Open data for speech synthesis of Austrian German language varieties

Michael Pucher[1], *Michaela Rausch-Supola*[1], *Sylvia Moosmüller*[1],
Markus Toman[2], *Dietmar Schabus*[3], *Friedrich Neubarth*[3]

[1]Acoustics Research Institute (ARI), Austrian Academy of Sciences (OAW)
{michael.pucher,michaela.rausch-supola,sylvia.moosmueller}@oeaw.ac.at
[2]Vienna University of Technology (TUW), Austria
m.toman@neuratec.com
[3]Austrian Research Institute for Artificial Intelligence (OFAI)
{friedrich.neubarth,dietmar.schabus}@ofai.at

Abstract

In this paper we summarize open data sets and open source software that we have released for Austrian German language varieties as a result of several research projects. We also describe some data sets that are released for research purposes only, due to licensing limitations. From the development of these resources we draw conclusions concerning the collection and licensing of such data with a special focus on the problem of speech synthesis where the voice identity of the speaker plays an important role. Furthermore we discuss recordings that we plan to perform in the future, where we aim to cover most Austrian dialects.

Index Terms: speech synthesis, language varieties, dialect, sociolect, open data, open source

1. Introduction

In this paper we describe a set of tools and data sets that we have already released for Austrian German language varieties (dialects, sociolects) concerning audio as well as audio-visual speech synthesis. This includes the following:

- The **open source** SALB front-end for speech synthesis using Hidden Markov Model (HMM)-based Speech Synthesis System (HTS) voice models (available from `http://m-toman.github.io/SALB/`) [1].
- "Leo", an **open data** HTS based voice model for Austrian German (available from `https://sourceforge.net/projects/at-festival/`) [2].
- **Open data** for building unit selection voices for Viennese dialects with the Festival speech synthesis system (available from `http://speech.kfs.oeaw.ac.at/vdvoices/`) [3].
- **Open research data** for audio-visual dialect synthesis - Goisern and Innervillgraten Audiovisual Dialect Speech Corpus – GIDS (available from `http://speech.kfs.oeaw.ac.at/gids/`) [4].
- **Open research data** for triple modality speech synthesis – Multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech MMASCS (available from `http://speech.kfs.oeaw.ac.at/mmascs`) [5].

Since all synthetic voices in the current speech synthesis paradigms (unit selection, HMM, Deep Neural Network (DNN)) are built using data of a specific speaker, they will reproduce the speaker's identity to a certain extent. This brings about some specific licensing problems that we will discuss in Section 8.

In the future we also aim to record dialect data of 40 different dialect regions in Austria in the field using a mobile recording studio. These recordings shall also be released under an open data or open research data license.

2. SALB front-end for speech synthesis using HTS voice models

Hidden-Markov Model (HMM) based speech synthesis provides a methodology for flexible speech synthesis while keeping a low memory footprint [5]. It also enables speaker adaptation from average voice models, allowing the creation of new voice models from sparse voice data [6], as well as techniques like interpolation [7][8] and transformation [9][10] of voices. A well-known toolkit for creating HMM-based voice models is HTS [11, 12]. Separate software toolkits are available to actually synthesize speech waveforms from HTS models. A popular, freely available framework is hts_engine [13]. Speech synthesis front ends on the other hand provide means for analyzing and processing text, producing the necessary input for the synthesizer. In HTS this input is a set of labels where usually each label represents a single phone and contextual information, including surrounding phones, position in utterance, prosodic information etc. While not exclusively being front ends and not specifically targeted for HTS, popular choices are Festival [14] or Flite [15]. Festival is a complex software framework for building speech synthesis systems focusing Unix-based operating systems.

Our main goal when creating the SALB front-end framework was to easily allow HTS voices to be used with the Speech Application Programming Interface 5 (SAPI5). This allows the framework to be installed on different versions of the Microsoft Windows operation system as speech synthesis engine, making HTS voice models available as system voices to applications like screen readers, e-book creators etc. The second goal was

[1] Published in [1]

[2] Published in [1]

[3] Published in [2]

[4] Published in [3]

[5] Published in [4]

simple integration of new languages and phone sets. The third goal was portability to mobile devices.

Flite has been adapted for HTS in the Flite+hts_engine software [13] and due to its small and portable nature it seemed like a good fit to our requirements. The structure of Flite makes integrating new languages rather cumbersome. [6] Therefore our framework integrates Flite for text analysis of English while additionally providing a second text analysis module that can utilize Festival style pronunciation dictionaries and letter to sound trees. Text preprocessing tasks (e.g. number and date conversion) can be added to the module in C++. Adding a completely new text processing module is also possible. The framework includes hts_engine for speech waveform synthesis and can be extended by other synthesizers.

3. “Leo”, a HTS based voice model for Austrian German

Category	Phones (IPA)
Vowels (monoph.)	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː øː æː ɐ œ œː ə u ʊ uː ʏ y yː
Vowels (monoph.) nasalized	ɒ̃ː ɔ̃ː æ̃ː œ̃ː
Diphthongs	aɪ̯ ɒːɐ̯ ɑːɐ̯ ɒɪ̯ ɑʊ̯ ɛɐ̯ ɛːɐ̯ ɪɐ̯ iɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ oːɐ̯ ɔʏ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ ʊːɐ̯ ʏɐ̯ yːɐ̯
Plosives (stops)	b̥ d̥ g̊ k ʔ p t
Nasal stops	m n ŋ
Fricatives	ç x f h s ʃ v z ʒ
Approximants	j
Trill	r
Lateral approx.	l

Table 1: Phone set used for Austrian German voice “Leo”.

With the framework we provide a free voice model of a male, professional speaker for Standard Austrian German called “Leo”. The model is built from 3,700 utterances recorded in studio quality using a phonetically balanced corpus. The phone set used in the voice can be seen in Table 1. A pronunciation dictionary with 14,000 entries, letter to sound rules and procedures for number conversion are also included. The voice is distributed with the framework, but can also be used with the Festival speech synthesis system.

4. Unit selection voices for Viennese dialects

Within the research project “Viennese sociolect and dialect synthesis” (VSDS), we developed three voices for speech synthesis modeling three Viennese varieties. In the light of personalization and regionalization of speech based interfaces it becomes indispensable to develop not only high quality speech synthesis for different languages but also for a representative set of language varieties, i.e., dialects that differ from the standard variety substantially enough to treat them alongside different languages. In performing this task, the focus lies on the necessity that the developed synthetic voices must be able to shift between the standard variety and specific dialects, similar to everyday language use [7]. In Vienna, language varieties are differentiated rather socially than regionally, therefore it would be correct to speak about sociolects. In the VSDS project, we developed three different voices: a voice representing “the Viennese dialect”, one representing colloquial Viennese, and one representing the youth language in Vienna. For the recordings, we could win two renowned actors and for the recordings of youth language, we arranged a casting among pupils of vocational schools.

4.1. Speaker selection

The selection of the professional speakers was based on several criteria, amongst others: reading speed and accuracy, the accuracy of their standard Austrian pronunciation, the degree of authenticity of their sociolect, the consistency of their pronunciation (in particular, we did not want them to shift between different sociolects without being told so), and the pleasantness of the voice.

4.2. Text selection

The quality of a unit selection voice highly depends on how well the recorded material covers the set of possible diphones and prosodic contexts. Most of our recording text script for the standard Austrian variety was selected from large corpora of non-proprietary texts, such as EU parliament debate transcripts, and from the Viennese city magazine “Falter” (with their friendly approval). We were aiming at diphone coverage with the following linguistic context features: lexical stress, syllable boundaries and word boundaries. During the initial iterations of text selection, we focused on the most frequent diphones without features while taking account of some back off strategies, for example that diphones bridging a word boundary can easily be backed off by inserting a short pause.

4.3. Recording

The recordings were made in an unechoic, acoustically isolated room with a HD-recorder (44100 kHz sampling rate, 16 bit encoding) and a professional microphone. We made sure that the recording parameters (distance to microphone recording level) were the same for each session. The recordings were semi-automatically segmented at sentence level using the acoustic software S_TOOLS-STx of Acoustics Research Institute (ARI) and a script written in Perl. The speech database contains transcriptions and soundfiles corresponding to single sentences. Importantly, these are not just cut from the original recordings, but they can be dynamically exported each time some alignments change.

4.4. Voices

The release “Unit selection voices for Viennese dialects” contains data for 3 Viennese voices (Table 2). Additionally the release contains base lexica for the phonetic encoding of each variety, which covers the most important and typical words of the respective Viennese variety, and a set of letter-to-sound rules for Austrian German. The voices can be used with the Festival speech synthesis system [14], in particular the open-domain unit selection Multisyn [16]. The provided data can also be used for training of HMM-based voices for HTS [12].

[6] We have published instructions on adding a new language to Flite: http://sourceforge.net/p/at-flite/wiki/AddingNewLanguage/

Voice ID	Variety	Age group	Database size
HPO	Viennese dialect	45-60	2:55
HGA	Colloquial Viennese	60-70	3:10
JOE	Viennese youth language	15-25	2:11

Table 2: Viennese dialect unit selection voices.

5. GIDS – Goisern and Innervillgraten Audiovisual Dialect Speech Corpus

Visual speech synthesis techniques have possible applications in computer games and films. Generating visual speech directly from audio data is nowadays a state-of-the-art technique in facial animation in the computer games industry [17]. We have developed a corpus of audio-visual speech recordings to investigate visual dialect text-to-speech synthesis where we generate an acoustic and visual signal of a certain speaker from given text.

5.1. Corpus

The Goisern and Innervillgraten Dialect Speech (GIDS) Corpus 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.

After a careful phonetic analysis we compiled sets of phonetically balanced sentences (656 for IVG and 665 for GOI) with respect to the phone set established for the dialect, the frequency of occurrence of each phone in the data, and the context specific variation of phones. The utterances of the recording script were extracted from a larger corpus of material consisting of 18-20 hours of recordings for each dialect with at least 10 speakers per dialect. These sentences consisted of spontaneous speech (elicited with key words) and translation tasks. We created a lexicon of words occurring in the script. The script was divided into a training and testing part. In the final audio-visual recordings we recorded 2 male and 2 female speakers per dialect, i.e., 8 speakers in total.

The recordings consist of optical 3D facial motion tracking data, captured with a NaturalPoint OptiTrack Expression system,[7] the greyscale video data also recorded by the same system, and studio quality audio.

For each of the recorded utterances, the corpus contains a RIFF wave audio file, facial marker data in the form of a matrix stored as a text file, a gray scale video from the optical system, the sentence of the utterance in plain text, a text file listing the phones spoken in the utterance including begin and end times of all phones, and a quin-phone full-context label file.

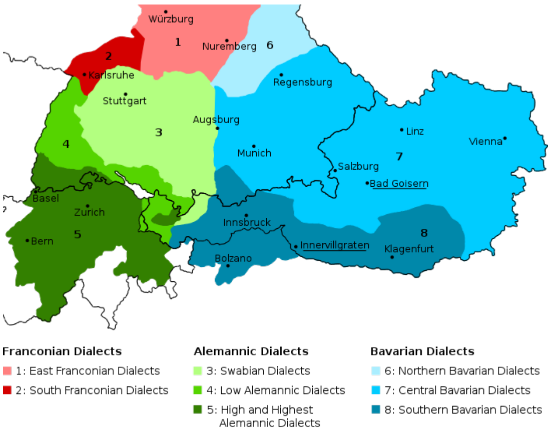
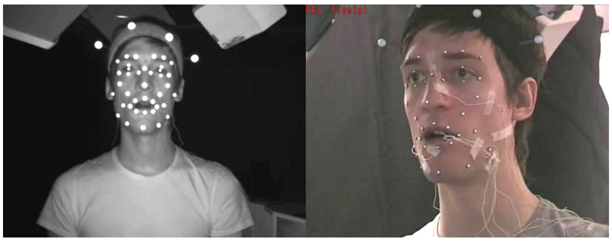


Figure 1: Upper German dialects



(a) (b)

Figure 2: Example still images from the gray scale video from the OptiTrack system (a) and the color video from the camcorder (b) showing visual markers and articulatory markers.

6. MMASCS – Multi-Modal Annotated Synchronous Corpus of Speech

The MMASCS corpus is a multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech. The tongue motion data is captured with Electro-Magnetic Articulography (EMA).

The MMASCS corpus combines facial motion capture data with intra-oral EMA data. In comparison to optical motion capturing only, this has the obvious advantage of also providing tongue motion data, which is impossible to capture optically. In comparison to EMA data only, it has the advantage of providing a larger number of tracked points on the lips, eyelids, eyebrows and other areas of the face. While it is in principle possible to use EMA coils also on the face surface, the inexpensive and easy-to-attach optical markers are much less intrusive for the speaker than the EMA coils with their cable connection (one cable per coil) to the articulograph. Another difference is that our data is for Austrian German speech. One can imagine that it might be interesting to investigate inter-lingual differences in speech motion, once a larger number of corpora (of EMA and/or facial motion data) in various languages is available (of course speaker-specific effects would need to be accounted for). Finally, our data is different in that it comprises data of speech at three different speaking rates (normal, fast and slow).

We have already used this corpus for evaluating a method to convert from non-acoustic to acoustic speech, where we could show that visio-articulatory features can improve the conversion compared to visual only features [19].

7. Recording of high quality dialect data in the field

In the future we also aim to record dialect data of 40 different dialect regions in Austria in the field using a mobile recording

[7] http://www.naturalpoint.com/optitrack/

studio. These recordings shall also be released under an open data or open research data license.

The selection of dialect locations and speakers as well as the phonetic and phonological analyses of 40 Austrian dialects, an essential prerequisite for dialect synthesis, is currently in progress within the SFB project "Deutsch in Österreich" ("German in Austria").

For speech synthesis we will create phonetically balanced recording scripts, record 1 male and 1 female speaker for each location, perform a semi-automatic transcription of the recordings, build and investigate acoustic models for statistical parametric synthesis, and build a synthesis front-end. To achieve high quality recordings, we will deploy a mobile recording studio. We will test the studio by recording two speakers in the course of the year 2016. The recording scripts will be adapted from our existing recording scripts for Standard Austrian German (SAG), Viennese (VD), Innervillgraten (IVG) and Bad Goisern (BG) dialect.

8. Licensing, repositories, and standards

Since we are also synthesizing a speaker's identity the data we are collecting is very personal and the speakers must be informed about possible applications of their data. Many of the speakers that we have recorded agreed to release their data within an open data or open source framework, but we can also observe that the use of speech synthesis technology is not yet as widespread that speakers are able to fully understand possible application scenarios. Independent of country dependent legal requirements as scientists and developers we have to make sure to give speakers that we are recording a realistic perspective on what can happen with their voice data. The Festvox documentation [20] contains some guidelines on these issues with a list of possible licenses from "free for any use" to "fully proprietary", but in the future we may need more sophisticated licenses that reflect the fast technological changes that we are witnessing.

Our data sets are available from our websites, but it would be beneficial to have a common repository for data distribution within the speech communication community.

The data format standards that we use for the creation of our data sets are mainly set by the popular speech synthesis frameworks such as HTS and Festival. Such a kind of implicit and bottom-up standardization seems natural for a field that is strongly driven by research, but might not be optimal from an industry point of view.

9. Conclusion

We have given an overview of open data sets and open source software that we have released for Austrian German language varieties and drew some conclusions concerning the collection and licensing of such data with a special focus on speech synthesis. Furthermore we discussed recordings that we plan to perform in the future, where we aim to cover most Austrian dialects.

10. Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" was funded by the Vienna Science and Technology Fund (WWTF). This research was also funded by the Austrian Science Fund (FWF): P22890-N23, P23821-N23, and by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB).

11. References

[1] M. Toman and M. Pucher, "An open source speech synthesis frontend for HTS," in *Text, Speech, and Dialogue (TSD) 2015*, Pilsen, Czech Republic, 2015, pp. 291–298.

[2] M. Pucher, F. Neubarth, V. Strom, S. Moosmller, G. Hofer, C. Kranzler, G. Schuchmann, and D. Schabus, "Resources for speech synthesis of viennese varieties," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010, pp. 105–108.

[3] D. Schabus and M. Pucher, "Comparison of dialect models and phone mappings in hsmm-based visual dialect speech synthesis," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, Vienna, Austria, Sept 2015, pp. 84–87.

[4] D. Schabus, M. Pucher, and P. Hoole, "The MMASCS multimodal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May 2014, pp. 3411–3416.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[7] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, feb 2010.

[8] C. Valentini-Botinhao, M. Toman, M. Pucher, D. Schabus, and J. Yamagishi, "Intelligibility Analysis of Fast Synthesized Speech," in *Proc. Interspeech*, Singapore, September 2014, pp. 2922–2926.

[9] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, pp. 528–531.

[10] M. Toman, M. Pucher, and D. Schabus, "Cross-variety speaker transformation in HSMM-based speech synthesis," in *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, Barcelona, Spain, Aug. 2013, pp. 77–81.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW)*, Bonn, Germany, Aug. 2007, pp. 294–299.

[12] "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.

[13] "hts-engine," http://hts-engine.sourceforge.net/.

[14] "Festival," http://www.cstr.ed.ac.uk/projects/festival/.

[15] "Flite," http://www.festvox.org/flite/.

[16] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[17] SpeechGraphics, "Speech Graphics - Audio-driven facial animation," http://www.speech-graphics.com/, 2015.

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

[19] M. Pucher and D. Schabus, "Visio-articulatory to acoustic conversion of speech," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, Vienna, Austria, Sept 2015, p. Article No.6.

[20] "Festvox - Who owns a voice," http://festvox.org/bsv/x794.html.