Text-to-Speech Engine with Austrian German Corpus

C. Kranzler[1,2]*, F. Pernkopf*[1]*, R. Muhr*[1]*, M. Pucher*[2]*, F. Neubarth*[3]

[1]Signal Processing and Speech Communication Lab (SPSC)
Graz University of Technology, Austria
[2]Telecommunications Research Center Vienna (ftw)
Vienna, Austria
[3]Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
{kranzler, pucher}@ftw.at, pernkopf@tugraz.at, rudolf.muhr@uni-graz.at,
friedrich.neubarth@ofai.at

Abstract

This paper deals with developing a unit selection speech corpus for the Austrian variety of German by (re)using the resources for German and adapt them to Austrian German. This means adaptation on different levels such as lexicon level, phone level, or speech data level, whereas a compromise between reusing the given resource and an exact time-consuming phonetic transcription has to be found. In experiments our Austrian German voice was evaluated by speech experts against the correctness of pronunciation while using an adapted Austrian lexicon.

1. Introduction

Although much research on the topic of speech synthesis has been done so far, the areas of regionalization, prosody, and emotional speech still pose challenges for current speech technology.

Speech synthesis can be divided into three different methods [8]: Articulatory synthesis, formant synthesis, and concatenative synthesis, whereas articulatory synthesis and formant synthesis are model-based methods and concatenative synthesis is a data-based method.

Articulatory synthesis models the physics of the human articulators such as the vocal tract geometry or the shape of the lips. For the description of the articulators, mathematical and electrical models are used.

Text-to-Speech (TTS) systems on the basis of formant synthesis use the source-filter model for speech production.

Concatenative speech synthesis is currently the most common method. It uses prerecorded speech segments which are concatenated. This method produces a much more natural sounding speech in comparison to model-based methods, i.e. articulatory or formant synthesis. One widely-used technique in the past was diphone synthesis, where the speech segments model the transitions between phones, i.e. *diphones*, where the borders are located in the middle of a phone. For diphone synthesis every existing diphone (or possible combination of two phones) of a language has to be recorded under controlled conditions, e.g. a consistent fundamental frequency ($f0$) of the speaker and negligible background noise.

At the synthesis stage, prosody manipulation techniques [1], [2] such as the commonly used Pitch Synchronous Overlap and Add (PSOLA) and Multi-Band Re-synthesis Overlap and Add (MBROLA) are necessary.

Non-uniform *unit selection synthesis* extends diphone synthesis beyond diphones to units of different length or type including multiple representations [6]. A huge amount of prerecorded speech units is available, which is concatenated during synthesis. Our Austrian German voice is based on this technique and further details are given in Section 3.

In this paper, we focus on regionalization. Although there are differences between the German language spoken in Germany (GG) and the varieties spoken in Austria or Switzerland, the variety in Germany is the accepted standard and most language corpora for TTS synthesis created so far are based on it. Here, we develop a speech synthesizer built upon a speech corpus of Austrian German (AG). Additionally, we present the main fundamental differences between GG and AG.

The paper is organized as follows: In Section 2 we give an overview of the differences between GG and AG before we introduce unit selection synthesis in Section 3. In Section 4 we present the results of two experiments. We conclude the paper in Section 5 and discuss future work.

2. Differences between AG and GG

Our Austrian German corpus for a unit selection TTS system is based on modifications of the GG Bonn Machine-readable Pronunciation lexicon (BOMP) [17]. These modifications represent the most important aspects of AG, and the TTS system produces a correct pronunciation in the sense of a regionalized variety of German. All these differences are based on linguistic research in the area of Austrian German [3], [4].

2.1. No voiced sibilants

In AG, what is pronounced as voiced sibilants /z/ or /Z/ in most German standard varieties, is phonetically realized as unvoiced unequivocally (/s/ and /S/). It has to be noted, however, that some speakers do not strictly adhere to this regularity, in that they produce a slightly voiced variant of the sibilants, but these speakers belong to specialized groups, in our case radio speakers. Still, the differences in voicing are represented in the speech data and the voiced version in the lexicon is pronounced unvoiced when generating the synthetic speech. For that reason we did not consider any modification to the lexicon, and the original GG forms can be used for several types of Austrian speakers.

2.2. /r/-vocalization

The next phenomenon when adapting the German BOMP lexicon to AG is the /r/-vocalization where the phone /r/ post-vocalically fully or partially vocalizes to a-schwa /6/. This is illustrated by the word "werben" ('solicit'). The transcription in BOMP is /vEr.b@n/, but in the pronounciation of the Austrian variety (and in most German varieties as well) the /r/ is replaced by (or augmented with) /6/, i.e. /vE6.b@n/. This calls for automatic rule-based re-coding in the lexicon to avoid confusion in alignment and selection between /r/ and /6/ as separate phones.

2.3. Elision of the *e-schwa*

In AG the *e-schwa* tends to minimize (or elide) before a nasal in non-onset position. For example the phonetic transcription of the word "fragen" ('to ask') in the GG BOMP lexicon is /fra:.g@n/. Depending on the speaker of AG, when the schwa reduces towards zero, this word should be rather transcribed as /fra:.gN=/ (with place-assimilation of the nasal towards the preceding obstruent). Luckily, this phenomenon needs no recoding since at the time of alignment the speech recognizer sets the length of the schwa to zero if it does not appear. At the synthesis stage if a *schwa* with zero duration should have some duration, the target and join costs as discussed in section 3.3 may circumvent such a problem. These cost measures avoid the selection of a *schwa* with zero duration on the basis of the phonotactic context, which also captures the place assimilation of the nasal.

2.4. GG /C/ versus AG /k/

Two phenomena can be distinguished here: (i) Pronunciation of words ending in 'ig': The GG standard pronunciation for the orthographical 'ig' at the end of a word is /IC/, whereas the correct AG form is pronounced as /Ik/, e.g. the word "lustig" ('funny') is transcribed as /lus.tIC/ in GG but the AG transcription is /lus.tIk/. (However, words written '-ich' are pronounced /IC/ in both varieties.) (ii) Foreign words: There is a difference between the GG pronunciation /C/ versus the AG pronunciation /k/ at the beginning of a word which is orthographically written as "Ch". For example the word "Chemie" ('chemistry') is transcribed as /Ce:.mi:/ in GG and /ke:.mi:/ in AG. Both phenomena are captured by rules applied to the lexicon to get the right transcription for the AG lexicon.

2.5. Special Austrian words

This task is problematic because speech is dynamic and words may appear and disappear over the years and across varieties. Nevertheless, there are some typical Austrian words which can be claimed to be an integral member of most varieties. These words are transcribed and included in the lexicon for TTS synthesis. Such words are for example "Obers" ('cream') which has the GG pendant "Sahne". A total of 52 typical Austrian words have been selected and added to the lexicon, but we are aware that this is just a preliminary endeavor since the number of these words is much larger and difficult to determine.

2.6. Letter-to-Sound (LTS) rules

If a lexicon with a high coverage is available, LTS rules can be generated on the basis of such a lexicon as training data. Beforehand it is necessary to define a set of potential mapping rules, i.e. onto which phone symbols an orthographic character can be mapped. So for instance the letter 'k' always maps to a /k/, but the letter 'h' for instance can map to a /h/ or to a null-element, which means that it is part of an orthographic combination and not mapped to a separate phone at all. An 'x' in German always maps to two phones, namely /k/ plus /s/, exceptions can only be found in loan words from Spanish or French. Very often the letter-to-phone mapping is ambiguous, for instance with the letter 'i' which can map to /I/ or /i:/.

Once these rules are established, the transition probabilities for the transformation from letters to phone symbols can be calculated based on the lexikon. Using a Classification and Regression Tree (CART) [9] one can predict the transition from letters to phones. In such a tree the root contains the letter and the leaves contain the phone mappings. Each branch contains a the probability of a certain phone mapping.

3. Unit selection synthesis

Unit selection is currently the most widely used synthesis technique due to a high quality of the resulting speech output. The quality depends on the amount of previously collected data (usually very large) or whether or not the data fit well the text domain the synthesized voice is going to be used in. For unit selection the pre-recorded units are selected and joined in an optimal manner (see section 3.3). Uniform unit selection systems restrict the units to be of a fixed type, e.g. diphones. The most common systems define the borders of these units in the middle of phones. Non-uniform unit selection systems facilitate the use of various types. The text is selected from the previously recorded material in a way that reduces the number of concatenations to a minimum. By doing so the voice sounds increasingly more natural.

3.1. Selection of the recording material

For producing high quality speech it is necessary to have a large amount of candidates available in the database to select the most appropriate units. All actually occurring diphones of the language should appear at least once in the speech database. This ensures that every possible combination of two phones can be synthesized. However, it is desirable to have more instances of the same diphone in various prosodic contexts, because this is a prerequisite to synthesize the same wording with different prosody. E.g., to synthesize a sentence with a prosody that indicates an affirmative or interrogative meaning, one needs different instantiations of the same phone string at the end of the sentence. Therefore, it is indispensable to select a representative mixture of sentences (e.g. questions, answers, enumerations) for the text serving as the base for the recordings. If there is a representative text corpus for a given language available, this text corpus can by analyzed and a subset of it selected for recording material to give a good coverage of the words for this specific language (variety) [11], [12].

3.2. Features

To guarantee a smooth transition at the unit borders objective features are necessary to characterize the units [7]. The most important feature is the fundamental frequency (f_0) which definitely should not differ too much between two joined units. Otherwise, the resulting speech sounds very artificial. Other important features are the duration of a phone, the Mel Frequency Cepstral Coefficients, and whether the element is phrase final or not. The units have target features, which are used to calculate the target costs, and join features associated with the join costs [6].

3.3. Concatenation of the units

During synthesis the optimal sequence of units for a given phone string has to be selected. In a first step the longest possible phone string contained in the inventory of the speech database is determined. Subsequently, the remaining units complementing this sequence best are selected. Therefore, the costs for selecting a unit and the costs for joining these units have to be calculated separatiely. Standardly this is done using the Hunt and Black algorithm [10]. We can assume that we have a set of desired items $S =< s_1, s_2, ..., s_T >$ and a set of units in the speech database $U = \{u_1, u_2, ..., u_M\}$, each of them containing a list of features. The task now is to find the best sequence U that fulfills the specification S. In a first step, we calculate the target cost $T(u_t, s_t)$ for all relevant units, which is the distance between a specification element s_t and a unit in the speech database u_t, and for all unit combinations the join cost $J(u_t, u_{t+1})$, which measures the transition between two adjacent units [10].

3.3.1. Search of the minimum cost path

Searching for the path with the minimum cost is a computationally intensive approach. Consider a diphone unit selection synthesis system where the number of the specified diphones of a sequence that should be synthesized is N. The number of speech units in the database to which the N diphones have to be compared to is M. Then, if we want to determine the costs of every possible path, the number of calculations is N^M.

Since we are looking for the path with the minimum cost, we can apply the Viterbi algorithm [13]. The Viterbi algorithm takes into account only the path with the minimum (or normally with the maximum) cost or rate. Just if there is a path from n to $n + 1$ with the same cost, then this path is also kept in memory, and in the next iteration it is determined, which one of the previous paths point to the path with the minimum cost. The computational complexity is $\mathcal{O}\left(M^2N\right)$. Several methods of pruning can be used to speed up the search [6].

3.3.2. Enhancements

Further improvement of the concatenation can be achieved by manipulating the acoustic features. Especially, the fundamental frequency of a unit can be modified by PSOLA or MBROLA [1] to make the transitions between units smoother. Normally, the units are concatenated as they are since there exist a large number of representations in the speech database.

Another possibility of tuning the unit selection synthesis is adding, e.g. stress, or removing features [6].

4. Experiments

To show the importance of using an adapted lexicon for a specific variety of a language we made two experiments with the original GG and the adapted AG lexicon. The method to compare these two is based on the individual generation of voices using the Festival unit selection system [14], [16].

4.1. Experimental setup

In these experiments we created two separate unit selection voices with the two different lexicons, i.e. one represents the relevant aspects of the Austrian German variety and the other one is a German German lexicon (i.e. BOMP) [17]. Hence, for producing the voice either the GG lexicon or the AG lexicon was used, both at the stage of segmentation/alignment and at the synthesis stage.

As input data we used five phonetically balanced corpora, selected from a well-established corpus collection, the Kiel PHONDAT corpus [5], which is phonetically balanced and covers all phones of Austrian German in different prosodic contexts. The amount of recorded data is very limited, however, it proved sufficient for carrying out the experiments to be discussed in detail in the remainder of this paper.

For testing the different transcriptions, 30 single word utterances have been selected randomly out of a total set of 11600 utterances covering the ambiguity between /C/ and /k/, which is not represented in the GG lexicon (details are given in Section 2). Out of these 30 utterances, 23 represent the orthographical suffix $-ig$ being pronounced /Ik/ in the AG lexicon and /IC/ in the GG lexicon and 7 utterances represent the orthographic string ‘ch’ at the beginning of a word which is pronounced /k/ in AG instead of /C/ in GG.

We presented these utterances to a group of speech experts being born and raised in Austria asking the question if a played utterance generated by the unit selection synthesis represents the correct variety (AG).

4.2. Results

The following two tables illustrate the results of the experiments. ’0’ means that the majority of test persons decided that the utterance does not represent the aspects of Austrian German, ’1’ denotes that it does.

In the first experiment the shift from /IC/ to /Ik/ was tested. Table 1 shows the utterances and their assessed correctness with respect to the Austrian variety of German. The two examples which were found to not sound Austrian neither with the GG nor with the AG lexicon exemplify the fact that the speaker did not always pronounce the words as they are transcribed. By having a closer look into the speech data, several of this deviations could be observed.

utterance	**gloss**	**GG**	**AG**
Ratlosigkeit	helplessness	0	0
Verantwortungslosigkeit	irresponsibility	0	0
Abbaugerechtigkeit	reduction-justice	0	1
angriffslustigstem	most agressive	0	1
Bergpredigten	sermons o. t. Mt.	0	1
Flüssigkeit	fluid	0	1
Geselligkeit	sociability	0	1
Heiligabend	Christmas eve	0	1
Heiligtum	sanctuary	0	1
Honigwein	mead	0	1
lässigst	most casual	0	1
Reisigbesen	besom	0	1
Reisigbündel	fagot	0	1
Rückgängigmachung	cancellation	0	1
Schallgeschwindigkeit	sonic speed	0	1
schwachatmigst	weakest	0	1
Spitzengeschwindigkeit	top speed	0	1
Zigfache	multiple	0	1
Achtzig	eighty	1	1
Ewigkeiten	eternities	1	1
Honigmond	honeyed moon	1	1
mehrsprachig	multilingual	1	1
richtiggehend	fully fledged	1	1

Table 1: Sample words for /IC/ → /Ik/.

The second experiment is concerned with the shift from /C/ to /k/ in the onset of a word. Table 2 shows the results. Here the AG lexicon always produced a pronunciation representing the aspects of Austrian German.

utterance	**gloss**	**GG**	**AG**
Alchemie	alchemy	0	1
Elektrochemie	electrochemistry	0	1
Chemikalien	chemicals	1	1
China	China	1	1
Chinin	quinine	1	1
chirurgische	surgically	1	1
cherubinisch	cherubinic	1	1

Table 2: Sample words for /#C/ → /#k/.

Table 3 shows the overall result of the first and the second experiment. Although the presented results are not significant in a strict statistical sense (due to the small number of utterances used in the experiments), they still show very clearly that the modifications of the lexicon significantly support the different pronounciation between AG and GG. We observed that using the wrong GG lexicon produced only 1/3 correct utterances, whereas the AG lexicon leads to a high proportion of correctly pronounced utterances in total. The mere fact that even in a small randomly selected sample of test cases the standard GG lexicon yields a relatively small number of correct output whereas the adapted AG lexicon is merely always correct (except for certain irregularities in the speech data) supports the decision to create a specially adapted lexicon for the AG variety.

Process	**Lexicon GG**	**Lexicon AG**	**total**
/IC/ → /Ik/	5 / 18	21 / 2	23
/#C/ → /#k/	5 / 2	7 / 0	7
Sum	**10 / 20**	**28 / 2**	**30**

Table 3: Correctness of utterances in #correct / #incorrect.

Another observation apart from the experiments was that using the AG lexicon produced a far more accurate alignment with HTK [18]. This must be due to the fact that the speech signal is assigned to the wrong phone label with the GG lexicon and therefore the aligner is trained with erroneous labels. Inevitably this leads to a voice of lower quality. Hence, using the correct lexicon not only induces the correct pronunciation of a language variety but also improves the quality of the alignment and so the quality of the synthetic voice [15].

5. Conclusion

In this paper, we introduced the development of the first non-uniform unit selection TTS system for Austrian German. Unit selection speech synthesis is perfect for modeling also a limited domain speech corpus, where a limited number of words and sentences can be synthesized. For the Austrian variety of German several adaptations on different levels are necessary. Although the number of recorded phonetically balanced sentences was rather small, the resulting speech database produces a comparably good output. It was used to test the positive effects of using an adapted lexicon for a specific language variety. Although the experiments were not carried out in a statistical manner, they clearly show the advantages of a lexicon that represents certain crucial differences between language varieties.

6. References

[1] Thierry Dutoit, *An Introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, 1997. ISBN 1-4020-0369-2.

[2] Peter Vary, Ulrich Heute, Wolfgang Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Stuttgart, 1998. ISBN 3-519-06165-1.

[3] Rudolf Muhr, Richard Schrodt, *Österreichisches Deutsch und andere nationale Varietäten plurizentrischer Sprachen in Europa*, Verlag öbv&hpt, 1997.

[4] Rudolf Muhr, *Österreichisches Aussprachewörterbuch, Österreichische Aussprachedatenbank*, Verlag Peter Lang, 2007. ISBN 978-3631554142.

[5] Klaus J. Kohler, *Lexica of the Kiel PHONDAT Corpus*, The Kiel Corpus of Read Speech, Vol. I, Arbeitsberichte der Universität Kiel, Nr. 27, 1994.

[6] Paul Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009. ISBN 978-0521899277.

[7] Sadaoki Furui, *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, 2001. ISBN 0-8247-0452-5.

[8] Erhard Rank, *Oscillator-plus-Noise Modeling of Speech Signals*, Dissertation at Vienna University of Technology, November 2005.

[9] Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004. ISBN 0-262-01211-1.

[10] Andrew J. Hunt, Allan W. Black, *Unit selection in a concatenative speech synthesis system using a large speech database*, Proceedings of the International Conference on Speech and Language Processing, Pages 373-376, 1996.

[11] Allen W. Black, Kevin A. Lenzo, *Optimal Data Selection for Unit Selection Synthesis*, 4rd ESCA Workshop on Speech Synthesis, 2001.

[12] Jan P. H. van Santen, Adam L. Buchsbaum, *Methods for Optimal Text Selection*, Proceedings of Eurospeech, 1997.

[13] Bernard Sklar, *Digital Communications, Fundamentals and Applications*, Prentice Hall, 2001. ISBN 0-13-084788-7.

[14] Robert A. J. Clark, Korin Richmond, Simon King, *Multisyn: Open-domain unit selection for the festival speech synthesis system*, Speech Communication Vol. 49, Issue 4, Pages 317-330, 2007.

[15] Friedrich Neubarth, Michael Pucher, Christian Kranzler, *Modeling Austrian dialect varieties for TTS*, Proceedings of Interspeech, Brisbane, Pages 1877-1880, 2008.

[16] Allan W. Black, Kevin A. Lenzo, *Building Synthetic Voices*, http://festvox.org/bsv/index.html, 2007.

[17] The Bonn Machine-Readable Pronunciation Dictionary (BOMP), *http://www.ifk.uni-bonn.de/forschung/abteilung-sprache-und-kommunikation/phonetik/sprachsynthese/bomp*, January 18th, 2007.

[18] Steve Young et. al., *The HTK Book (for HTK Version 3.4)*, http://htk.eng.cam.ac.uk/docs/docs.shtml, 2006.