# Evaluation methods for dialect speech synthesis of similar dialect pairs

Michael Pucher, Carina Lozo, Sylvia Moosmüller

*Acoustics Research Institute, 1040 Vienna, Austria, Email: {michael.pucher, carina.lozo, sylvia.moosmueller}@oeaw.ac.at*

## Abstract

In this paper we investigate, which evaluation metric should be applied in the evaluation of dialect synthetic voices. In our evaluation we show that there are instances of dialect voices with high overall quality and low adequacy ratings, as well as voices with low overall quality and high adequacy ratings. This shows that as least these two metrics should be used in such an evaluation. For the evaluation of the adequacy of a voice we use task specific synthetic prompts and a situative priming of the listener. Synthetic dialect or sociolect voices can extend our ability to design realistic spoken dialog systems by allowing us to incorporate different personas.

## Introduction

Dialect synthesis is a challenging area of research and contrasts the synthesis of standard varieties not only as to the non standard nature of dialects but also in collecting proper corpus data. Previously we evaluated a method for synthesizing new dialects with existing dialect models of a similar dialect by using a simple phone mapping. Then we used a small amount of training data to transfer the original duration and fundamental frequency (F0) of a speaker in order to evaluate how the basic mapping model can be improved. In this contribution we focus on the evaluation methods of these synthesized dialects. To improve dialect synthesis we should not only adapt the existing acoustic models but also the evaluation methods. It is expected that the presentation of synthesized dialect to the listener is crucial to the rating of these systems. Due to the versatile connotations of dialects we assume that a sterile evaluation setting seems inappropriate to the listener and needs to meet the situative demands. In order to gain more insights on how the listener can be sensitized to synthesized dialect we propose an adapted evaluation method based on the data from [2] which considers potential fields of application of synthesized dialect.

We agree with [1] that "there is no such thing as a voice user interface with no personality". We further know from sociolinguistic studies that the perception of a language variety (sociolect, dialect, accent) influences our evaluation of a speaker's attributes concerning his/her competence, intelligence, friendliness, and so on. The concept of "persona" can be defined as the "standardized mental image of a personality or character that users infer from the application's voice and language choices" [1].

In a spoken dialog system the speech synthesis component is surely an important part of the system's persona. Spoken dialog systems are used in many applications already today (call center automation, personal assistants, screen readers, web readers, traffic information systems, car navigation systems) and will be used in future applications like human-robot interaction. Different applications however require different personas. An application for electronic banking for example may exclude voices with a certain 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 on how to evaluate dialect and sociolect synthetic voices for persona design is addressed in this paper. Dialect or sociolect voices furthermore 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.

**Table 1:** *Standard Austrian German (SAG), Viennese dialect (VD), Innervillgraten dialect (IVG), and Bad Goisern dialect (GOI).*

| Variety | Variety type |
|---|---|
| Std. Austrian German (SAG) | Standard |
| Viennese dialect (VD) | Sociolect |
| Innervillgraten dialect (IVG) | South Bavarian dialect |
| Bad Goisern dialect (GOI) | Middle Bavarian dialect |

## Synthetic voices

The used dialect voices were developed as speaker dependent voices using the HSMM-based speech synthesis system published by the EMIME project [8]. Sound samples were recorded at 44100 Hz, 16 bits/sample. The training process was also performed using these specifications. Cutting and selection was performed manually. Noise cancellation and volume normalization was applied to the recordings. Synthesized samples used in the evaluation were also volume normalized. A 5 ms frame shift was used for the extraction of 40-dimensional mel-cepstral features, fundamental frequency and 25-dimensional band-limited aperiodicity measures. More details on the voice training can be found in [6].

The data for training the voices was collected in different projects, which are described in detail in [3]. The data for the Viennese (VD) and Standard Austrian German (SAG) voice was collected within the the research project "Viennese sociolect and dialect synthesis" (VSDS), where we developed three voices for speech synthesis modeling three Viennese varieties. One voice representing "the Viennese dialect" also used for this study, one representing colloquial Viennese, and one representing the youth language in Vienna. In this project we also collected data of a speaker of Standard Austrian German, which was also

used here.

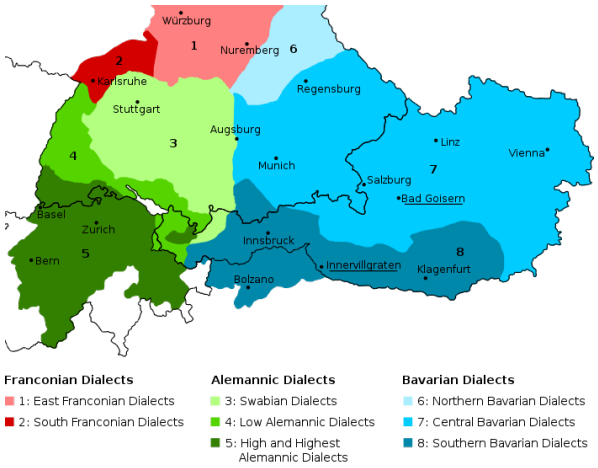


**Figure 1:** Upper German dialects

The voices for Middle Bavarian dialect of Bad Goisern (GOI) and South Bavarian dialect of Innervillgraten (IVG) are part of the Goisern and Innervillgraten Dialect Speech (GIDS) Corpus that is a collection of audiovisual speech recordings for research purposes. It consists of a total of 7068 sentences spoken by eights speakers from two Austrian villages, Bad Goisern (BG) and Innervillgraten (IVG). For each speaker, about two thirds of the recorded sentences are in the speaker's respective dialect and the rest is in Regional Standard Austrian German (RSAG). The dialect of Bad Goisern in the Salzkammergut region belongs to the (South)-Central Bavarian dialects, and the dialect of Innervillgraten in the East Tyrol region belongs to the Southern Bavarian dialect family as shown in Figure 1.

## Evaluation

We assume that a sterile evaluation setting for dialect synthesis seems inappropriate to the listener and needs to meet the situative demands [7]. The evaluation was compiled via an online survey tool [5], where the listener was presented with one audio sample at a time, whereas the rating was obligatory to be performed before the next sample would be presented. We had 26 listeners, three Germans, 23 Austrians from age 17 to 67, 15 female and 10 male listeners (one didn't state their gender) conducting the evaluation. The Austrian listeners were distributed across the federal states of Austria Lower Austria, Upper Austria, Styria and Vienna. Since we had a particular interest in the relation between dialect background of a listener and his or her evaluation of the adequacy of a synthesized dialect, we asked the listeners to rate themselves as dialect or non-dialect speaker. Overall 78% of the Austrian listeners would rate themselves as dialect-speakers. In the evaluation we presented 13 synthesized sentences to the listener. We used four different synthetic voices (VD, GOI, AT and IVG) and seven different contexts (navigation, reservation, public traffic, gaming, weather and public service) in the survey. First we asked the listener to rate the adequacy of the synthesized sentence in the given context on a slider offered by the evaluation interface. The two ends of the slider were named "very inappropriate" (0%) to "very appropriate" (100%). (1) shows an example of a public traffic task for VD.

(1) Stellen sie sich vor, Sie fahren mit der S-Bahn durch Wien und die Ansage der Haltestellen erfolgt mit dieser Stimme (Imagine that you drive through Vienna with the train and the announcement of stops is done with the following voice).

*Bitte bewerten Sie, ob ihnen diese Stimme in der angegebenen Situation als passend erscheint (Please evaluate if you find this voice appropriate in this context).*

That followed a Mean Opinion Score (MOS) test on the quality of the synthetic voice, since we wanted to look into the connection between the adequacy-rating and the quality-rating of a voice. Each sample was rated on an ordinal scale (1 - "sehr gut (very good)", 2 - "gut (good)", 3 - "neutral (neutral)", 4 - "eher schlecht (poor)", 5 - "sehr schlecht (very bad)").

## Results

A statistical analysis of the retrieved data was conducted with [4]. In Figure 2 the rating in % is shown for each audio sample and task.

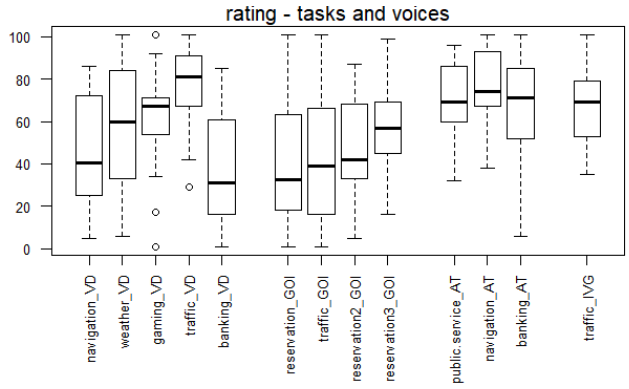


**Figure 2:** rating in % by tasks and voices

For the Mean Opinion Score (MOS) of the voice quality shown in Figure 3 we found significant differences between the single voices VD and GOI, VD and AT, GOI and AT, GOI and IVG, AT and IVG ($p<0.05$). Only in the comparison between the VD and IVG voice we could not find a significant difference.

Overall the AT voice is significantly better rated than the dialect voices. A performed ANOVA of the data showed that there is a statistical correlation between the rating of adequacy and the quality of the synthetic voice. Contradictory to our expectations we found no positive correlation between the dialect background of the listener and the rating of adequacy (Figure 4, left), also there was no significant difference between young (<30 years) and older (>30 years) listeners. However it is worth mention-

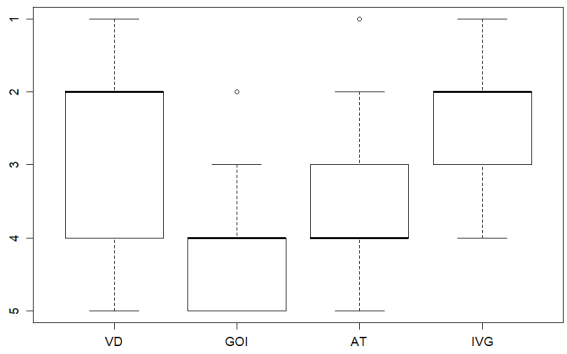


**Figure 3:** Mean Opinion Score of the different voices (1 - "very good" to 5 - "very bad")

ing that the mean rating of both participants with dialect background and participants from the younger age group was slightly higher than the participants without dialect background and participants from the older age group (shown in Figure 4, right).

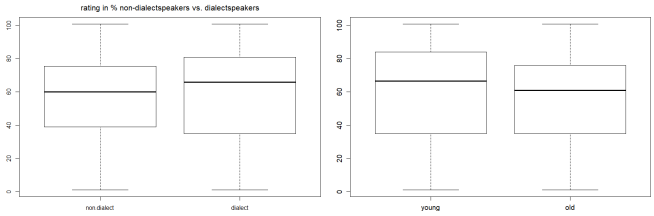


**Figure 4:** rating in % by dialect-background (left), rating in % by age (right)

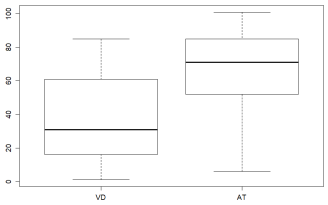


**Figure 5:** rating in % for banking task - AT vs. VD.

Considering the rating per task concerning dialect vs. Standard voices we can see in Figure 5 that the AT sample for the banking task is higher rated than the VD sample. The scores for these ratings show a significant difference ($p<0.05$). Taken together with the MOS ratings (Figure 3), this shows that voices with high MOS ratings can also achieve low adequacy ratings (Figure 2).

The low MOS score of the GOI dialect voice, on the other hand, is contrasted by a high adequacy rating for this voice for certain tasks. For example, we have high adequacy rating (mean=56%) for a GOI reservation task in contrast to its poor voice quality MOS (mean=4.1). This shows that the adequacy and MOS ratings are both needed for the evaluation of dialect synthesis.

## Conclusion

In contrast to common evaluation methods the more application-oriented approach to evaluate dialect synthesis we presented in this contribution shows interesting results. The findings in the evaluation illustrated that there is low adequacy for high quality voices in certain tasks and high adequacy for low quality voices in certain tasks. We have showed, that for further studies on dialect synthesis it is indispensable not only to improve the synthesis systems or voices but also the evaluation methods. When the listener is able to put a synthetic voice into a specific context, it seems the voice is better accepted than in a sterile setting. The versatile connotations of regional dialects can be utilized to achieve a higher acceptance of synthetic voices in everyday life and thereby smooth the way for various fields of applications.

## References

[1] M. M. H. Cohen, J. P. Giangola, and J. Balogh. *Voice User Interface Design*. Addison-Wesley, 2004.

[2] M. Pucher, C. Lozo, and S. Moosmüller. Phone mapping and prosodic transfer in speech synthesis of similar dialect pairs. In *28. Konferenz Elektronische Sprachsignalverarbeitung 2017, Saarbrücken*, pages 180 –185, Germany, 2017.

[3] M. Pucher, M. Rausch-Supola, S. Moosmüller, M. Toman, D. Schabus, and F. Neubarth. Open data for speech synthesis of austrian german language varieties. In *12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, pages 147–150, Munich, 2016.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[5] Soscisurvey. Website, 2018. `https://www.soscisurvey.org/`.

[6] M. Toman, M. Pucher, S. Moosmüller, and D. Schabus. Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis. *Speech Communication*, 72:176–193, September 2015.

[7] P. Wagner and S. Betz. Speech synthesis evaluation - Realizing a social turn. In *Tagungsband Elektronische Sprachsignalverarbeitung (ESSV)*, pages 167–172, 2017.

[8] J. Yamagishi and O. Watts. The CSTR/EMIME HTS system for Blizzard challenge 2010. In *Proceedings of the Blizzard Challenge Workshop*, pages 1–6, Kansai Science City, Japan, Sept. 2010.