic command speech, the style of speech used by military officers to give orders. Command speech is spoken to be understood clearly and to convey authority [1]. Based on US Army manuals on command speech and our own spectral analyses of speech samples, it appears that there are at least three distinct types of command speech: stylized marching commands, commands to troops at a distance, and commands to individuals nearby. All are spoken clearly and with authority, but the commands at a distance have the additional characteristics of elevated maximum pitch, increased pitch range, and elevated amplitude. The ranges vary with individual speakers, but a typical f0 range for a male speaker is from 70 Hz to 280 Hz. Pitch accents are predominantly H* in the ToBI markup system, with exaggerated decrease of pitch at boundary tones. It is difficult to create a convincing command sound with conventional synthesizers simply by adjusting available prosody controls such as pitch and amplitude. What results instead is a high pitched or sing-song voice that fails utterly to convey the desired air of authority.

As indicated above, the ESpeech voice needs to convey a range of emotional attitudes. A number of researchers have documented correlations between basic emotions such as anger and sadness and prosodic characteristics, particularly mean f0, f0 range, and speech rate [3][7][10][15]. However such broad characterizations of emotion are not sufficient for creating realistic emotions for animated agents, and as a result richer computational models based on emotional appraisals have been developed [14]. Such a model is incorporated into the MRE system, and is useful in order to account for the variability in emotional speech that arises in the MRE scenario. The sergeant does not simply have a generalized emotional state such as anger or sadness, but has differing emotional attitudes toward various objects and events, both present and future. The sergeant can be simultaneously distressed about the current state of the boy, hopeful that he will be cared for in the future, and fearful that the time spent caring for the boy will put the overall mission in danger. As a result, the prosodic profile of the sergeant's voice should vary to convey different emotions depending upon the current topic of conversation. These emotional shifts can even occur in mid-sentence, if for example two people are mentioned each of whom invokes different emotional attitudes.

One prediction of the appraisal-oriented view is that emotions may influence the way in which individual words are emphasized via prosodic contours. This is manifested in the samples that we have studied as follows. There is a tendency to use more prominent pitch accents such as L+H* on emotionally charged words, in place of the more common H* accents. Hirschberg and colleagues [16] have noted that the L+H* accent is used to evoke a salient scale; degree of emotional appraisal could serve as such a scale. When the emotional appraisal is one of sadness, which tends to produce a lower f0, the low L pitch in the L+H* accent tends to fall even lower. Such pitch accents are not used consistently, and we suspect that they are more common in dramatized speech intended to convey emotions clearly; nevertheless this seems appropriate in MRE, which is a kind of dramatization.

The extent to which emotion is conveyed in the samples we have studied depends upon the style of speech that the speaker is employing; for example, command utterances do not convey sadness or fear. One way to account for this is to regard emotional expression as dependent upon the illocutionary force of the utterance [2]. Commands are classic instances of illocutionary acts; other illocutionary acts that arise this domain include warnings, alerts, and reports. It may prove useful to support multiple illocutionary acts in the ESpeech synthesizer, although to date our focus has only been on commands.

## 4. CONCATENATIVE SYNTHESIS OF MILITARY SPEECH

ESpeech is built using the FestVox synthesis toolkit [5] and is used in conjunction with the BEAT toolkit [8] for communication via speech and gesture. The overall system takes as input a partially ordered set of utterances to say and gestures to perform, annotated with XML tags. BEAT schedules the gestures and invokes Festival to synthesize waveforms for the utterances, passing along parameters derived from the XML annotations indicating the manner in which the utterance should be spoken. A timed sequence of visemes is computed from the timing data in the Festival multi-level data structure [18] after waveform synthesis in order to control the mouth movement of the animated character.

ESpeech currently uses a limited domain synthesis (LDS) approach [6]. We plan to augment LDS with a domain-independent diphone synthesizer for words that are out of vocabulary or lack the appropriate prosodic characteristics. LDS is good at recreating the speaking style of the original speaker; the challenge is to use LDS to produce a variety of speaking styles and emotions as needed.

Although the sergeant needs to be able to switch speaking styles from one utterance to the next, as he addresses different individuals in the scene, it is not necessary for him to switch speaking styles in mid-sentence. Therefore ESpeech incorporates multiple sets of units, each representing a different speaking style. The XML markup specifies which speaking style should be used for a given utterance. Four styles are currently defined: shouted commands, shouted conversation, spoken commands, and spoken conversation. The enumeration of speaking styles was determined in part through empirical testing of the synthesizer; if the synthesizer was found to generate utterances containing samples that sounded inappropriate together, we divided the samples into disjoint subsets indexed by speaking style. The amount of coverage provided for each speaking style is dictated by the needs of the training scenario. Currently the greatest vocabulary coverage is provided for spoken conversation, 316 words; second most coverage is provided for shouted commands, 72 words.

## 5. GENERATING PROSODY AND CONVEYING EMOTION

The ability of a limited domain synthesizer to generate prosody and convey emotion is dependent upon the recorded samples and the way in which the synthesizer selects and combines the samples. The following guidelines were found to be helpful in generating prompts that yielded suitable samples.

Words were classified according to where in an utterance they needed to be able to appear. Three categories of position were used: sentence initial, sentence medial, and sentence final. Sentence final position is distinctive because sentence final words carry boundary tones, which contrast strongly with other parts of the sentence. Sentence initial and sentence medial are distinct for two reasons. Pitch generally tends to decrease over the course of declarative utterances. Also, given and new information both tend to vary in pitch and occur in different parts of the sentence. This distinction was most important for conversational speaking styles; command speech tends to be more uniformly high pitched except for boundary tones and words receiving heavy emphasis.

In order for the synthesizer to generate speech with convincing emotion it is helpful to record samples that exhibit these emotions. To achieve this we constructed prompts in which words would appear in a similar sentence position and have similar affect as in the intended utterances. This was particularly important for key emotionally charged words, such as the word “serious” in the sentence “The boy has serious injuries.” In order for this approach to work effectively it is important that the limited domain synthesizer use the emotion-bearing words consistently. This turns out to be straightforward in practice, both because the limited domain restricts the range of emotions that might be expressed about a given object and because different emotions can be associated with different lexical items. For example, if “serious” is used as an emotion-bearing word, then the phrase “not serious” should be avoided because it would have inappropriate prosody, and alternative word such as “minor” should be used instead. This approach is particularly appropriate for MRE, which incorporates a natural language generator that is capable of choosing different words in different emotional states [9].

To ensure that emotions were conveyed as clearly as possible, we found it useful to have the voice talent rate each utterance on a set of emotional scales before recording, in order to help them focus on what emotion they were trying to convey.

To further improve the quality of the synthesis it was useful in some cases to record words with different preceding phonemes, so that the synthesizer could select a sample that matches the phonetic characteristics of the preceding word as closely as possible. Even then the waveform synthesis phase would sometimes select units with inappropriate prosody. One common problem was the insertion of words carrying boundary tones in the middle of phrases. This is not a serious flaw, since it simply sounds as if the speaker is pausing momentarily in mid-sentence. More serious are the occasional instances where sentence-medial samples are placed at the ends of sentences, which sounds distinctly unnatural. Our planned future work includes looking for ways of improve unit selection in order to eliminate these anomalies.

## 6. EXPERIMENTAL EVALUATION

A preliminary evaluation of the ESpeech synthesizer was recently completed. Subjects were presented a random mixture of natural and synthesized speech, and were not informed as to which was which. The subjects rated the test utterances in terms of a) overall sound quality, b) naturalness of intonation, c) the extent to which the utterance sounded like a command, and d) emotional expression along a number of dimensions – concern, confidence, fear, and hope. Ratings were on a Likert scale (e.g., 1 = very unnatural, 5 = very natural); for some of the emotion ratings the subjects also had the option of indicating “can’t say” and not providing any rating. Two versions were prepared of each set of experimental materials so that utterances that were natural in one version were synthesized in the other, and vice versa. Subjects were volunteers, primarily USC staff and students.

Two experimental conditions were constructed. In one subjects heard test sentences in audio form only, presented in random order. In the other, subjects viewed a videotape of a trainee interacting with a modified version of the MRE scenario, in which natural speech was randomly substituted for synthetic speech. The video was viewed on a desktop video player, and the sound, including both the sergeant’s voice and background sounds and comments by other characters, were played through stereo speakers. This compromise delivery approach recreated many of the characteristics of the learning experience, and made the test easy to administer, although the immersiveness of sound and video were reduced and the subjects were passive observers rather than active participants.

Ten subjects took the video version of the test, and fifteen took the audio version. The video version had 18 test utterances, and the audio version had these plus seven others.

For video, the rating for natural commands was 4.475, N=40. The rating for synthetic commands was higher than the natural voice rating, 4.6, N=40, although the difference is not significant. For audio only, natural commands were rated at 4.36, N=97, and synthetic commands were rated at 4.02, N=92. The difference in ratings between natural and synthetic was significant, at $p<0.02$, as was the difference between audio synthetic and video synthetic, at $p<0.0001$.

For conversational speech the disparities were greater. Intonation rating for video natural utterances was 4.22, N=45, for video synthetic 3.16, N=45. The ratings for audio only were significantly higher: for natural utterances 4.57, N=74, and for synthetic utterances 3.64, N=101.

The following factors may help account for the differences in command vs. conversational ratings. Command intonation has less variability, making it easier to synthesize. Sound quality was rated slightly better overall for command utterances (synthetic command rating = 3.78, synthetic conversational rating = 3.56), and intonation quality was positively correlated with sound quality (0.616). The subjects for the most part did not have military training, which may have affected their ability to judge the naturalness of military command speech.

The relationship between visual and story context and ratings of quality was more complex than we anticipated, and we are analyzing this further. The natural utterances were recorded in isolation, rather than in the context of the story, and we suspect this had a negative impact on naturalness of conversational speech in the story context.

Recognition of speaking style was dependent upon video and story context. In the audio version, ratings of command utterances (natural and synthetic) as command-like was 4.78; for video the rating was 4.99, significantly greater at $p<0.0001$.

Speaking style influenced what emotions were perceived, and also what emotions were perceivable. In 59% of the command utterances subjects indicated that it was impossible to assess the degree of hope in the voice; for video only this proportion rose to 94%. Command utterances were rated high confidence (4.64) and low fear (1.95). Again these ratings were more extreme in the video version (e.g., fear = 1.19).

## 7. CONCLUSIONS AND FUTURE WORK

Using limited domain synthesis techniques, it is possible to generate military speech that is comparable in quality to natural speech, particularly for command speech. Style of speech can strongly influence perceived emotional content. Video and story context influence perceived emotion and attitudes.

In our future work, we plan to revise the pitch prosody target generation and unit selection so that the synthesizer is able to do a better job of planning prosodic contours, particularly in conversational speech. We are constructing a domain-independent diphone synthesizer for out-of-vocabulary words and words that lack suitable limited-domain samples. Further tests are planned, with military personal as evaluators. We also plan to construct additional voices, building on lessons learned in this initial phase of the project.

## 8. ACKNOWLEDGMENTS

Thanks to Jamison Moore and Murtaza Bulut for contributing their efforts to the work presented in this paper, and to the volunteer evaluators. Thanks also to the voice talent for his time and effort in the creation of the Sergeant voice, and to David Miraglia for his help in preparing of the video testing materials. This work was supported by the U.S. Army Research Office under contract #DAAD19-99-D-0046.

## 9. REFERENCES

[1] ADTDL. Commands and the Command Voice. In http://www.adtdl.army.mil/cgi-bin/atdl.dll/fm/22-5/ch2.htm, ch. 2, 2001.

[2] J.L. Austin. *How to Do Things with Words*. Harvard University Press, 1962.

[3] Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614-636.

[4] M. Beutnagel. The AT&T Next-Gen TTS System. *Joint Meeting of the ASA*, Berlin, pp. 15-19, 1999.

[5] A.W. Black and K. Lenzo. Building voices in the Festival speech synthesis system. http://festvox.org, 2000.

[6] Black, A.W. (2000). Limited domain synthesis. *ICSLP '2000*, Beijing, China.

[7] Cahn, J.E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society* 8:1-19.

[8] Cassell, J., Vilhjálmsson, H., & Bickmore, T., (2001). BEAT: the Behavior Expression Animation Toolkit. *Proceedings of SIGGRAPH 2001*, ACM Press.

[9] Fleischman, M. & Hovy, E. Emotional variation in speech-based natural language generation. Submitted, 2002.

[10] Johnson, W. F., Emde, R. N., Scherer, K. R., & Klinnert, M. D. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry, 43*, 280-283.

[11] Johnson, W.L., Rickel, J.W., and Lester, J.C.(2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education* 11, 47-78.

[12] J. Lasseter. Principles of traditional animation applied to 3D computer animation. *Computer Graphics* Proceedings, Annual Conference Series, ACM Conference Series, ACM SIGGRAPH, 35-44, 1987.

[13] Lepper, M.R., Woolverton, M., Mumme, D., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S.P. Lajoie and S.J. Derry (Eds.), *Computers as cognitive tools*, 75-105. . Hillsdale, NJ: Lawrence Erlbaum Associates.

[14] Marsella, S., & Gratch, J. (2001). Tears and Fears. Proceedings of the Cognitive Science Society.

[15] Pell, M.D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. Submitted.

[16] J. Pierrehumbert and J. Hirschberg. The meaning of intonational contours in the interpretation of discourse. In P.R. Cohen, J. Morgan, and M.E. Pollack (Eds.), Intention in Communication, 271-311. Cambridge U. Press, 1990.

[17] W. Swartout, R. Hill, J. Gratch, W.L. Johnson et al. Towards the Holodeck: Integrating graphics, sound, character, and story. *Proceedings of the 2001 International Conference on Autonomous Agents*, 409-416. New York: ACM Press, 2001.

[18] P. Taylor, A.W. Black, and R. Caley. The architecture of the Festival Speech Synthesis System. 3rd ESCA Workshop on Speech Synthesis, pp. 147-151, Jenolan Caves, 1998.

[19] D. Traum and J. Rickel. Embodied agents for multi-party dialog in virtual worlds. Proc. of the 1st International Conference on Autonomous Agent and Multi-Agent Systems. New York: ACM Press, 2002.

[20] Valbert, H., Moulines, E., & Tubach (1992). Voice transformation using the PSOLA technique. *Speech Communications* 11:175-187.

[21] Zhao, L., Costa, M., & Badler, N.I. (2000). Interpreting movement manner. *Proceedings of Computer Animation 2000.*