Regionalized Text-to-Speech Systems: Persona Design and Application Scenarios

Michael Pucher, Gudrun Schuchmann, and Peter Fröhlich

ftw., Telecommunications Research Center, Donau-City-Strasse 1,
1220 Vienna, Austria
pucher@ftw.at, schuchmann@ftw.at, froehlich@ftw.at

Abstract. This paper presents results on the selection of application scenarios and persona design for sociolect and dialect speech synthesis. These results are derived from a listening experiment and a user study. Most speech synthesis applications focus on major languages that are spoken by many people. We think that the localization of speech synthesis applications by using sociolects and dialects can be beneficial for the user since these language variants entail specific personas and background knowledge.

Keywords: speech synthesis, dialect, sociolect, persona design.

1 Introduction

When we hear a speaker with a certain dialect we are familiar with we think that this speaker has a certain regional knowledge of the region that is associated with this dialect. This is demonstrated through the case of a taxi reservation system, which was launched in Austria. Users could order taxis for Vienna while the actual call center was located in the region of Salzburg. Since some of the call center agents spoke like people from Salzburg, users did not think that they were competent enough to know the streets of Vienna.

This example is another manifestation of the similarity attraction effect. [1] found the similarity attraction effect for language accents, where users preferred voices with a similar accent to their own. [1] could show that synthetic voices were more attractive, when they had the same gender and personality as the speaker. Previous studies did not entail dialects and sociolects [1]. Some work has been performed on cultural dimensions in user interface design were it is assumed that the language part of a user interface can be translated from one language to the other [2]. Such an assumption must necessarily neglect information that is encoded in a specific voice.

Spoken dialog systems that are using text-to-speech (TTS) technology are successfully deployed in a number of application domains like voice-web portals, online banking, and taxi service [3]. State-of-the-art data driven speech synthesis methods can achieve natural sounding speech output. The data-driven unit selection method can be used with small [5] and large speech corpora [5]. We think that these methods can be used to adapt speech synthesis to dialects and sociolects.

In our project on sociolect and dialect speech synthesis we aim at exploiting the effects of localization or regionalization [2] of voice user interfaces also for TTS

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 216–222, 2009.

systems. We are developing synthetic voices for different sociolects, first contextualizing it in the region of Vienna. This is the first attempt to develop multiple synthetic voices that represent the space of sociolects for a certain language. To achieve this goal we will develop 3 synthetic sociolect voices with speakers from Vienna. The adaptation methods and application scenarios that we develop are applicable to synthetic voices for dialects and sociolects of other languages.

The TTS paradigm of producing spoken output from written text is of course problematic for synthesizing dialects and sociolects. These language variants do not have a standardized written form, which means that it is not possible to apply standard TTS methods directly. In our work on building such voices we develop methods that allow us to use resources from standard German and apply those for Viennese variants. This is also a reason why the development of synthetic dialect voices has not been investigated intensively. Another reason is of course that some dialects are only spoken by relatively small groups of people. This is however not generally true if one thinks of Bengal or Arabic dialects. In speech recognition there has been work on the dialectal adaptation of speech recognizers [6].

Persona design is used to enhance the user experience of spoken dialog systems [8]. With this design method system prompts, dialogs, and system behavior can be designed in a way such that a consistent persona is perceived by the user. The choice of a synthetic voice constrains the type of persona that can be realized with this voice. The synthetic voices have a certain gender, age, dialect, sociolect, and voice characteristic. An application for electronic banking for example excludes voices with a certian age and sociolect. The persona underlying such an application has to be mature and earnest. For an entertainment application on the other hand these excluded voices may produce a more realistic persona. The question when to use and when not to use dialect and sociolect synthetic voices is addressed in this paper. Furthermore these voices allow for the extension of the standard user models, which represent the well-educated, adult, middle-aged, computer-literate, male user. A persona that is often implicit in synthetic voices of standard language.

We performed listening experiments to find salient features in voices of Viennese speakers that define the sociolect space for persona design. Furthermore a user study was conducted to find application scenarios appropriate for dialect/sociolect TTS.

2 What's in a Voice: Salient Voice Features for Persona Design

The method of unit selection speech synthesis uses a large corpus of recorded speech of a specific speaker. It achieves a high degree of naturalness, but it is difficult to change speaker and voice characteristics, such that the range of realizable personas is constrained. For this reason the selection of the right speaker is crucial. To represent the sociolect space for synthetic voices it is necessary to determine the dimensions of this space. [8] shows age and education as fundamental sociolect dimensions for Viennese German. [8] found acoustic correlates for these dimensions. We investigated the validity of these dimensions through a listening experiment.

In our listening experiments we evaluated what kind of information might be retrieved from user utterances of Viennese speakers. We carried out a perception test

Table 1. Accuracy of speaker/dimension association (in percent) and correlation between estimated and "real" values

Accuracy	*Mean*	*Min*	*Max*	*Correlation*
Age	0.55	0.88	0.12	0.85, p<0.01
Gender	0.94	1.0	0.88	0.82, p<0.01
Education	0.57	0.81	0.31	0.61, p<0.01
District	0.08	0.19	0	Not significant
Std. German	0.54	0.96	0.12	Not significant
Coll. Language	0.58	0.96	0.04	Not significant
Dialect	0.61	0.92	0.15	0.21, p<0.01

with 9 voices perceived by 26 listeners to see whether the parameters in [8] are also perceivable. And in order to rank these aspects in importance to our speaker selection process. Tokens of 5 – 10 sec were played twice and listeners had to put down their judgment of *age*, *gender*, *district*, and the speaker's claimed competence in *standard German* ("Hochdeutsch"), *colloquial language* ("Umgangssprache") and *dialect*. The group of speakers consisted of 6 females and 3 males, aged from 17 to 80 (average 42,0), they were distributed with respect to educational level (3 university degree holders, 4 A-levels, and 2 below A-levels) and district (8 districts in 9 subjects). The group of listeners consisted of 15 males an 11 females, average age 31 (min: 16, max: 59) , who were rather highly educated (20 university degree holders, 2 A-levels, 4 below A-levels). At large there is a wide-spread though not representative coverage of the dimensions under investigation.

The listeners were asked to guess the age, gender, educational level, and district where the speaker comes from. This was compared to the reality. For the dimensions associated with language we did use a subjective measure because we are mainly interested in the stereotypes of language perception. So we decided to differentiate between self-perception and perception-by-others. The listeners were asked which type of language the speakers stated they were able to speak: Standard German (*Hochdeutsch*), colloquial language (*Umgangssprache*), and dialect (*Dialekt*). This was compared to the self-perception of the speaker.

Table 1 shows the accuracy for the different dimensions. The *min / max* values are the percentages of the best/worst estimated speakers for this dimension. For the *age* dimension we accepted judgments within an interval of 15%(+/-) as correct. There are significant correlations for real and estimated *age*, *gender*, *education*, and *dialect*, while there are no correlations for *district*, *Standard German* and *colloquial language*.

2.1 Implications for Persona Design

The correlations within *age* and *educational level* are in accordance with [8]. This shows a tendency towards these dimensions as basic dimensions of Viennese sociolects. The lack of correlation within the *district* dimension contrasts with the cliché of regional differences in Vienna.

Taking into account the correlations within the *gender* dimension we decided to realize 3 personas/voices that represent this 3-dimensional (*age*, *gender*, *education*) sociolect space.

Concerning the language dimension there is only a weak but significant correlation within the dialect dimension. These results show that self-perception and perception-by-others concerning language competence do not match and that these features are not reliable cues for speaker and persona selection.

3 Application Scenarios for Regionalized Voice Interfaces with Sociolect and Dialect Speech Synthesis

The user study aims to select appropriate application scenarios for a dialect speech synthesis system. We conducted 26 interviews with experts and non-experts regarding our research objective. Participants were aged between 17 and 80 (mean 39,6), 12 male 14 female. The 11 experts were from various fields: media art, usability, computer science, sociolinguistics, dialect phonetics.

Five application scenarios were presented to the users who had to decide if a sociolect/dialect voice is appropriate. Then they were encouraged to comment on the decision and to suggest further application scenarios. The user comments were merged into 9 negative and 7 positive classes (Figure 3 and 4). The application scenarios were a *district information system*, a *talking clock*, a *computer game*, an *application for administrative information*, and a *health information system*. As can be seen from

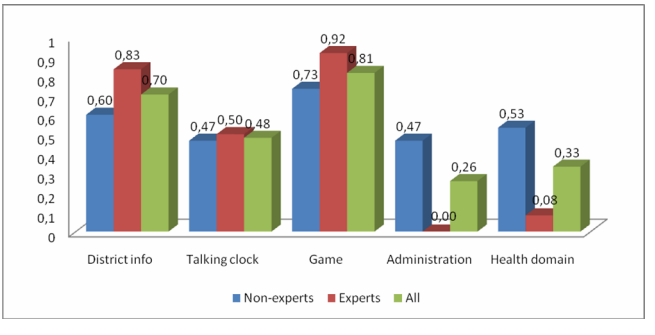

Fig. 1. Application scenarios for dialect voices

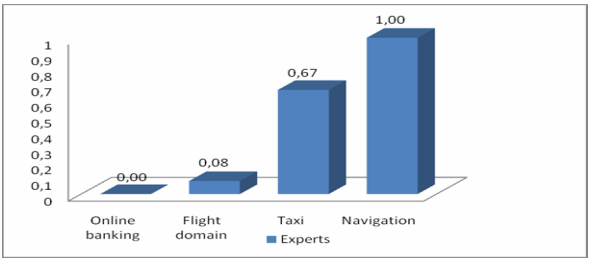

Fig. 2. More application scenarios for dialect voices

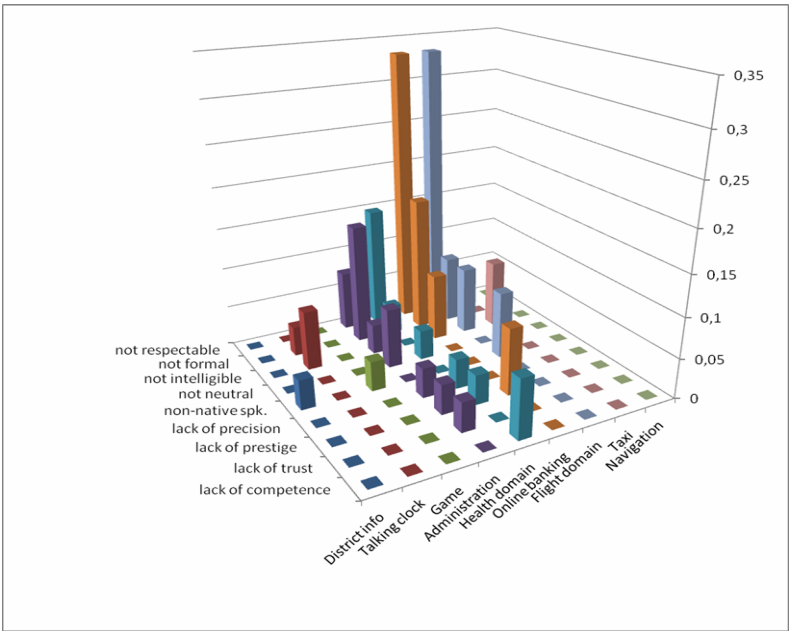

Fig. 3. Reasons to reject dialect voices in these scenarios

Fig. 1 there was a tendency to distinguish between two types of applications with *game*, *district info*, and *talking clock* being in the one group and *administration* and *health domain* being in the other group. A grouping can also be found with four more application scenarios that we evaluated with the expert users (Fig. 2). The *online banking* and the *flight domain* application were found unacceptable by the expert users.

Fig. 3 and Fig. 4 show the attributes that were chosen for applying or rejecting dialect voices. The most frequent reasons for rejecting a dialect persona were that it is *not respectable* and *not formal* enough in that situation. The reasons to apply dialect voices are *fun* (*game*, *navigation*), optional (*district info*, *talking clock*, *navigation*), *personal* (*navigation*), *regional* (*district info*, *taxi*, *navigation*), and that they are able to represent the adequate persona as in the *taxi* reservation system, where the dialect voice may represent the taxi driver.

Application scenarios for sociolect and dialect speech synthesis that were suggested by the users were an application for a site of Vienna's local government, for reserving barbecue places, for audio books of Viennese songs and stories, and for Second Life avatars.

The main difference between experts and non-experts is that the latter are not as consistent in their judgments as the former. For the dialect voices in the *health domain* for example there was only one expert who voted for it. The positive attribute for this expert was *democratic* (*BürgerInnennähe*) which means that a dialect persona creates familiarity between patient and doctor.

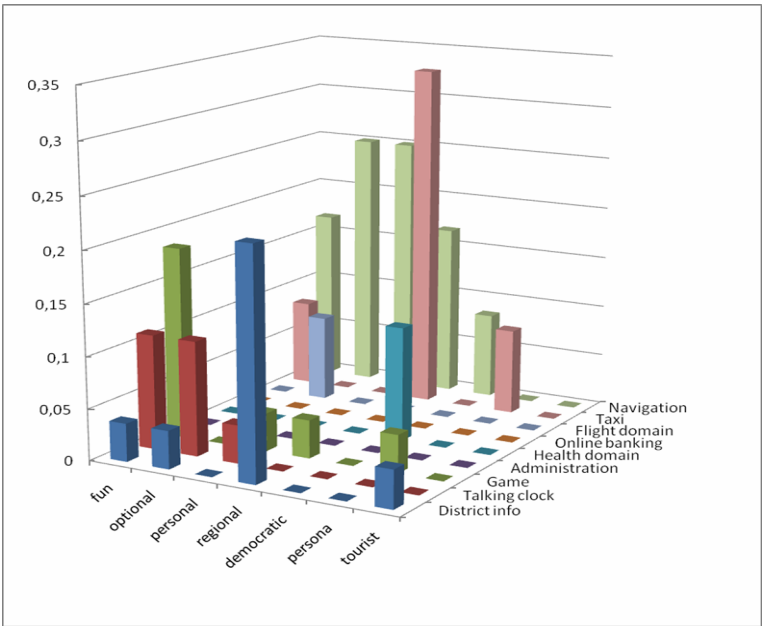

Fig. 4. Reasons to apply dialect voices in these scenarios

3.1 Implications for Application Scenarios

For the *district info*, *talking clock*, *game*, *taxi*, and *navigation* applications we got positive feedback. This shows that there is a range of scenarios where the dialect voices can be deployed. The *optional* attribute shows that an adaptive voice user interface that allows for switching between a dialect and a standard persona is valued.

We also saw that there is a class of applications where a *respectable* and *formal* persona is required. A dialect voice is not considered appropriate for this persona.

4 Conclusions

Our listening experiments showed that age, gender, and educational level are the appropriate dimensions for defining the sociolects in Vienna. Considering these dimensions speakers can be selected to representatively cover the space of sociolects.

Users consider dialect TTS applications *fun*, *regional*, and *personal*. But they reject them when a *respectable* persona is expected. Additional application scenarios can incorporate dialect personas if they are *optional* to a standard persona.

The above results support us in developing multiple synthetic voices, which realize different personas representing the Viennese sociolect space. Therefore the target personas are going to cover both genders, three different age groups, and three sociolects namely Viennese dialect, Viennese youth language (slang), and Viennese standard German.

In our future work we will evaluate our released synthetic voices in a similar study such that a comparison between natural and synthetic dialect speech is possible. Since we will release these voices under an open source license we encourage other researchers to perform further studies.

One open problem for unit selection speech synthesis is the synthesis of emotional speech. We think that dialect and sociolect speech synthesis evokes emotions in the listener because only a specific social group is associated with the persona, and each other social group has specific perception patterns for this persona. For these reasons one has to be considerate when realizing a sociolect speech synthesizer. One expert suggested to choose personas based on sociolects that are stereotyped and known through mass media. This area of research is open to sociolinguistic investigations concerning the perceptual judgments between sociolects.

Our approach is transferable to other fields of research where considering social and regional aspects can increase the naturalness of human computer interaction. We advocate the development of socially enabled agents with realistic personas.

Acknowledgements

This work has been supported in by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (ftw.) is supported by the Austrian Government and by the City of Vienna within the competence center program COMET.

References

1. 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 2007, pp. 1553–1556 (2007)
2. Nass, C., Lee, K.M.: Does computer-generated speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology 7(3), 171–181 (2001)
3. Marcus, A., Gould, E.W.: Crosscurrents – Cultural dimensions and global web user-interface design. Interactions 7(4), 32–46 (2000)
4. Voice Award (2007), http://www.voiceaward.de/
5. Pucher, M., Neubarth, F., Rank, E., Niklfeld, G., Guan, Q.: Combining non-uniform unit selection with diphone based synthesis. In: Proc. Eurospeech 2003, pp. 1329–1332 (2003)
6. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP 1996, pp. 373–376 (1996)
7. Baum, M., Erbach, G., Kubin, G.: SpeechDat-AT: A telephone speech database for Austrian German. In: Proc. LREC workshop very large telephone databases (XL-DB) (2000)
8. Cohen, M.H., Giangola, J.P., Balogh, J.: Voice user interface design. Addison-Wesley, Reading (2004)
9. Moosmüller, S.: Soziophonologische Variation im gegenwärtigen Wiener Deutsch. Franz Steiner Verlag, Stuttgart (1987)