VOICE CREATION FOR CONVERSATIONAL FAIRY-TALE CHARACTERS

Joakim Gustafson[1] *and Kåre Sjölander*[2]

[1] Voice Technologies, Expert Functions, TeliaSonera, Farsta, Sweden. joakim.gustafson@teliasonera.com
[2] Centre for Speech Technology, KTH, Stockholm, Sweden. kare@speech.kth.se

ABSTRACT

The NICE fairy-tale game system allows users to interact with conversational fairy-tale characters in a 3D world environment. Apart from engaging in conversation, the characters are able to perform physical actions in this simulated world. The goal is to create believable fairy-tale characters with distinct personalities. The personality of the characters will be conveyed by their appearance, their voices, how they express themselves and what they are doing. This paper describes the requirements a fairy-tale game domain poses on a spoken output generation system. The implementation of a unit selection synthesizer that meets these requirements is also described.

1. INTRODUCTION

The NICE fairy-tale game system lets users explore a 3D fairy-tale world, where they meet animated fairy-tale characters with conversational skills [19]. The users' task is to help the fairy-tale characters to solve problems that they present verbally. They will solve the problems by collaborating with one of the fairy-tale characters, Cloddy Hans, who is helpful, but a bit stupid. This paper describes the requirements on a verbal output generation system for believable conversational fairy-tale characters in a computer game application. For the fairy-tale characters to be lifelike and believable in their roles in the game they have to be provided with natural sounding voices with distinct personalities. The characters also have to be able to engage in clarification dialogues that make the number of possible system utterances prohibitively large. For these reasons, a limited domain unit selection synthesis system has been developed, with the following design criteria:

- Swedish language, feasible to add new languages
- high quality sounding with personal speaking style
- easy to build new voices
- easy to use as system component in a dialogue system
- faster than real time speech generation
- prosodic control (emphasis and speaking rate)
- voice design using acoustic signal processing (giant/gnome)

These requirements led to the implementation of **Snacka** a new corpus based unit selection synthesizer, which extends on the Snack sound toolkit [27]. The system only requires a set of speech recordings with matching orthographic transcriptions files. With this as input the system is able to generate a synthesis voice without further manual intervention. The aim in creating the system was to have symmetric analysis and synthesis components. Modules for tasks such as letter-to-sound, text analysis, co-articulation modelling, have been designed with both tasks in mind and are used by both components. In this way it is assured that there is a high degree of match between unit database and synthesis output, which increases synthesis quality. A special quality assessment tool has been developed that makes it easy to check domain sentences. Sentences that sound strange can be examined, automatic segmentation errors can be corrected, and bad units can be pruned manually. These changes can be used instantly for speech synthesis, without rebuilding the whole voice in most cases. It is also possible to perform acoustic voice transformations in order better match the personality of the fairy-tale characters. The paper will also briefly describe the implementation of the Snacka synthesis system.

2. INTERACTIVE STORIES

The goal of the NICE system is to let users interact multimodally with conversational agents in an adventure game application. When text adventure games originally appeared in the 70's, they could achieve a limited sense of omniscience, since their goal-oriented users wanted to be immersed into the adventure, which refrained them from trying to deceive the system. The immersion was limited, due to the systems' limited understanding capabilities. Although limited, they did use situation dependent parsers that would not allow actions in the wrong contexts and that asked clarification questions if the user provided text input that was ambiguous in a certain situation. Paradoxically, today's commercial 3D adventure games have much more limited input capabilities. Typically, users can only navigate in the 3D world, selecting objects via mouse input and selecting what their avatar should do or say from predefined menus. In this way, today's adventure games limit the users' role to deciding in what order thing should happen. Thus, the lack of linguistic input capabilities limits the types of plots that these 3D adventure games can use.

Over the recent years interactive story-telling research systems have been developed that in some cases allow linguistic input. Hayes-Roth [20] lists a number of principles that are important for these interactive story-telling systems. The user has to be given an illusion of immersion by participating in an interesting story where they feel that they are actively participating by interacting with the characters in a meaningful and natural way.

Animated characters that are used in story-telling systems and adventure games have to be believable. According to Loyall [22] the following set of features are important for believable agents: *personality, emotion, self motivation, change, social relationships, consistency of expression* and *illusion of life*. Allen [1] lists six determinants of agent personality: *agent concerns*; *motivational profile*; *sensitivity of emotional states*; *reactive and deliberate processes; persistence in attaining goals*; and *choice of expressive behavior*. Rizzo et al [25] present a practical model of agent personality, where the characters' goals and actions constitute their personalities. A character with selfish goals will perform selfish actions that will make observers judge it as having a selfish personality.

3. CONVERSATIONAL SKILLS

It is important for lifelike characters to have conversational skills. They have to be able to communicate their goals and plans to the user, and they should be able to cooperate with the user to solve problems. To convey personality and to build a collaborative trusting relationship with the users, the characters also have to be able to engage in small talk [14]. Collaboration in dialogue is the process where the participants coordinate their action towards a shared goal[18]. Allwood et al [2] describe four basic communicative functions that correspond to the Clarks's [15] levels at which problems can arise during the grounding process:

- contact at the vocalization and attention level
- perception at the identification level
- understanding at the meaning level
- attitudinal reaction at the proposal and uptake level

Traum [28] presents a computational model of how participants in dialogue come to reach a state of mutual understanding of a speaker's utterance. This model uses the following grounding acts: *initiate, continue, repair, request repair, display, acknowledge, request-acknowledge,* and *cancel.* In conversation the coordination of turns is crucial. Allwood [3] defines a turn as a speaker's right to the floor, and that this right is regulated by a number of turn management sub-functions that can be expressed verbally or non-verbally. There are two simultaneous information channels in a dialogue: the information channel from the speaker, and the back-channel feedback from the listener. The back-channel feedback indicates attention, feelings and understanding, and its purpose is to support the interaction [30]. It is communicated by anything from short vocalizations like "mm" to utterances like "I think I understand", or by facial expressions and gestures [17]. Brennan and Hulteen [13] have suggested a feedback model for human-computer interaction that builds on a grounding process. The system should provide positive feedback in successful contexts and negative feedback when problems have been detected. There are a number of studies of linguistic adaptation during error resolution (for an extensive overview see Bell [5]). On the acoustic level the following adaptations have been reported: increases in duration and pitch range, hyperarticulation, louder and often with clearer articulation. Other linguistic error adaptation strategies include rephrasing, use of cue words like "no" and explicit clarification meta utterances like "what do you mean?" Disfluencies are indicators of problems in dialogue, but they have also been found to be useful for the listener in processing the speakers turns. Initial fillers are used to manage turn taking, pauses and fillers indicate feeling-of-knowing and some fillers like "uh" can speed monitoring of a subsequent word [12]. The aim of the NICE project is to build characters with these kinds of conversational abilities.

4. LIMITED DOMAIN SPEECH SYNTHESIS

In commercial spoken dialogue system a lot of effort is put into persona design of the voices. Companies want voices that support branding and that either reflect the company or attract certain user groups. These systems mostly use recorded prompts to obtain naturalness and quality. Nass and Lee [24] examined if it was possible to convey personality with speech synthesis. They provided a synthetic voice with stereotypical extrovert vocal features (high loudness, increased pitch, a great frequency range and a fast speaking rate). The users where then asked to give their impression of the voice. The result was that they often described the voice with extrovert personality adjective items. In the Nice system the characters can to engage in domain-oriented dialogues, that include clarification system utterances that include parts of a problematic user turn. This makes the number of possible system utterances too large to be pre-recorded. In limited domains it is possible to build natural sounding synthetic voices using a unit selection synthesizer, like Festival/FestVox [9,11]. Limited domain synthesizers [10] have been developed for a number of applications, e.g., in a task oriented dialogue system in the travel domain [26] and for animated characters in a military training domain [21]. The limited domain synthesis approach was decided to be the most promising method for creating voices for the fairy-tale characters in the NICE project.

5. THE NICE FAIRY-TALE GAME SYSTEM

The NICE Fairytale System is a multimodal spoken dialogue system where the user can explore a fairy-tale world inspired by the Danish author H.C. Andersen. The users interact multimodally with embodied conversational fairy-tale characters. The role of the user is to help the characters to solve different kinds of problems. Since the users will meet different characters that have different character traits it is important that their appearance and voices are designed accordingly. Some of the fairy-tale characters are shown in Figure 1. The users can talk to the characters and point at objects in the 3D environment, but they cannot manipulate object themselves. To make it possible for the user to always manipulate physical objects, a helping character has been introduced in the game. His name is Cloddy Hans (loosely inspired by the character from H.C. Andersen's story with the same name). Cloddy Hans has the following personality traits:

- practical rather than intellectual
- friendly and honest
- slow in both mind and action
- unselfish, no goals of his own

This means that Cloddy Hans and the user form a team where Cloddy Hans can perform physical action, but he does not know what to do, while the user knows what has to be done, but he cannot perform the physical actions needed. This means that they have to cooperate to solve the problems that the fairytale characters present them with.

The fairytale game will start with an initial scene where the user meets Cloddy Hans in H.C. Andersen's fairy-tale lab. If the user wants to get to know Cloddy Hans they can start with engaging in small talk with him.

Figure 1. The fairy-tale world and some of the characters.

Figure 2. Cloddy Hans in the fairy-tale lab.

The initial scene is a "grounding game" with the purpose of being a training session before the user enters the fairy-tale world. Cloddy Hans explains to the user that he is trying to build a fairy-tale with H.C. Andersen's fairy-tale machine. He want the user to help him, first by identifying a number of symbols that denotes what types of objects should be put in what slot of the machine, see Figure 2.

Then he will ask the user what some of the fairy-tale objects and characters on a shelf are called, since he does not know that. Finally, he will ask the user to first tell him what object to take from the shelf, and then in which slot to put it. During this interaction he will ask the user to point at objects in cases where he has trouble understanding. When a certain number of objects have been put in the appropriate slots, Cloddy Hans will pull the lever to create the story. A door will open and the user can enter the fairy-tale world, where the selected objects and fairy-tale characters appear and have the functions that the user and Cloddy Hans have agreed upon.

By choosing a very simple predefined task in the initial scene, the system is able to anticipate what the user will have to say to solve it. The real purpose is not to solve the task, but to engage in a collaborative grounding conversation where the user and Cloddy Hans has to agree on what the fairy-tale objects can be used for and how they should be referred to. This process also lets the players find out (by trial-and-error) how to adapt in order to make it easier for the Cloddy Hans to understand them, e.g. by using multimodal input in certain contexts. Hopefully, this will make the interaction smoother in the subsequent game in the fairy-tale world, since the objects and characters that appear in it already have been grounded in the initial scene.

Table 1. An overview of the task independent dialogue speech acts, that has been recorded for the Cloddy Hans voice.

Plan Regulating	Error Handling	Turn Handling	Attitudal Feedback	Extralinguistic sounds
agree | disagree	report (hearing | not hearing)	feedback continuer	(positive | negative) filler words | filled pause	clear throat
Ask for (accept | reject | correct) request	report (understanding | not understanding | misunderstanding)	floorholder (easy | hard) question	yes | no	cough
(accept | reject) offer	report (knowing | not knowing)	backchannel question	apology | non-apology	exhalation
	report (correct | wrong) action	neutral filled pauses	grateful | ungrateful	inhalation
	error acknowledgement		attitude to (grateful | ungrateful)	laughter
	ask for (clarification | repetition | rephrase)		attitude to (success | failure)	sigh
	(open | bounce) question		attitude to (good | bad)	

6. CORPUS DESIGN FOR OUTPUT GENERATION

Since the system will be used for collaborative grounding dialogues, it is also important that the system is able to say everything it had been designed to understand. In the task-oriented dialogues it is thus important that Cloddy Hans can talk about the physical actions he will have to perform to solve a certain task. In the small-talk domain he must be able to ask the users the same questions that the system has been prepared for to answer. Another design criterion was that Cloddy Hans should be able to understand and generate both grounding and turn regulation utterances. Finally, to facilitate rephrasing as an error-handling strategy, all domain and meta utterances have been provided in a version with alternative wording.

The fairy-tale characters will be able to generate both verbal and non-verbal behaviour. Cloddy Hans is currently able to perform the following non-verbal behaviours: *physical action*; *emotional display*; *state of mind*, *turn regulation cues*, *back-channeling gestures*. Cloddy Hans's verbal output includes the following types: *responses to instructions from the users* (confirmations; acknowledgements and clarification); *initiatives to fulfil plans; social utterances*, *meta utterances* (grounding and error handling), *attitudal feedback* and *turn regulation utterances* (floor-holders, back-channels and filled pauses). The turn regulation utterances and attitudal feedback are used to buy time while the system generates the next Cloddy Hans utterance, but more importantly, they are used for the purpose of conveying his uncertain personality.

Table 1 shows an overview of the general dialogue regulating speech acts that will be used in all tasks and plots. The plan regulating acts include utterances like *"what do you want me to do?",* the error handling acts include *"Could you repeat that?",* turn handling utterance include *"okay"*, the attitudal feedback include both phrases like *"too bad"* and filled pauses like *"uhm"* recorded with both positive and negative prosody.

A number of domain utterances have been designed that let Cloddy Hans explain the overall plot as well as talk about the task at hand. To facilitate grounding, Cloddy Hans has been given the possibility to ask clarification questions about everything the user can ask him to do. The recorded corpus includes sentences where all objects and slots have been placed in both medial and final position of the utterances, making it possible to ask clefted clarification questions like *"Is it the axe you want me to put in the useful slot?"* as well as *"Was it in the useful slot you wanted me to put the axe?"*. All utterances were also tagged with prosodic phrase boundaries and emphasized words prior to recording. Utterances with slots and objects in all combinations of position and emphasis have been recorded. Finally a number of sentences with only function words were recorded to make sure that they were sufficiently covered, and to increase the coverage of Swedish di-phones a set of old Swedish sayings were recorded.

7. THE UNIT SELECTION SYSTEM

An important role of the actual realization of the verbal utterances in the fairytale domain is that it conveys the characters' personality. The friendly, but dunce and slow Cloddy Hans should not have the same way of expressing himself, or not even the same type of voice, as the evil, smart and selfish prince. A unit selection synthesizer was developed to achieve a natural voice quality and prosody. A unit selection voice for Cloddy Hans in the Fairytale machine scenario domain has already been recorded. The guiding principle in designing the system was that corpus analysis and synthesis processing should be symmetric. That is, given a recorded sequence of words in the corpus and a sequence that needs to be synthesized that matches the first one verbatim, every segment contained in the recording should be generated as possible candidates for the search. In order to attain this, modules for tasks such as letter-to-sound, text analysis, co-articulation modelling, etc, were designed with both tasks in mind and used

during both the analysis and synthesis stages. This increases the degree of match between the available units in the recorded corpus and the output units needed for synthesis, which in turn increases the quality the synthesized speech [6].

One important system module is the speech segmentation component, which is used to automatically process the speech corpus. The module uses a Hidden Markov Model based method to generate accurate phone and word label transcriptions automatically. Input is taken from a corpus containing sound files and matching orthographic transcription files. One feature of the module is that it can process long files. In principle the input can consist of one long sound file and one transcription file. Initial experiments during development of the synthesis system used talking books, which were available in this format.

A unit selection inventory is created and clustered using decision trees. The units use acoustic vectors with Mel cepstrum, F0, and first order regression coefficients. Standard features such as phone context, stress, phone position, syllable position, lexical stress, focus, etc, are used in the decision tree questions [8]. The basic synthesis unit used is half-phones [7] with variable boundaries using optimal-coupling [16]. Pronunciation modelling is based on co-occurrence statistics. The speaker used for corpus collection might exhibit a certain preference to pronounce a word in a certain way in a given context. This information is stored for later use by the synthesis module. A prosody model is trained on the corpus, also using decision trees [29]. All of the above modules have been implemented using the Snack sound toolkit [27] extended with new primitives for tasks such as decision tree modelling, Viterbi search, etc. This means that the whole voice building process can be contained in one single script. No external tools are needed, since the needed functionality is available as commands in the scripting language itself. The build script creates a voice definition file that contains all information about the units of the corpus needed for synthesis. The sound segments are taken from original sound files that also are part of the voice.

A separate synthesis module was developed for use in the game system. At synthesis time a target description is generated from an XML-tagged text obtained from the output generator. A set of XML tags is used for emphasis mark-up, prosodic phrase mark-up, extra-linguistic sounds and other meta-information. The target description is used to find matching unit candidates. The best sequence of unit candidates is found using Viterbi search. These optimal units are finally joined together to form the output utterance.

Special XML-tags are used to coordinate the verbal output with body gestures and actions, e.g. "*do you want <EMPH>me</EMPH> to take* <PointAt object="Axe">*that*</PointAt> *and put it* <PointAt slot="Useful">*there*</PointAt>".

The synthesizer creates a transcription in addition to the generated sound. This transcription contains time stamped phones and animation tags which are used by the animation system for lip-synchronization and gesture coordination. While still highly experimental the system is capable of faster than real-time speech synthesis.

A special tool, based on WaveSurfer, has been created for easy testing and quality assessment of synthesis voices. Typical domain sentences can be generated using simple point-and-click selection. If a faulty utterance needs to be analysed this tool can be used to correct problems with the automatic segmentation. Also, bad units can be pruned manually. In many cases the utterance can be re-synthesized instantly without rebuilding the full voice.

Voice transformation was applied in order to better match the personality of the Cloddy Hans character. Speaking rate and vocal tract length modifications were tried using a version of the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) algorithm [23]. The transformation was carried out on the whole recorded voice corpus after creating the voice description. In an earlier study similar transformations were shown to influence the way users interacted with Cloddy Hans [4].

8. CONCLUSIONS AND FUTURE WORK

The unit selection synthesizer is currently being used for user tests at the Telecommunication museum in Stockholm. Adults and children are invited to interact with Cloddy Hans, who appears in full size on a back-projected video wall. The system is run in a semi-automatic mode where a human wizard can override the utterance analysis at different point and/or decide to generate other actions than the system selected automatically. It could either be that the wizard decides to initiate a clarification sub-dialogue if the speech recognition confidence score was too low, or to let Cloddy Hans take the initiative to do things if the user is uncertain. After the interactions the users are asked what they thought about their interactions with Cloddy Hans, where they among other things are asked how they judge his personality and what they thought about his voice. Preliminary results from these evaluations will be included in the final version of this paper.

Future work on the synthesizer includes improvements on the prosodic models used as well as experimenting with ways to change the speaking style, for example trying to make repetitions acoustically different from the original utterances.

9. ACKNOWLEDGEMENTS

The work described in this paper was supported by the EU/HLT funded project NICE (IST-2001-35293), www.niceproject.com. Part of the work was carried out at CTT, the center of speech technology at KTH. The authors would like to thank Mattias Heldner for lending his voice to Cloddy Hans.

10. REFERENCES

[1] Allen, S. “Control states and motivated agency,” In André, E., ed., Proceedings of the i3 Spring Days’99 Workshop on Behavior Planning for Life-like Characters and Avatars, 43–61, 1999.

[2] Allwood, J., Nivre, J. and Ahlsén, E. “On the Semantics and Pragmatics of Linguistic Feedback,” in Journal of Semantics, 1992.

[3] Allwood, J. "Reasons for management in dialog" in Beun, R.J., Baker, M. and Reiner, M. (eds.) Dialogue and Instruction. Springer-Verlag.pp 241-50, 1995.

[4] Bell, L., Gustafson, J. and Heldner, M. "Prosodic adaptation in human–computer interaction", Proceedings of ICPhS 03, Barcelona, Spain. 2003.

[5] Bell, L. “Linguistic adaptations in spoken human-computer dialogues - Empirical studies of user behavior,” Doctoral Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, 2003.

[6] Beutnagel, M., Conkie, A. and Syrdal, A. “Diphone synthesis using unit selection.” In: The 3rd ESCA/COCOSDAWorkshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998.

[7] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. and Syrdal A. “The AT&T Next-Gen TTS System”, 137thAcousticalSociety of America meeting, Berlin 1999.

[8] Black A., and Taylor, P. “Automatically clustering similar units for unit selection in speech synthesis”, Proceedings of Eurospeech 97, Rhodes, Greece. 1997.

[9] Black A., Taylor P., and Caley R., “The Festival speech synthesis system,” http://festvox.org/festival. 1998.

[10] Black, A. and Lenzo, K. “Limited Domain Synthesis”, ICSLP2000, Beijing, China, 2000.

[11] Black, A. and Lenzo, K. “Building Voices in the Festival Speech Synthesis System,” http://festvox.org/bsv, 2000

[12] Brennan, S. “Processes that shape conversation and their implications for computational linguistics,” Proceedings, 38th Annual Meeting of the ACL. Hong Kong, 2000.

[13] Brennan, S. and Hulteen, E. “Interaction and feedback in a spoken language system: a theoretical framework,” Knowledge-Based Systems(8): 143-151, 1995

[14] Cassell, J., Bickmore, T. “Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents,”, User Modeling and Adaptive Interfaces(12) 1-44, 2002.

[15] Clark, H. “Managing problems in speaking”, Speech Communication, 15:243-250, 1994.

[16] Conkie, A. and Isard, S., “Optimal coupling of diphones," *Progress in Speech Synthesis*, Springer, 1997, pp. 293-304.

[17] Goodwin, C. “Conversational Organization: interaction between speakers and hearers,” New York/London, Academic Press, 1981.

[18] Grosz, B. and Sidner, C. “Attention, Intention, and the Structure of Discourse,” Computational Linguistics 12(3), pp. 175–204, 1986.

[19] Gustafson, J., Bell, L., Boye, J., Lindström, A. and Wiren, M (2004) "The NICE Fairy-tale Game System", forthcoming, proceedings of SIGdial 04, Boston, 2004.

[20] Hayes-Roth, B. “Character-based Interactive Story Systems,” IEEE Intelligent Systems and Their Applications 13.6: pp 12-15, 1998.

[21] Johnson, W., Narayanan, S., Whitney, R. Das, R., Bulut, M. and LaBore, C. “Limited Domain Synthesis of Expressive Military Speech for Animated Characters,” In Proceedings of the IEEE TTS Workshop, 2002.

[22] Loyall, B. “Believable Agents: Building Interactive Personalities,” Ph.D. thesis, Technical Report CMU-CS-97-123, School of Computer Science, CMU, 1997.

[23] Moulines, E. & Charpentier, F., “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, Speech Communication Vol. 9 (5/6), pp. 453-467, 1990.

[24] Nass, C. and Lee, K. “Does computer-generated speech manifest personality? An experimental test of similarity-attraction,” Proceedings of the CHI 2000 Conference, The Hague, The Netherlands, 2000.

[25] Rizzo, P., Veloso, M., Miceli, M. and Cesta, “A Personality-driven social behaviors in believable agents,” In Proc. AAAI 1997 Fall Symposium on "Socially Intelligent Agents", pp 109-114, 1997.

[26] Rudnicky, A., Bennett, C., Black, A., Chotomongcol, A., Lenzo, K., Oh, A., Singh, R. “Task and domain specific modelling in the Carnegie Mellon Communicator system,” Proceedings of ICSLP, 2000.

[27] Sjölander K and Beskow J. “WaveSurfer - an Open Source Speech Tool,” In Proceedings of ICSLP, 2000.

[28] Traum, D. “A Computational Theory of Grounding in Natural Language Conversation,” TR 545 and Ph.D. Thesis, Computer Science Dept., U. Rochester, December 1994.

[29] Taylor, P. "Analysis and synthesis of intonation using the tilt model," Journal of the Acoustical Society of America, vol. 107, no. 3, pp. 1697--1714, 2000.

[30] Yngve, V. ”On getting a word in edgewise,” In Papers, from the Sixth Regional Meeting, Chicago Linguistic Society, pages 567–577, 1970.