Design and Development of Spoken Dialog Systems Incorporating Speech Synthesis of Viennese Varieties

Michael Pucher[1], Friedrich Neubarth[2], and Dietmar Schabus[1]

[1] Telecommunications Research Center Vienna (FTW),
Donau City Str. 1, 1220 Vienna, Austria
{pucher,schabus}@ftw.at

[2] Austrian Research Institute for Artificial Intelligence (OFAI),
Freyung 6/6, 1010 Vienna, Austria
friedrich.neubarth@ofai.at

Abstract. This paper describes our work on the design and development of a spoken dialog system, which uses synthesized speech of various different Viennese varieties. In a previous study we investigated the usefulness of synthesis of varieties. The developed spoken dialog system was especially designed for the different personas that can be realized with multiple varieties. This brings more realistic and fun-to-use spoken dialog systems to the end user and can serve as speech-based user interface for blind users and users with visual impairment. The benefits for this group of users are the increased acceptability and also comprehensibility that comes about when the synthesized speech reflects the user's linguistic and/or social identity.

Keywords: Spoken dialog system, speech synthesis, dialect.

1 Introduction

Speech varieties can be regarded as a primary means for expressing a persons social affiliation and identity. In the context of human computer interaction (HCI) future applications that use speech as their main modality must be capable of reproducing different language varieties. As noted in [2] HCI must shift towards a paradigm where technologies are available to support the production of culture and identity. This is even more important as computers enter everyday life in a more ubiquitous way. Future computers will not only serve as tools or provide information, but will also be monitoring, participating and communicating with users whose experience with such situations and interfaces may be low. Systems that can adapt to the user by displaying a familiar cultural and social identity will definitely increase the acceptance of such systems.

A second aspect, specifically relevant for blind users or users with visual impairment, is that the communicative modality based on spoken language has to serve the needs for both, fast information transmission (fast speech) and comprehensibility. If a user with little experience is confronted with a system that

K. Miesenberger et al. (Eds.): ICCHP 2010, Part II, LNCS 6180, pp. 361–366, 2010.

speaks to her in a language variety that is not familiar to her, comprehensibility is likely to decrease, on a par with acceptability. In the case of a pluricentric language like German, this is obviously the case - at present interfaces using synthetic speech are available, but only generating speech reflecting the standard language spoken in Germany, not the standard spoken in Austria. Non standard varieties are normally not even taken into consideration.

Within the project ”Viennese Sociolect and Dialect Synthesis” [1] we developed four synthetic voices, each of them representing a linguistic variety spoken in Vienna. The voice representing Standard Austrian German is already integrated into a web-reader application.

2 Persona Design for Spoken Dialog Systems

2.1 Spoken Dialog Systems

Spoken dialog systems are computer programs that allow a user to interact with a system using speech. Spoken dialog systems are the most advanced form of voice user interfaces (VUI) [3] since they allow for full speech interaction. These systems are composed of three main components: speech synthesis for generating speech output, speech recognition for processing the acoustic input, and dialog management. Equipped with speech synthesis the dialog system is able to transform text available in written form into spoken language. Speech recognition is employed to transform spoken user utterances into written text or words using grammars. The speech recognition component can also be extended by a natural language understanding component. The dialog management component defines the interaction behavior or dialog logic of the dialog system. A standardized definition language for spoken dialog systems is VoiceXML [7], a markup language which is built around the web-based form filling paradigm. The main components of a VoiceXML application are prompts (define what is said), grammars (define what can be said), and forms (define the dialog logic). Figure 1 in section 3 shows a state diagram of our dialog system. All green blocks are defined within one VoiceXML page using four different forms. One form for the selection of the restaurant and a separate form for each restaurant. Within a restaurant form it is possible to make a phone call to the selected restaurant.

2.2 Persona Design

The personality or persona of a spoken dialog system must be considered , since there is no such thing as a voice user interface with no personality [3]. The perception of sociolectal and dialectal varieties influences our evaluation of a speaker’s attributes like competence, intelligence, and friendliness. The persona can be defined as a standardized mental image of a personality or character that users infer from the applications voice and language choices [3]. Speech synthesis is an essential part of a spoken dialog system’s persona. In a previous study [4] we found the following positive and negative attributes that are associated with sociolect / dialect voices. It should be mentioned that this study was carried

Table 1. Positive and negative attributes of sociolect / dialect voices

Positive (application)	Negative (application)
fun (game)	not respectable (banking)
optional (navigation)	not formal (banking)
personal (navigation)	not neutral (administration)
regional (taxi)	not intelligible (flight domain)
democratic (administration)	non-native spk. (flight domain)
persona (taxi)	lack of precision (health)
tourist (district info)	lack of prestige (banking)
lack of trust (administration)	lack of competence (health)

out with special focus on varieties within Austria, but it can be assumed that the results will match users' judgments about many dialect voices that stand in opposition to an officially recognized standard variety.

3 Spoken Dialog System

Within our research project on synthesis of Viennese varieties [1] we developed four synthetic voices that represent representative points in a 3-dimensional space of language varieties, which is defined through age, gender, and education [6].

1. Austrian German standard (35–50, male, +)
2. Viennese dialect (45–60, male, –)
3. Viennese youth language (15–25, female, +/-)
4. Viennese standard (55–70, female, +)

On the basis of this selection of prototypes, we created four different personas within our spoken dialog system. In the evaluation of possible scenarios [4] we found that a restaurant information system is well suited for Viennese dialect synthesis. The mapping of positive / negative properties to standard / dialect served as design guideline for the dialog design. The speaker of the Austrian standard variety (1) was used as a moderator, also providing some help functionality. Every other speaker recommends a different type of restaurant, which is associated with the speaker's sociolect.

Figure 1 shows the VoiceXML [7] dialog flow diagram for our regionalized and localized restaurant guide. Speaker specific prompts were used for the dialog

Table 2. Sociolect / restaurant type association

Speaker sociolect	Restaurant type
2. Viennese dialect	Viennese cooking
3. Viennese youth language	Low prices / cool places
4. Viennese standard	Luxury restaurants

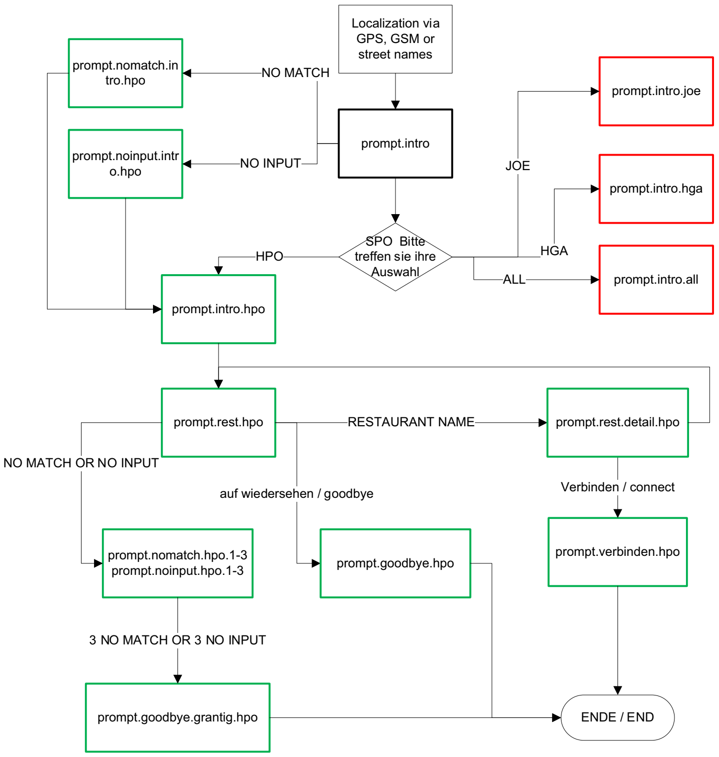

Fig. 1. VoiceXML dialog flow

and no match / no input strategies. One additional dialog with all speakers and interactions between them was also realized. On our project webpage [1] it is possible to try out all four synthetic voices as well as the dialog system.

4 Blending between Standard and Dialect

To enable applications that use continuous varieties between standard and dialect we have investigated interpolation methods that can realize such in-between varieties [5]. The blending between an Austrian German and Viennese utterance is shown in Figure 2. This can be realized using the parametric synthesis method of Hidden Markov Model (HMM) based speech synthesis. With this method it is possible to interpolate models of different varieties and thereby create in-between varieties. By interpolating the duration model with a zero duration model we can also model units that are present in one variety but not in the other (such as vowel deletion/epenthesis) as shown in Figure 2. To model context-dependency of units a broad linguistic context is taken into account consisting of linguistic features on different linguistic levels.

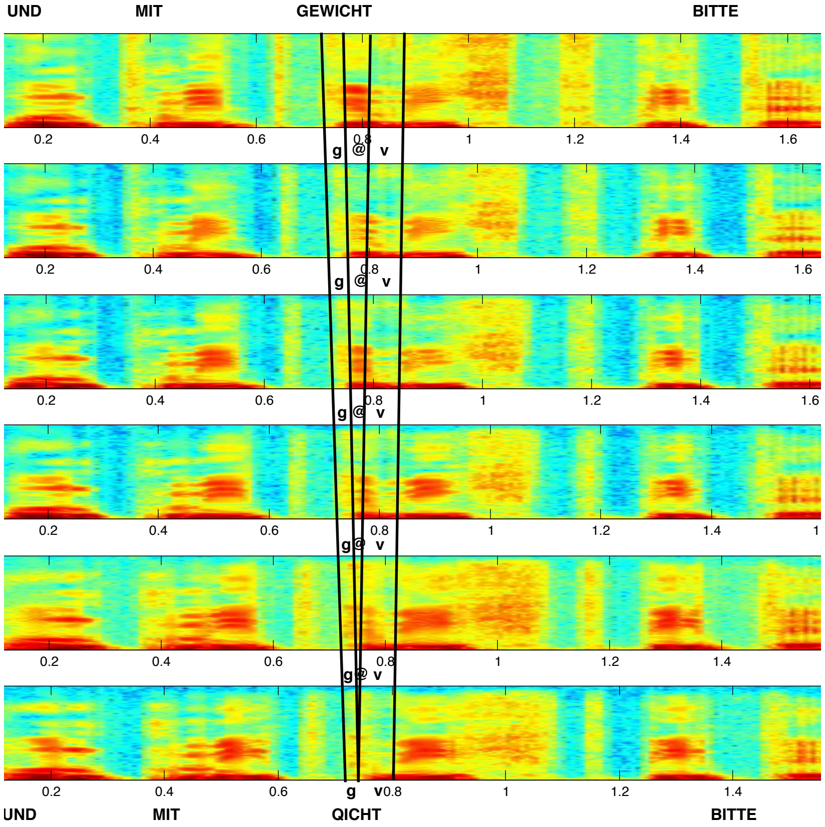

Fig. 2. Blending between German "Und mit Gewicht bitte" and Viennese "Und mit Qicht bitte"

These blending methods could be used in a dialect / sociolect adaptive spoken dialog system. By using dialect recognition there could be even a matching of the speaker's dialect and the dialect of the spoken dialog system. In this way it would be possible to gradually adapt to the user's dialect. The variety of the dialog system could be chosen to be similar [8] or different from the speaker's variety. [8] shows that users prefer accents that are similar to their own accent. This, however, is not necessarily true for dialects or sociolects.

5 Conclusion

We discussed the importance of regionalized adaptive user interfaces and showed the design of a spoken dialog system that incorporates multiple varieties which represent the sociolect space of Vienna. In future work we want to develop additional spoken dialog systems that are suitable for sociolect / dialect synthesis.

Furthermore we also want to investigate less known varieties like Turkish-Viennese, which enable new personas.

Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" was funded by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET. OFAI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology and by the Austrian Federal Ministry for Science and Research.

References

1. Viennese Sociolect and Dialect Synthesis, `http://dialect-tts.ftw.at`
2. Cassell, J.: Social Practice: Becoming Enculturated in Human-Computer Interaction. In: Stephanidis, C. (ed.) Universal Access in HCI (UAHCI), HCI 2009. LNCS, vol. 5616, pp. 303–313. Springer, Heidelberg (2009)
3. Cohen, M.H., Giangola, J.P., Balogh, J.: Voice user interface design. Addison-Wesley, Reading (2004)
4. Pucher, M., Schuchmann, G., Fröhlich, P.: Regionalized Text-to-Speech Systems: Persona Design and Application Scenarios. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) Multimodal Signals: Cognitive and Algorithmic Issues. LNCS (LNAI), vol. 5398, pp. 216–222. Springer, Heidelberg (2009)
5. Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V.: Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis. Speech Communication 52(2), 164–179 (2010)
6. Moosmüller, S.: Soziophonologische Variation im gegenwärtigen Wiener Deutsch. Franz Steiner Verlag, Stuttgart (1987)
7. VoiceXML 2.0 recommendation, `http://www.w3.org/TR/voicexml20/`
8. Dahlbäck, N., Wang, Q., Nass, C., Alwin, J.: Similarity is more important than expertise: Accent effects in speech interfaces. In: Proc. SIGCHI conference on human factors in computing systems, pp. 1553–1556 (2007)