Voice Chat with a Virtual Character: The Good Soldier Svejk Case Project

Jan Nouza, Petr Kolár, Josef Chaloupka

SpeechLab, Technical University of Liberec
Halkova 6, 461 17 Liberec, Czechia
jan.nouza,petr.kolar,josef.chaloupka@vslib.cz

Abstract. In this paper we present our initial attempt to link speech processing technology, namely continuous speech recognition, text-to-speech synthesis and artificial talking head, with text processing techniques in order to design a Czech demonstration system that allows for informal voice chatting with virtual characters. Legendary novel figure Svejk is the first personality who can be interviewed in the recently implemented version.

1 Introduction

It is good for any research if its state-of-the-art can be demonstrated on applications that are attractive not only for a small scientific community but also for wider public. This type of application may go even beyond traditional existing or commercial areas.

A nice example of an interesting show-product demonstrating capabilities of the recent speech technology was presented on the IVVTA workshop in 1998 [1]. It showed a naturally looking voice dialogue with a virtual character. The designers of the system chose Albert Einstein as the target person to whom one could address questions about his life and work. The system employed a continuous speech recognition engine that was able to accept spoken questions and translate them into a sequence of semantic symbols. According to the list of given key-words the computer selected a (more or less) appropriate response and replayed it to the human interviewer. The Einstein's reactions had form of video sequences prerecorded by an actor. The primary aim of that demonstration was to show the current advances of the speech research, yet various applications in education and entertainment offered itself.

Being inspired by that idea we decided to develop a similar system that integrates the results of our research in speech and text processing. Recently we have been able to build up such a demo product using our middle-size vocabulary speech recognition engine applicable for continuously spoken Czech utterances [2], a TTS module developed in the Institute of Radioengineering in Prague [3] and a virtual animated face that can be synchronized with synthesized or natural speech signal [4]. We put these modules together and linked them by a simple 'chat manager'. The role of the first virtual personality was given to legendary Czech character Svejk.

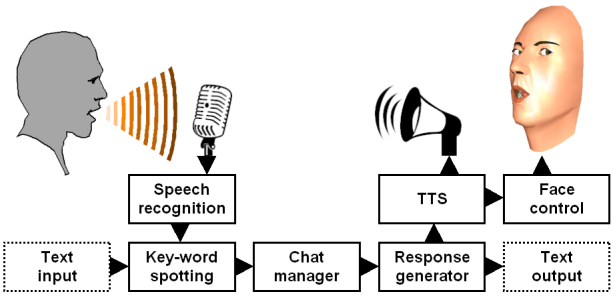

Fig. 1. Schematic diagram of the interaction between human and virtual character

2 Design of Voice Chatting System

The system is depicted in Fig. 1. Let us note its two operation modes. The simpler one uses text input and output. In this mode the user types his or her questions and receives answers again in printed form. Optionally, the virtual characters's responses can be generated by an artificial face producing synthesized speech. This mode is very helpful especially during the development and debugging phase. In the second mode the system accepts spoken input and translate it into the text form. The main components of the system are briefly described in the following subsections.

2.1 Speech input module

The module provides the speech recognition of the human user's questions. These are supposed to be fluently spoken utterances. Two options for translating them into a sequence of text symbols are available now. The first one employs a real-time key-word spotting technique [5]. It can identify about 1000 key-words while the rest of the utterance is covered by filler models. No language modeling is applied except of controlling the performance through a word/filler insertion penalty. The second type of the speech input module employs our own speech recognition engine developed for Czech language [2]. Here the speech decoding is supported by a bigram language model. In general, the latter approach outperforms the former in case the input utterances remain inside the given vocabulary and keep the rules of standard Czech.

2.2 Talking head module

This module was built up with the use of the Baldi engine [6]. We have modified and complemented the original software so that it can be applied for Czech. Moreover, it is capable of animated talking both in synthesized as well as in 'natural' mode. In the latter case it employs phoneme recognition to generate an appropriate sequence of visual patterns (visemes). The head can take on a mask (a texture) of any person.

2.3 Chat manager module

Within this particular project this is the most crucial part. In ideal case it should provide natural language (analysis and synthesis) processing capabilities. In practice our goal is more modest. The recent version supports a simplified analysis of the text string coming out from the recognition module. The string is decomposed into lexical tokens by applying a lemmatization scheme [7] and mapped onto a semantic model prepared for the given application. No disambiguation procedures neither syntactic analysis are applied at the moment. The major reason for this simplification consists in the fact that the speech recognition module is not perfect and produces errors of different kind: mainly substitutions of similarly sounding words, confusions in suffixes, false word insertions, deletions, etc. We must also except the fact that the voice interaction has form of chit-chat rather than a meaningful dialogue.

3 The Good Soldier Svejk project

The personality we have chosen for the initial case project is the legendary Czech literature character Svejk. Our choice had several pragmatic reasons. First, Svejk is so famous that its role in the project must attract both specialists as well as wide public. Second, Svejk is a true prototype of a chatting person ready to response with pleasure to any topic. His rather strange and off-topic reactions are well-known and typical for his behavior. This gives our system an exceptional chance to hide - or at least to excuse - most confusions caused by erroneous functions of the speech and language processing modules. Third, we have the complete Hasek's novel in electronic form, which allows us to make a detailed analysis of the novel text, both on lexical as well as on linguistic level. (The text contains 198,976 words, from which 33,218 are different. If we omit all words that occurred less than three-times we can reduce the working lexicon to some 8,000 words and simplify thus the recognition task.)

The chat strategy implemented so far has the following scheme: The dialogue is launched by a short introductory part, usually an exchange of greetings. After that an 'open conversation' starts. Here, the system tries to extract from the input utterance the key-words that identify any of the pre-selected themes. Some topics are covered by a single Svejk's response or comment, other cause a move into a 'target-oriented' dialogue. In the latter case Svejk starts to chat with his human partner. Within a series of questions the system tries to fill some slots in an internal form in order to select an appropriate response or story to be told by Svejk. The slots can be filled in a random order and they can handle also situations like „don't know" or „not identified". Explicit confirmation at some instants helps in recovering from unintentional detours in the chat caused by speech recognition or text analysis errors. Svejk's replies are generated from a list of prepared sentences, phrases and stories that can be partly modified according to the current context. At any dialogue turn the system checks if the user still remains in the target oriented branch or whether he or she wants to change the topic or even quit the chat. In case the system does not understand the questions or if it become confused about the themes it lets Svejk tell one of his famous stories.

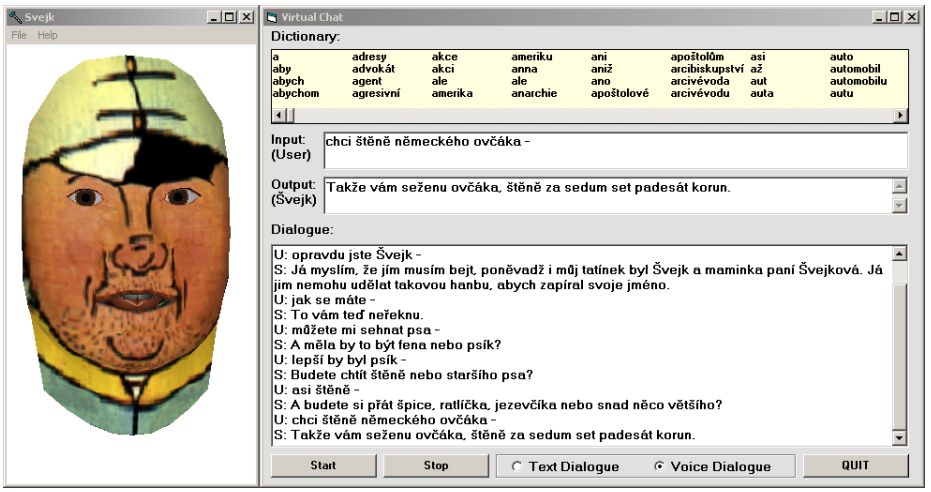

Fig. 2. Chatting with Svejk who is represented by animated talking face

4 Conclusions and Acknowledgments

The main goal of this project is to study possibilities of linking speech and text processing techniques and present them in the way that is attractive even for wider public. The project was supported by the Grant Agency of the Czech Republic (grant no.102/02/0124) and through research goal project MSM 242200001.

References

1. Furness T.A: Interactive Voice Technology in Virtual Environments. Invited talk at IVVTA (Interactive Voice Technology for Telecom. Applications) Workshop. Torino, 1998.
2. Nouza J.: Strategies for Developing a Real-Time Continuous Speech Recognition System for Czech Language. In this volume.
3. Pribil, J.: Czech and Slovak TTS System Based on the Cepstral Speech Model. In: Proc. of the 3th Int. Conference DSP '97, Herl'any (Slovakia), September 3-4, 1997, pp. 23-26.
4. Chaloupka J., Nouza J., Pribil J.: Czech-Speaking Artificial Face. In Proc. of Biosignal Conference. Brno, June 2002 (in print)
5. Nouza J.: A Scheme for Improved Key-Phrase Detection and Recognition in the InfoCity System. Proc. of 5th ECM2S workshop.Toulouse, May 2001, pp.237-241.
6. Cole et al: Intelligent Animated Agents for Interactive Language Training. Proc. of STiLL'98, Stockholm, 1998, pp.163-166.
7. Hajic, J.: Unification Morphology Grammar. PhD Thesis. Charles University, Faculty of Mathematics and Physics, Prague, 1994.