Resources for speech synthesis of Viennese varieties

Michael Pucher[1], Friedrich Neubarth[2], Volker Strom[3], Sylvia Moosmüller[4], Gregor Hofer[3], Christian Kranzler[1], Gudrun Schuchmann[1], Dietmar Schabus[1]

[1] Telecommunications Research Center Vienna (FTW), Austria
[2] Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[3] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[4] Acoustics Research Institute (ARI), Vienna, Austria
E-mail: {pucher,kranzler,schuchmann,schabus}@ftw.at, friedrich.neubarth@ofai.at, vstrom@inf.ed.ac.uk, sylvia.moosmueller@oeaw.ac.at, gregor.hofer@ed.ac.uk

Abstract

This paper describes our work on developing corpora of three varieties of Viennese for unit selection speech synthesis. The synthetic voices for Viennese varieties, implemented with the open domain unit selection speech synthesis engine Multisyn of Festival will also be released within Festival. The paper especially focuses on two questions: how we selected the appropriate speakers and how we obtained the text sources needed for the recording of these non-standard varieties. Regarding the first one, it turned out that working with a 'prototypical' professional speaker was much more preferable than striving for authenticity. In addition, we give a brief outline about the differences between the Austrian standard and its dialectal varieties and how we solved certain technical problems that are related to these differences. In particular, the specific set of phones applicable to each variety had to be determined by applying various constraints. Since such a set does not serve any descriptive purposes but rather is influencing the quality of speech synthesis, a careful design of such a (in most cases reduced) set was an important task.

1. Introduction

Within the research project "Viennese sociolect and dialect synthesis" (VSDS) [1], 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 (Pucher et al., 2010).

In Vienna, language varieties are differentiated rather socially than regionally, therefore it would be correct to speak about sociolects. [2] 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.

This is the first attempt to develop multiple synthetic voices that represent different dialects of a certain language, as opposed to synthetic voices speaking with an accent, such as Alan[3] (English with a Scottish accent) or Anjali[4] (English with an Indian accent). These voices are based on a Standard English pronunciation dictionary and therefore can produce only systematic deviations from the standard pronunciation on the phone level.

Linguistic level	Austrian German Standard	Viennese	Coding level
sound	ə	ɐ	sound
symbol set – phon(em)es	ɑɐ̯ a	a: / æ: / œ: a / ɔ	lexicon setup
phonology	ɑɐ̯ l # [vɑɐ̯l] *weil* 'because'	œ: [vœ:]	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

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

Dealing with a dialectal language variety is a far more complex task. In Figure 1, we illustrate the various levels of linguistic information where differences between a dialect and some standard can be found. In our project we concentrated on speech synthesis, no attempts were made to implement an automatic translation between the standard and the dialect variety. However, many lexical and phrasal items specific for a dialect are stored in the lexical resources when they occurred in the input texts for the speech recordings.

[1] See: http://dialect-tts.ftw.at
[2] In urban varieties, the term "dialect" coincides with the sociolect spoken by the lower social classes. Henceforth we will use "dialect" for all non-standard language varieties.
[3] See: http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html
[4] See: http://www.research.att.com/~ttsweb/tts/demo.php

Representing a dialect primarily requires a specialized pronunciation dictionary, tailored for each recorded speaker, which reflects deviations from the standard variety on the relevant linguistic levels. To compile such a dictionary manually is a rather time-consuming task. Therefore, we developed methods to derive Viennese dialect dictionaries from a standard Austrian German dictionary using various sets of transformational rules and added only those entries manually that could not be captured by the rules or were ambiguous. Since the correct transcription is crucial for the success of automatic segmentation of speech, we still had to exact control over all items that actually occurred in the recording texts. In the process of speech synthesis, however, for out-of-vocabulary words it is necessary to rely on the automatic transformation methods that are relevant for all other dialects and sociolects.

2. 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. All these criteria are highly subjective, so we were also looking for more objective ones.

It has to be mentioned beforehand that we decided to engage professional speakers for the dialect voices (with the exception of youth language), rather than a genuine dialect speaker. This decision is based on the fact that genuine dialect speakers are usually not familiar with the task to read non-meaningful texts fluently, error-free, with constant prosody and in a studio situation. Moreover, recording a speech database for unit selection speech synthesis requires the reading of thousands of sentences. Consequently, instructing and monitoring a genuine speaker would take significantly more time and efforts than working with a professional speaker, such that employing a genuine dialect speaker for this task might put the success of the entire project at risk.

Regarding criteria for speaker selection, it is well known that automatic phone segmentation works much better for some speakers than for others. A clear and consistent pronunciation certainly helps, but there are other factors as well, which are not entirely clear yet. The best way to find out how well a speaker is suitable for unit selection speech synthesis is to actually make a voice and evaluate it in a listening test (Syrdal et al. 1998). That way, pitch tracking and pitch marking quality is tested as well, which also varies for reasons not always as obvious as for e.g. creakiness of voice.

The recording material for our test voices was selected among very short (3 words) sentences from EU parliament debates. These sentences contained no proper names, numbers, or abbreviations. 10 of these sentences were selected as test sentences and we assured that the training data covered all the diphones contained in the 10 test sentences.

The phone strings were derived from a standard German lexicon; the linguistic context features were: lexical stress, syllable boundaries, and word boundaries. It turned out that merely 93 sentences (consisting of 3 words each) provided enough data to cover the phonetic material in the test sentences. This was partly due to a partial overlap of phone strings and the relative shortness of the sentences. Because the phone segmentation had to work with a "flat start" (Young et al. 2006), this was intended.

For the "dialect" and the "colloquial" voice, we casted 10 professional speakers (mainly actors: 5 male and 5 female), recorded them and made tiny Standard Austrian voices, each approximately 2 minutes long, synthesized the test sentences and assessed them in a group meeting. Although this evaluation was still subjective, it certainly helped us making our decision with more confidence.

In addition, we also made 10 dialect unit selection voices, based on the material mentioned above, but transcribed with an orthography approximating Viennese . The dialect test voices gave us a better idea of how consistent the speakers were when reading in an 'unusual' orthography.

For choosing the appropriate voices we had to decide which speaker comes closest to an authentic Viennese dialect speaker, hence the realizations of dialectal speech produced by the professional speakers were compared with recordings of authentic Viennese dialect speakers. It turned out that none of the professional speakers was able to produce a prototypical Viennese dialect in such a way that it matches with an authentic dialect voice.

This result is based on the fact that speakers mimicking a certain variety usually capture the prominent features of such a variety (Torstenssen et al., 2004, Neuhauser, 2008), i.e., stereotyping takes place. Moreover, in stereotyping a variety, unusual linguistic patterning can be observed (Schilling-Estes, 2002). In our case, e.g., the mono-lateral realization of the lateral, a very salient feature of the Viennese dialect, was over-generalized by merely all speakers to phonetic contexts in which the mono-lateral realization is not allowed (Moosmüller, in print).

On the other hand, one also has to take into account that the overall majority of listeners have not too much direct acquaintance with authentic speakers of different varieties, consequently, in their expectations they rely on stereotypes rather than on authenticity. Therefore, in choosing an appropriate speaker, one has to balance the expectations of the listeners and the claim for the production of an authentic variety.

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

Table 1: Release voices

Although none of the professional speakers were authentic Viennese dialect speakers (i.e., they were not brought up in that variety), and all of them "misapplied" some phonetic features of the dialect, there were also some advantages. In particular, arguing from the

perspective of listener expectations, a certain degree of stereotyping is even preferable. To balance the degree of stereotyping and authenticity, we finally decided for an actor who came closest to an authentic Viennese dialect speaker, and an actress who had a very natural colloquial speaking style.

For the "youth language" variety, we proceeded in a similar way, with the only difference that we first pre-selected a specific group defined by age, school-type, gender, and variety spoken within the family.

3. 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. On the other hand, we paid particular attention to prosodic phrase boundaries: in order to cover diphones in phrase-final yes-no questions with rising intonation (in ToBI H-H%), we constructed 672 sentences of the form "article-noun-question mark". In order to cover diphones in front of continuation rises, we gave sentence-internal pauses a symbol different from sentence-framing pauses.

Thus we avoided to add yet another linguistic context feature for "boundary tone", e.g. in ToBI L-L%, H-H%, L-H% or "default". During synthesis, these tones are determined by punctuation and a list of interrogative pronouns serving as additional features at sentence level.

This quite large sample of texts, however, was designed on the basis of transcriptions corresponding to the Austrian standard, and could only be used for the colloquial voice and the Viennese youth voice, since both of these varieties resemble the Austrian standard enough on the transcription level. In particular, the relevant differences are phonetic to the largest extent, and therefore represented in the recorded speech itself. Still certain differences had to be respected, in particular that there are no preterit forms in either of the varieties and that certain lexical items do not exist, but have a distinct correspondent. Therefore, sentences ungrammatical in the Viennese varieties were either filtered out (partly automatically) or altered according to Viennese.

For the voice representing the prototypical Viennese dialect we had to employ additional measures. First, our recording text script for Viennese additionally contained a manually compiled set of sentences from various sources existing in various orthographic encodings, all of them representing "the Viennese dialect": e.g., sentences extracted from poems by H.C. Artmann, from songs by "Dr. Kurt Ostbahn", from a translation of the comic "Asterix" etc. Although these were clearly authentic Viennese texts, they were not sufficient with respect to diphone coverage, so we had to resort on texts from the standard variety. The speaker was instructed to translate the texts adapted to Viennese into proper dialect on the fly, a task that was unexpectedly easy to perform. The transcriptions of the text were transformed into dialect accordingly, utilizing the rules mentioned in section 1.

Initially, the pronunciation lexicon of the dialect covered only the texts from the authentic dialect sources, yet, it was growing until the very end of the project. Only then we decided between five competing phone inventories. Therefore, we had no choice but to assume that good diphone coverage in Standard Austrian directly correlates with a good coverage in Viennese dialect.

4. 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 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.

5. Voices

The release "Speech database for unit selection synthesis of Viennese varieties" contains data for 3 Viennese voices (Table 1). 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 tested at our website, and will be released for the Festival speech synthesis system (Black & Clark), in particular the open-domain unit selection Multisyn (Clark et al. 2005, 2007).

Category	Austrian German	Viennese dialect
vowel	a aː (ɔ̝) e̝ (e̝ː) eː i iː o̝ oː u uː y yː ø øː	a aː ɔ̝ ɔ̝ː e eː ɛ ɛː i iː ɪ o oː u uː ʊ y yː øː œ œː
di-/monoph-	a͡e a͡o o͡e	æː ɒː œː ɔ̝ɪ̯ oɪ̯ uɪ̯
thong/nasal	(æ̃ː) (œ̃ː) (ɔ̃ː)	ãː ɔ̝̃ ɔ̝̃ː ĩ æ̃ õ
r-vocalized	e̝ɐ e̝ːɐ iɐ iːɐ o̝ɐ oːɐ uɐ uːɐ yɐ yːɐ øɐ øːɐ	ɔ̝ɐ ɔ̝ːɐ e̝ɐ e̝ːɐ iɐ iːɐ o̝ɐ o̝ːɐ ʊɐ ʊːɐ (yːɐ) øːɐ
schwa	ə ɐ	ə ɐ
plosive/spirant	b d g p t k	b d g ß ð ɣ p t k
fricative	f v s ʂ ʃ ʒ ç x h	f v s sː ʃ ʒ ç x h
liqu./nas./glide	ʀ l m n ŋ j	ʀ l l̩ m m̩ n n̩ ŋ ŋ̩ j
pause/glottis	*'sil' 'pau'* ʔ	*'sil' 'pau'* ʔ

Table 2: Phone sets for Austrian German and Viennese.

Table 2 shows the maximal sets of phonetic labels for speech segments on the phone level. They are represented with IPA symbols; however, within our project we only

worked with a version of German-SAMPA adapted to the needs to represent also Viennese dialects. The coding for Austrian German is in accordance with the phonetic analysis presented in (Muhr, 2007), the coding for Viennese dialect reflects an analysis we achieved during the project. Phones in brackets indicate that these are not genuine members of the native set.

These sets are the basis for transformed and reduced sets used in the phonetic coding of the lexica for speech synthesis (Pucher, Neubarth & Strom 2010). We designed a set of transformational rules that would merge certain phone classes, or split certain diphthongs resulting from r-vocalization. We evaluated the resulting subsets in three experiments. The first is concerned with phone error rate of letter-to-sound rules. Figure 2 shows phone error rates for 5 random splits of the lexicon derived from texts from Artmann. Each of the 5 phone sets was tested with held-out data from this lexicon, and also with the entire lexicon derived from the Viennese translation of Asterix-comics.

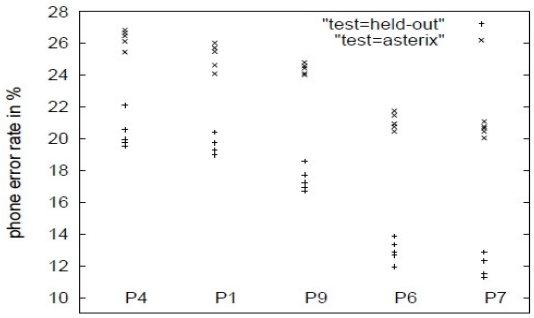

Figure 2: Phone error rates for letter-to-sound-rules and phoneme sets

The other two tests were performed with the actual voice: on a sample of 27 test sentences designed in such a way that they would contain lexical material the mentioned rules would be sensitive to, we counted the number of missing diphones that would have to be replaced by less appropriate units. Here the results were just the opposite of those from the first experiment. In the third test, we had 8 listeners perform a pair-wise comparison of all test sentences synthesized with all 5 potential voices. Although not all comparisons were statistically significant, the coding with average results from the first two tests (P9) fared slightly better than the other ones.

6. Summary

We described the building process of synthetic voices for Viennese varieties. The methodological approach can be generalised to the building of synthetic voices for social and regional varieties in general. We have already demonstrated the use of our synthetic voices within a dialog system designed as a restaurant guide, where types of restaurants are associated with a certain social variety. We hope that the public releases of our voices will find interest among other researchers and developers, and that new applications are realized with these resources.

In our future work we want to focus on the rapid prototyping of dialect and sociolect synthetic voices, which can be realized with adaptive parametric speech synthesis approaches.

7. Acknowledgements

The project “Viennese Sociolect and Dialect Synthesis” was 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.

8. References

Black, A. W., Clark, R., *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/.

Clark, R., Richmond, K., King, S. (2005). Multisyn voices from ARCTIC data for the Blizzard challenge. In *Proc. of Interspeech 2005*.

Clark, R., Richmond, K., King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication 49/4*, pp. 317-330.

Moosmüller, S. (in print), *Stereotyping the Viennese Dialect.* Stauffenburg Verlag: Tübingen.

Muhr, R. (2007). *Österreichisches Aussprachewörterbuch Österreichische Aussprachedatenbank*. Peter Lang Verlag, Frankfurt.

Neuhauser, S. (2008). Voice disguise using a foreign accent: phonetic and linguistic variation, *The International Journal of Speech, Language and the Law, 15/2*, pp. 131-159.

Pucher M., Neubarth F., Strom V. (2010) Optimizing Phonetic Encoding for Viennese Unit Selection Speech Synthesis. In A. Esposito et al. (eds.) *Development of Multimodal Interfaces*, Proc. of the 2nd COST 2102 Intern. Training School, Dublin, March 2009.

Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V. (2010). Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis, *Speech Communication, 52/2*, pp. 164-179. http://dx.doi.org/10.1016/j.specom.2009.09.004

Schilling-Estes, N. (2002), Investigating Stylistic Variation, in: Chambers, J.K. et al. (eds.), *The Handbook of Language Variation and Change*, Oxford, Cambridge: Blackwell Publishers, pp. 375-401.

Syrdal, A., Conkie, A., Stylianou, Y. (1998). Exploration of acoustic correlates in speaker selection for concatenative synthesis, In *Proceedings of ICSLP`98*.

Torstenssen, N., Eriksson, E.J., Sullivan, K.P.H. (2004). Mimicked accents. Do speakers have similar cognitive prototypes? In *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Macquarie University, Sydney: Australian Speech Science & Technology Association Inc., pp. 271-276.

Young, S.J. et.al. (2006). *The HTK book version 3.4*, Manual. Cambridge University Engineering Department, Cambridge, UK.