Modeling Austrian dialect varieties for TTS

Friedrich Neubarth [2], *Michael Pucher* [1], *Christian Kranzler* [1,3]

[1] ftw. (Telecommunications Research Center Vienna), Vienna, Austria
[2] OFAI (Austrian Research Inst. f. Artificial Intelligence), Vienna, Austria
[3] SPSC (Signal Processing and Speech Communication Lab), TU Graz, Graz, Austria
friedrich.neubarth@ofai.at, pucher@ftw.at, kranzler@ftw.at

Abstract

In this paper we discuss certain strategies for building adapted TTS systems for dialectal or regional varieties from a given standard source. The basic question is how much re-coding is necessary for a given transfer and to what extent it is possible to rely on the speech data alone. It will turn out that there are ambiguities that cannot be resolved without a certain amount of linguistic engineering. For exemplification we present two experiments dealing with Austrian standard German and Viennese dialect on the level of phonetic lexicon and orthography.

Index Terms: speech synthesis, dialect, adaptation, lexicon

1. Introduction

In our project on dialect (or sociolect) speech synthesis we aim at exploiting the effects of localization or regionalization [1] of voice user interfaces also for TTS systems. We are developing synthetic voices for different dialects, first contextualizing it in the region of Vienna/Austria. 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 are establishing 3 synthetic sociolect voices with speakers from Vienna. The adaptation methods and application scenarios that we develop should be 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 varieties 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 available from standard German. In a first step those are minimally adapted and applied to standard Austrian German (AT). Building upon this source, in a further step the lexicon and other procedures are adapted towards each of our Viennese varieties. To achieve this goal is not a trivial task and this may be one 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. However, this is not generally true if one thinks of Bengal or Arabic dialects.

2. Standard versus dialect

Generally the standard variety of a given language is taken as the defining level for all other language varieties. But often it turns out that there is more than one specific standard for many languages. This reflects the concept of a pluricentric language – a situation often found when language and the national identity of its native speakers do not coincide [2], [3]. In the case of German the German standard is often taken as prevailing over the Austrian and the Swiss standard, but this has rather to do with the number of speakers in the respective countries. There are ample differences in vocabulary and pronunciation which have to be reflected in the resources for the TTS system. Moving from a standard variety to dialect (or sociolect) varieties complicates the situation even more. For dialectal varieties, there are no standards for written input (orthography), the way how the dialects differ from the standard are not straightforward, and the differences may turn out to be gradual. Finally, it has to be ensured that the particular lexicon conveys those items comprised by a specific variety and blocks others.

Additionally speakers gradually shift from dialect(s) to standard. This creates a problem when one has to decide upon an appropriate speaker, since the material to be recorded has to be both consistent and characteristic for the language variety to be represented.

3. Representation of differences

In order to deal with potential shifts between standard and dialectal varieties, but also to be able to represent different dialectal varieties it is indispensible to establish a method for the transfer of linguistic information, independently of practical considerations (e.g., morphology, pronunciation, PoS-tags etc.). The two constraints guiding such a method are minimization of efforts and full coverage of ambiguities. So, for example, if one variant differs from the other only in vowel qualities and a neutralization of plosives, nothing has to be done beyond recording.

However, this is rarely the case. Rather, we expect differences on all levels of representation. The question always is whether these are fully or in part deducible from the common source, or whether one has to encode them for each variety separately. Table 1 gives a rough overview of what kinds of differences we have to deal with and which levels of representation these differences are assigned to.

Regarding lexical specificities it is evident that one needs a cascaded structure of lists of lexicon entries for items that occur only in a certain dialectal variety (or a set of related varieties). For words not occurring in one of those lists it has to be decided whether they still belong to the native stratum, or if they are pronounced in a shifted variety, i.e., the way the standard is pronounced within a certain variety. Since it has to be assumed that these lists are rarely complete this decision has to be estimated by inductive methods on the basis of (i) a morphological analysis and (ii) the orthography used in the input (see section 5 below) [4].

One level below the task gets more interesting. If there is a direct correspondence between the forms of the standard and a certain variety then the latter should be deductible

Accepted after peer review of full paper

from the basic form. However, to what extent this is possible depends on the number of potential ambiguities that arise with such a transformation.

Linguistic level	Austrian German Standard	Viennese	Coding level
sound	ɔ	ɛ	sound
symbol set – phon(em)es	a͡ɛ a	aː / æː / œː a / ɔ	lexicon setup
phonology	a͡ɛl# [va͡ɛl] *weil* 'because'	œː [vœːɐ]	rules
morphological	*pass-te* 'would fit' *Gläs-chen* 'glass dim.'	*pass-ert* *Glas-erl*	lexicon transfer
morpho-syntactic	*lesen können* 'can read' *ertrinken* 'drown'	*derlesen* *dersaufen*	
lexicon: – open class	*trinken* 'drink' *fett, dick* 'fat' *Kopf* 'head'	*saufen* *blad* *blutzer*	lexicon specific
– functional: – articles – pronouns	*der* 'the' *hinaus* 'to-out' *heraus* 'from-out'	*d' / da / der* *ausse* *aussa*	
phrasal: – clitica – infl. compl. – idioms – no preterit	*weil du weggehen sollst!* 'because you should leave' *er ging* 'he went'	*weilst di über d' häuser haun sollst!* *ea is gangen.*	text trans-lation

Table 1: *Levels of representation concerning differences between AT standard and Viennese dialect*

In the next section we introduce certain peculiarities of the phonemic system of the Austrian standard, as compared to the German standard. In the following we will investigate a (preliminary) transformational system from standard Austrian German to one Viennese variety. Certain ambiguities will not be resolvable beyond individual coding, but many others turn out as feasible challenges to a method that also comprises a certain extent of morphological coding together with a phonological implementation of well-formedness constraints.

The envisaged transformations are formalized as rules in the form of regular expressions over strings of phones. They do not necessarily have a phonological reality, nor are they purely phonetic in nature. Therefore we will refer to them as 'phone string rules', in order to avoid confusion.

4. Phone string rules

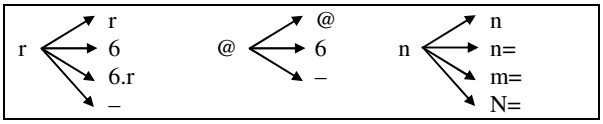
4.1. Rules for Austrian Standard (AT)

There are certain rules the application of which seems dispensable since the relevant phonetic properties will be unambiguously represented in the speech recordings.

- Glottal stop /ʔ/ is inserted at the beginning of onsetless morphological domains. (This is subject to variation, in the German standard, ʔ is realized more often, in the extreme with all onsetless syllables.)
- no voiced sibilants: /z/ → /s/, /ʒ/→ /ʃ/ (but in the variety of actors or radio speakers voiced fricatives are retained if no voiceless obstruent or fricative precedes within the same prosodic domain)

Other rules are conditioned by the phonological context. Here, it has to be tested, whether the unit selection can grasp the necessary contextual domain. For the TTS encoding we use German Sampa (GSAMPA) with some adaptations for Viennese dialects.

- syllabic nasals (accompanied with place assimilation) and syllabic /l/
- /r/-vocalization



word	gloss	German	Austrian
Lehrer	teacher	leː.r@r	lE:6.r6
werben	solicit	vEr.b@n	vE6.bm=
mehrere	several	meː.r@.r@	mE:6.r6.r@
Barbar	Barbarian	bar.'baːr	ba.'baː
fahren	drive	faː.r@n	faː.n=

Table 2*: /r/-vocalization and syllabic nasals in AT*

There is one case where the German standard variety displays a contextual merge of phonemes which has to be 'undone' in the Austrian standard.

- /IC/→ /Ik/ (if orthographically written as *-ig*, e.g., *richtig, König*, but not when written as *-ich*); /C/→ /k/ (e.g., *Chemie, China*)

word	gloss	German	Austrian
billige	cheap	bI.lI.g@	bI.lI.g@
billig	cheap	bI.lIC	bI.lIk
freilich	admittedly	fraI.lIC	fraI.lIC
Chemie	chemistry	Ce:.mi:	ke:.mi:

Table 3: *Palatal velar fricatives as velar stops in AT*

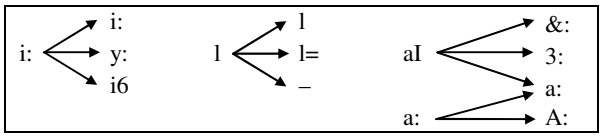
Since the output is ambiguous (the crucial information is found in the orthography), the relevant distinction has to be coded in the lexicon for the AT-variety.

4.2. Additional rules for Viennese dialect (VD)

In the Viennese varieties we have to deal with a number of additional processes. The first two turn out as rather unproblematic, since they can be covered by the local context and are fully represented in the acoustic data.

- neutralization of plosives in onsets, e.g., /t/ → /d/ (except /g/, /k/, but still in complex onsets: /kr/ → /gr/)
- lenition of lenis plosives intervocalically (and before syllabic nasals and /l/), e.g., /d/ → /D/.
- /l/-vocalization: as can be seen in the examples below, the output depends on the syllabic structure (vowel-shift ± deletion of /l/).
- schwa-deletions, most notably with the perfect prefix *ge-*. Before stops the whole prefix is deleted.
- vowel shifts: while some of them could be covered by the acoustic data, others are ambiguous in their output and have to be individually coded. (E.g., /a/ in AT is /A/ or /a/ in VD – foreign words are always /a/, native words are mostly /A/.)

With these facts in mind, it is plausible to argue for a component that implements these differences as a set of phone string rules.



word	gloss	German	Viennese	
weil	because	vaIl	v&:	(/&/ = [œ])
bleiben	stay	blaI.b@n	bl3:.Bm=	(/3/ = [æ])
zwei	two	tsvaI	tsva:	
Faser	fiber	fa:.z@r	fa:.s6	
Raben	ravens	ra:.b@n	rA:.Bm=	(/A/ = [ɔ])
viel	much	fi:l	fy:	
viele	many	fi:.l@	fy:.l@	
lieb	dear	li:p	li6p	
nie	never	ni:	ni:	

Table 4: *Vowel shifts and /l/-vocalization in VD*

5. Orthography

As there is no standard orthography for dialectal varieties and sociolects there are multiple spelling variants for many words in these varieties. These orthographic variants can be located on a scale between standard orthography and phonetic accurateness. (E.g., *gelacht – glacht – glocht – glochd* 'laughed'.) For text-to-speech synthesis it is necessary to manage textual input containing words that are proprietary to the dialect. One general problem in this context is the mutual homomorphism between different lexicon items as can be seen in Figure 1. To provide a robust text analysis module we decided to develop an orthography generation module that is capable of (re-)producing a ranked list of orthographic variants given dialectal text input. In this way the users do not have to know some 'correct' Viennese spelling variant that would be mandatory to synthesize speech.

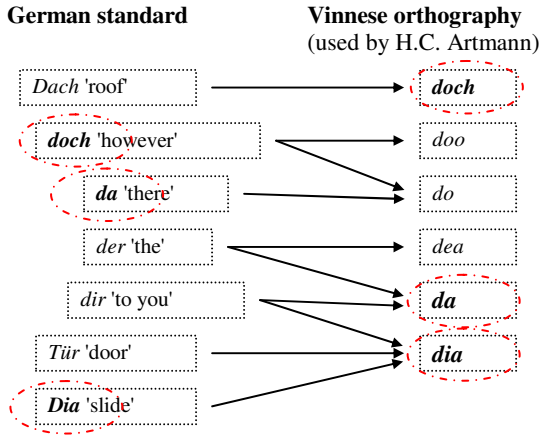


Figure 1: *Interferences between different orthographic varieties*

To derive the correct word identity based on the input orthography it is useful first to decide if the relevant string is part of the dialect or belongs to the standard. In the first case we determine all ambiguous (homomorphous) and unknown lexemes in the string and try to find the most similar words in the lexicon. Based on this information the word identity can be decided upon.

6. Experiments

6.1. Lexicon experiments

The experiments concerning the variety-specific lexicons reflect the fact that we are focusing on automatically labeled speech databases. The same lexicon is used for the segmentation and for the synthesis stage. Mismatches are thus only possible if a certain phone string matches multiple pronunciations in words present in the speech database.

In order to define the appropriate level of granularity of the phonetic labeling we conducted several experiments applying various versions of the lexicon for Austrian German and Viennese to the Festival unit selection synthesis system [5].

For a preliminary test on the required accurateness of transcription in the lexicon we built two unit selection voices based on the same recordings of an Austrian German speaker (comprising approx. 300 sentences) using a German lexicon (DE) and an Austrian German lexicon (AT). 30 utterances (consisting of a single word) were synthesized with both voices. They were selected to cover the already mentioned ambiguity between /C/ and /k/ not present in the primary source of the DE-lexicon. These were retrieved from a list of over 11600 words in the DE-lexicon that are misrepresented according to the AT standard. 23 of these utterances represent the context of suffix *–ig* that is pronounced as /Ik/ instead of /IC/, 7 others have /k/ instead of /C/ in onset position. For the evaluation we checked for both voices whether the synthesized prompts correspond to the AT pronunciation rules. In this way we evaluated the robustness of the voices and if it is necessary to encode these particular differences between Austrian German and German.

In the first part of the experiment where the /Ik/-context was tested it could be observed that using the correct lexicon improves the results. In the following tables '1' stands for a correct and '0' for an incorrect output of the unit selection engine with respect to the Austrian standard pronunciation.

utterance	gloss	DE	AT
Heiligabend	Christmas eve	0	1
Honigwein	mead	0	1
richtiggehend	fully fledged	1	1
Ewigkeiten	eternities	1	1
Ratlosigkeit	helplessness, perplexity	0	0
Verantwortungslosigkeit	irresponsibility	0	0

Table 5: *Examples for /IC/ → /Ik/*

The voice with the AT-lexicon produced only 2 errors, whereas the 'wrong' DE-lexicon had 18 errors in a sample of 23 utterances. Moreover, the 2 errors demonstrate that the speaker did not always pronounce the words in conformity with the AT-lexicon.

Utterance	Gloss	DE	AT
Alchemie	alchemy	0	1
Elektrochemie	electrochemistry	0	1
China	China	1	1
cherubinisch	cherubinic	1	1

Table 6: *Examples for /#C/ → /#k/*

Table 6 shows four examples from the second part of the experiment testing the syllable onset context, where we observe that both lexicons produce correct utterances if the position of the relevant phone is at the beginning of the

word. The reason for that is that the speech database only contains units with the correct pronunciation for this specific context. If the relevant phone string is in the middle of the word we actually do get errors with the lexicon that misrepresents the actual pronunciation (DE).

Although the sample in our experiment may not be statistically representative, the overall results show that in cases where ambiguities or deviations from the primary source are not predictable from the phonological context generating a lexicon that fits to a specific language variety strongly improves the accuracy of pronunciation.

Process	Lexicon DE		Lexicon AT	
IC → Ik	21,74%	5 / 18	91,30%	21 / 2
#C → #k	71,43%	5 / 2	100%	7 / 0
Sum	**33,33%**	10 / 20	**93,33%**	28 / 2

Table 7: *Correctness of utterances (in % correct #correct / #incorrect)*

In general we perceived that the synthetic voice with the German lexicon was of lower quality than the voice with the Austrian German lexicon. By inspection of the alignment we found more alignment errors in our test samples with the German lexicon. Hence, not only the unit selection can be improved adapting the lexicon to a specific language variety, but also the alignment and hence the quality of the speech database will be improved by adopting such an approach.

6.2. Orthography experiments

In the second set of experiments we investigated the question if our envisaged approach for orthography prediction is feasible and robust. For comparison we used different methods for finding the corresponding orthography given some other spelling variant. The two orthographic variants under consideration are a Viennese orthography created by H.C.Artmann, which is inspired very much by actual phonetics (VD_ART) [6], and one that is close to Standard German (VD_ST). (E.g., *lenxt / längst* 'long ago'; *himö / Himmel* 'sky'.) For the derivation in either direction we used a simple minimum-edit-distance approach (MIN), a minimum-edit-distance with character weights (MIN-C), where we used zero substitution costs for "similar" characters like 't' and 'd', and a decision tree-based derivation (CART) [7]. For the first two approaches we selected the most similar word from the lexicon. For the decision tree-based derivation we considered two different methods/tasks, where the first task was the correct prediction of a variety using the decision tree. In a second run we used the decision tree to derive a first orthographic variant, and the MIN/MIN-C to find the most similar string in the lexicon (CART-MIN, CART-MIN-C). The expectation was that the open domain task CART would perform worse than the other closed domain tasks. We consider the closed domain tasks as more relevant to our application where we want to find out the word identity given some arbitrary orthographic form.

Table 8 shows the prediction results for our first lexicon with 1423 entries. For evaluating the decision-tree-based method we used 90% (1281) for training and 10% (142) for testing. The MIN/MIN-C algorithms were applied to the whole lexicon.

The CART results are rather poor since this is an open domain task and there are a lot of three-character transformations that are not supported by the setup that we used for building the decision trees, such as '*sch*' → '*s*', '*l*' → '*erl*', '*x*' → '*ges*', or '*m*' → '*ben*' in the VD_ART/VD_ST transformation. In a future implementation we aim at including these transformations as well, since they occur quite frequently in the orthographic variants associated with Viennese dialects.

Method/Task	VD_ART to VD_ST		VD_ST to VD_ART	
MIN	37%	539 / 884	41%	590 / 833
MIN-C	44%	637 / 786	47%	669 / 754
CART	17%	25 / 117	16%	24 / 118
CART-MIN	63%	90 / 52	59%	84 / 58
CART-MIN-C	68%	97 / 45	63%	90 / 52

Table 8: *Orthography prediction rate (in % correct, #correct / #incorrect)*

The CART results include the failed alignments and the failed predictions. However, it could be shown that the combined strategy of using the CART prediction first and then searching for the MIN/MIN-C closest string in the whole lexicon works reasonably well with the limited amount of training data used.

7. Conclusions

We analyzed the differences between Austrian German and Standard German on multiple levels from phonetics to syntax. We showed that it is necessary to model certain phonetic differences between Austrian German and the German Standard for unit selection speech synthesis. This is true even more for modeling dialectal varieties. Furthermore we showed that a combined two-level approach is best for orthography prediction of dialects such as Viennese when having to cope with a limited amount of training data and a great variance of spelling conventions. This procedure is also interesting for other similar languages.

8. Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" is 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.

9. References

[1] Marcus, A. and Gould, E.W., "Crosscurrents – Cultural dimensions and global web user-interface design", Interactions *7*(4):32-46, 2000.

[2] Muhr, R. and Schrodt, R., "Österreichisches Deutsch und andere nationale Varietäten plurizentrischer Sprachen in Europa", Wien, 1997.

[3] Moosmüller, S., "Die österreichische Variante der Standardaussprache", in Krech, E.-M. and Püschel, U. [Eds] Beiträge zur deutschen Standardaussprache, 204-214, Hanau, Halle: Werner Dausien, 1996.

[4] Fitt, S. and Richmond, K., "Redundancy and productivity in the speech technology lexicon – can we do better?", Proc. Interspeech 2006.

[5] Clark, R. A. J., Richmond, K. and King, S., "Multisyn: Open-domain unit selection for the Festival speech synthesis system", Speech Communication, 49(4):317-330, 2007.

[6] Artmann, H. C., "Sämtliche Gedichte", Jung und Jung, Salzburg und Wien, 2003.

[7] Black, A. W. and Lenzo, K. A., "Building synthetic voices", http://festvox.org/festvox/festvox_toc.html.