Audiovisual Interface for Czech Spoken Dialogue System

Pavel Ircing, Jan Romportl and Zdeněk Loose
Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
Email: {ircing, rompi, zloose}@kky.zcu.cz

***Abstract*—Our paper introduces implementation details of the application that serves as an audiovisual interface to the automatic dialogue system. It comprises a state-of-the-art large vocabulary continuous speech recognition engine and a TTS system coupled with an embodied avatar that is able to some extent convey a range of emotions to the user. The interface was originally designed for the dialogue system that allows elderly users to reminiscence about their photographs. However, the modular architecture of the whole system and the flexibility of messages that are used for communication between the modules facilitate seamless transition of the application to any domain of the dialogue.**

I. Introduction

The research and development in the areas of automatic speech recognition and text-to-speech synthesis is, among other incentives, driven by the aspiration to make human-computer interaction more natural and user-friendly. Especially elderly users might feel intimidated to use keyboard for interacting with computer and allowing them to talk to machine and listen to its replies might help them to overcome psychological barriers they have towards using the computer. The feel of natural conversation is usually further strengthened by an embodied representation of a virtual dialogue partner ("talking head", "avatar").

Our paper describes an application that encompasses all the modalities described above and serves as an interface between the user and the central modules of the dialogue system (such as natural language understanding, dialogue management and natural language generation). As it was designed within a research project whose aim is to develop a system that would allow (elderly) users to reminiscence about their family photographs, it also includes a window for displaying the photos and some very rudimentary photo handling capabilities. The core modules of the presented software, however, naturally constitute the automatic speech recognition (ASR), text-to-speech (TTS) and the visual avatar modules. Those will be described in more detail in the following chapters, together with the messaging framework that is used for communication throughout the system.

II. Communication Framework

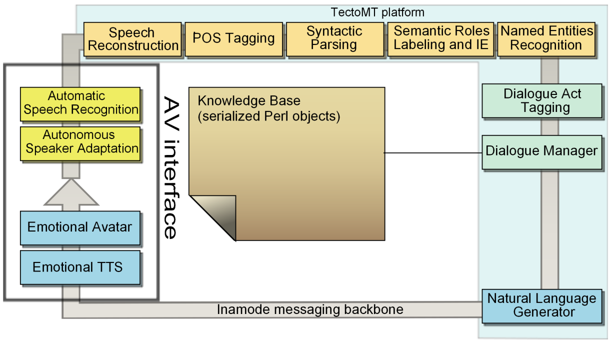

TectoMT platform
Speech Reconstruction
POS Tagging
Syntactic Parsing
Semantic Roles Labeling and IE
Named Entities Recognition
Automatic Speech Recognition
Autonomous Speaker Adaptation
Emotional Avatar
Emotional TTS
AV interface
Knowledge Base (serialized Perl objects)
Dialogue Act Tagging
Dialogue Manager
Natural Language Generator
Inamode messaging backbone

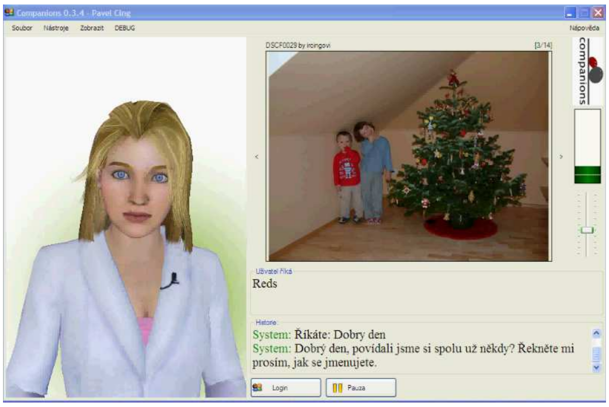
Fig. 1. *Workflow diagram.*

All the modules of the presented system communicate with each other as well as with the other components of the dialogue system through the means of the system Inamode, developed by Telefónica I+D, one of our partners in the project. It is basically a system comprising a central messaging hub and flexible mechanisms for information flow control. These mechanisms parse XML messages coming to the hub's input socket (via the TCP/IP protocol) from other modules and transduce them to output messages which are then sent back to the modules. This allows control and synchronization of activity of the modules in a heterogeneous environment where every module can run on a different computer and system.

Inamode thus creates and controls the workflow of the dialogue system in a form of a loop schematically illustrated by the figure 1. The modules constituting the presented audiovisual interface are shown on the left side of the picture — the rest of the dialogue system modules is grouped under the TectoMT platform developed by our project partners [1]. However, this loop represents only time-wise structuring of particular activities and does not limit communication channels among the modules. Any module can send a message to any other module at any time, not being limited by the cyclic communication. More specifically: every module can send its XML message to the central hub where this message is transduced accordingly and then resent to all other modules; any module potentially interested in this message can then accept it.

III. Audiovisual Interface Architecture

The audiovisual interface application is written in C# and runs under the Windows XP operating system. The screenshot is shown in Figure 2 — it contains (clockwise from left)



Reds

System: Říkáte: Dobry den
System: Dobrý den, povídali jsme si spolu už někdy? Řekněte mi prosím, jak se jmenujete.

Fig. 2. *Application screenshot.*

the embedded Internet Explorer window with the embodied avatar (see the section V for details), the window with a photograph that is being discussed and finally two windows showing the current user/avatar utterance (depending on whose conversation turn is actually taking place) and the history of dialogue turns.

Each user of the audiovisual interface has its own profile that allows to store multiple user-specific settings (such as the set of photographs, fonts used in the interface windows, etc.) and, most importantly, points to the personal acoustic model set that is incrementally adapted to each individual user by an unsupervised speaker adaptation technique described in the following chapter.

IV. ASR ENGINE

The speech recognition engine embedded in the interface is an in-house software based on the state-of-the-art probabilistic ASR technologies [2]. Its acoustic model employs Hidden Markov Models (HMMs) trained using large speech corpora (over 220 hours of transcribed speech, more than 700 speakers), which grants robust acoustic recognition independent of the speaker. However, since the described application happens to be inherently personalized, there are actually mechanisms for tailoring the acoustic model to each user.

First, depending on the gender of the user, the general male or female acoustic model set is selected. This baseline model is then gradually adapted to a user by our original unsupervised speaker adaptation procedure [3] that is completely hidden from the user and does not require any explicit training session when the user is forced to read predefined adaptation sentences. In our adaptation process, on the contrary, all recognized word sequences that have high-enough confidence scores serve as the adaptation data. The ASR engine accumulates statistics from this “confidently recognized” portions and once the amount of accumulated data exceeds an empirically specified threshold, a new set of feature MLLR adaptation matrices is computed and those matrices are then used to transform the input speech vectors from that point onwards. Several iterations of this adaption procedure can be performed, depending on the amount of accumulated adaptation data. Moreover, further iteration of the adaptation algorithm is run when the user closes a dialogue session.

Another notable feature of our ASR system is the speech decoder itself. It is currently able to handle a lexicon comprising more than half a million words. It may seem unnecessarily too much but bear in mind that every word form is treated as a separate lexicon entry in the ASR system and that the Czech as a highly inflectional Slavic language produces abundance of word forms for a single lemma. The decoder works with standard n-gram language models and allows 2-pass recognition process where only the bigram model is employed in the first pass and the resulting lattices are rescored with higher-order n-grams in the second pass. Note that despite of this 2-pass technique the decoder still operates within the real-time constraint [4]. Since the domain of the dialogue system that our application interfaces to is not very well-defined from the lexical point of view (users could reminiscence about their Christmas photos as well as the photos documenting their trip around the world) we wanted our language model to have as broad a coverage as possible. Thus we have used general language model with a lexicon containing 300 thousand words, trained using newspaper articles (313 million tokens) and TV news transcripts (14 million tokens).

The current version of the communication interface operates in “closed microphone” mode — that is, the user and the system are not allowed to interrupt each other and have to wait until the other finishes its utterance. One turn of the dialogue therefore looks as follows:

1) The ASR system starts recognition (microphone is opened), transcribed text is continuously being displayed in the application window.
2) Once the silence detector spots a long-enough pause in user’s speech, it closes the microphone and sends the Inamode message to the remote dialogue manager (DM) application [1]. Example of the message is shown in the table I.
3) DM processes the message, generates another message for the TTS and avatar module (see the section V for details) that contains the dialogue system response and the text of this response also displays in the application window.
4) After the avatar finishes the response, the ASR system is started again.

V. TTS AND EMBODIED AVATAR

As it was already mentioned, DM produces a message and sends it to the central messaging hub. This message contains full information about the generated response of the dialogue system:

[1] This DM application includes, besides the actual dialogue manager, also quite sophisticated natural language understanding and natural language generation modules (see Figure 1). These modules, however important and interesting, are outside the scope of this paper. It will be referred to as DM simply for the sake of readability

TABLE I
EXAMPLE OF AN ASR MESSAGE

```
<message zone_id="companion_cz"
 sender="eca"
 message_id="eca_1"
 message_type="asr_nbest">
<payload>
<emma:emma version="1.0"
 <emma:one-of id="nbest"
 emma:medium="acoustic"
 emma:mode="voice"
 emma:start="1426" emma:end="9442">
<emma:interpretation id="alternative1"
 emma:confidence="1">
<answer>
na teto fotce jsem ja a moje deti.
</answer>
</emma:interpretation>
</emma:one-of>
 </emma:emma>
 <meta from="zdenek@loose/pilsen" />
</payload>
</message>
```

TABLE II
EXAMPLE OF A DM MESSAGE

```
<message zone_id="companion_cz"
 sender="nlgmain"
 message_id="nlgmain_1251622550263">
<payload>
 <NLGProxy
  application="senior_companion_cz"
  place="LIVING_ROOM"
  embodiment="petra"
  perform="DEFAULT"
  emphasize="weak"
  affect="neutral"
  gesture="main"
  communicative_function="SURPRISE |
HAPPY-EMPATHY"
  text="Vy mate deset vnoucat? |
No to je uzasne."
/>
</payload>
</message>
```

1) The text of all sentences in the generated response of the system. DM generates a whole turn at once and sends all its sentences for further processing in one message.
2) Communicative functions of the generated sentences. Communicative functions are the way in which we deal with emotional speech synthesis. Motivation for their usage is discussed in [5] and the process of recording a speech corpus for emotional unit selection speech synthesis driven by communicative functions is described in [6]. DM is set up on the basis of this corpus and by its internal mechanisms it is able to assign an appropriate communicative function to each generated sentence.
3) Identification of a particular ECA. The system can potentially handle more independent conversations in parallel, moreover with ECAs in different environments. Therefore, DM must determine which ECA is supposed to accept the generated message.
4) Potential visual gestures of the avatar. The avatar control mechanism is able to transform communicative functions into suitable visual gestures. However, in some cases DM might want to supersede these implicit gestures, and then it indicates them in the generated message.

An example of such a message is in the table II. This message is accepted and processed by a proxy module which coordinates activity of two other modules: a TTS system and an avatar.

The TTS system is based on the Czech state-of-the-art speech synthesis system ARTIC which employs a unit selection synthesis technique [7]. This system delivers highly natural speech by concatenating speech units without signal modifications. The default version generates neutral “news-reading” voice which must be enhanced by some expressiveness demanded by its involvement in the natural dialogue system. This is achieved by utilizing the special expressive corpus annotated by communicative functions [6] and using communicative functions as features in the target-cost function of the unit selection algorithm.

Apart from the TTS system ECA comprises a visual avatar. The interface makes use of a 3rd party avatar — a graphical visualization of a female head and torso with the capability of articulation, facial expressions and body gestures which are triggered by special commands with both categorical and continuous parameters. A window with the avatar can be embedded in a web browser, which means that this virtual character is able to run in various environments and modes of usage.

The avatar as well as the TTS system are connected to the central messaging hub, as described in the section II. Both modules accept incoming messages from and send outgoing messages to the aforementioned proxy module which coordinates their activity. This proxy module accepts DM output messages and parallelizes the workflow by allowing the TTS system and the avatar to work simultaneously: TTS synthesizes sentences one by one and sends them to the input buffer of the avatar while the avatar is playing them stepwise. This measure is employed to minimize the latency of the system’s response. The proxy module also communicates with a gesture module in order to transform the communicative functions assigned by DM into appropriate commands triggering gestures and facial expressions of the avatar (these gestures, however, can be potentially superseded by requirements of DM). This process involves certain randomness, thus the visual behavior of the ECA is more natural.

Synchronization of the avatar’s articulation with synthesized speech is ensured by the TTS module which generates an accompanying string of phones and their timestamps. Since the avatar had been originally designed for English and Spanish, it normally utilizes a different set of visemes for its articulation. Therefore, the TTS system converts the string of Czech phones into its best estimation by a string of Spanish viseme codes. This final string is then aligned with the respective audio and streamed to the avatar in the Ogg Vorbis format.

VI. Evaluation

The two core components of our multimodal interface require rather different evaluation strategies. The performance of the ASR system can be rather easily measured using the standard Word Error Rate (WER) given that there are manually transcribed evaluation data available. We have actually gathered over 60 hours of interviews with 65 elderly people discussing their photographs within the Wizard of Oz experimental setting (see [6] for details). This data were employed for various purposes in the process of the dialogue system development, such as examining the reactions of the elderly subjects to the dialogue with what they believe was a computer program, inspiring some basic structures of the transition network used in the dialogue manager and finding the operator's (wizard's) responses that were best at keeping the conversation going. We have also tried to use this data for the ASR testing but unfortunately it turned out that their manual transcription is slightly inconsistent with the text data that were used for language modeling. However, we have decided to use it at least for evaluating the effect of the speaker adaptation since those language model and test data discrepancies can be expected to hurt the performance of both baseline and adapted acoustic models equally and thus the relative difference in the WER would show the effect of the adaptation. We have performed two sets of adaptation experiments. In the first one, we started recognition of each speaker's utterances with the general (baseline) acoustic models and gradually adapted them using the procedure described in the section IV. This means that the adaptation actually took place only after processing certain portion of the user's utterances, depending on the speed of appropriate adaptation data accumulation. In case of some speakers the amount of adaptation data did not exceed the threshold even after exhausting all the available sentences and thus the only adaptation iteration was forced at the end of the recognition. The average WER across all speakers improved by 0.86% absolute. In the second experimental setup, we used the adapted models obtained in the first set of experiments to recognize all the utterances of individual speakers again. The adaptation algorithm was switched off in this run. Using the adapted models from the beginning improved the WER by 1.36% over the previous incremental adaptation (that is, 2.22% over the baseline). The unsupervised speaker adaptation technique thus seems to be quite efficient.

On the other hand, the evaluation of the avatar's performance is more complicated and generally must be measured on the user-response basis. Emotional TTS synthesis will be evaluated separately from the whole dialogue system by listening tests. These listening tests will follow the scheme we have developed for prosodic phrase and semantic accent annotation, including statistical modeling of the results using the maximum likelihood approach [8]. The overall naturalness of audiovisual experience resulting from the TTS and avatar activity can be measured only indirectly by intersubjective assessments of testing users from among seniors. It still remains to be decided whether such tests should be performed within the Wizard-of-Oz framework (so as to exclude the influence of ASR and DM performance), or rather using the whole dialogue system. We can only say at this point that the avatar's speech, gestures and the overall appearance is general was well-perceived by the audience on several demonstration sessions (one of them being the COMPANIONS project annual review) but of course this evaluation is by no means thorough enough.

VII. Conclusions

The work presented in this paper shows how the cutting-edge speech and embodied avatar technologies can be used in a real-time application that has a good potential to make human-machine conversation more natural and user-friendly. The research and development of the system has also confirmed that the Senior Companion scenario is very interesting and promising, not only by its cultural and social potential, but also by its suitability as a laboratory scenario for speech and dialogue technology research. The dialogue domain is reasonably predetermined so that it fits the upper limits of the state-of-the-art spoken dialogue systems; moreover, the target group of users is very appreciative and talkative, which makes it easier to maintain the dialogue dynamics. On the other hand it has shown that even under such circumstances the domain cannot be restricted at all because the senior users often talked about so much different topics that the limits of the state-of-the art must by pushed forward so as to handle these communication situations.

Acknowledgements

This work was funded by the COMPANIONS project (www.companions-project.org — EC grant number IST-FP6-034434) and partially also by the UWB internal project SGS-2010-054.

References

[1] Z. Žabokrtský, J. Ptáček, and P. Pajas, "TectoMT: highly modular MT system with tectogrammatics used as transfer layer," in *StatMT '08: The Third Workshop on Statistical Machine Translation*. Morristown, NJ, USA: ACL, 2008, pp. 167–170.

[2] A. Pražák, J. Psutka, J. Hoidekr, J. Kanis, L. Müller, and J. Psutka, "Automatic online subtitling of the czech parliament meetings," *Lecture Notes in Artificial Intelligence*, vol. 4188, pp. 501–508, 2006.

[3] Z. Zajíc, L. Machlica, and L. Müller, "Refinement approach for adaptation based on combination of MAP and fMLLR," *Lecture Notes in Artificial Intelligence*, vol. 5729, pp. 274–281, 2009.

[4] A. Pražák, L. Müller, J. Psutka, and J. Psutka, "Live tv subtitling - fast 2-pass LVCSR system for online subtitling," in *SIGMAP 2007*, Lisabon, 2007, pp. 139–142.

[5] E. Zovato and J. Romportl, "Speech synthesis and emotions: a compromise between flexibility and believability," in *Proceedings of 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[6] M. Grůber, M. Legát, P. Ircing, J. Romportl, and J. Psutka, "Czech senior COMPANION: Wizard of oz data collection and expressive speech corpus recording," in *LTC 09*, Poznan, Poland, 2009, pp. 266–269.

[7] D. Tihelka and J. Matoušek, "Unit selection and its relation to symbolic prosody: a new approach," in *INTERSPEECH 2006 – ICSLP*, vol. 1. Bonn: ISCA, 2006, pp. 2042–2045.

[8] J. Romportl, "Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis," *Lecture Notes in Artificial Intelligence*, vol. 5246, pp. 493–500, 2008.