Optimizing Phonetic Encoding for Viennese Unit Selection Speech Synthesis

Michael Pucher[1], Friedrich Neubarth[2], and Volker Strom[3]

[1] Telecommunications Research Center Vienna (ftw.), Vienna, Austria
pucher@ftw.at
[2] Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
friedrich.neubarth@ofai.at
[3] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
vstrom@inf.ed.ac.uk

Abstract. While developing lexical resources for a particular language variety (Viennese), we experimented with a set of 5 different phonetic encodings, termed phone sets, used for unit selection speech synthesis. We started with a very rich phone set based on phonological considerations and covering as much phonetic variability as possible, which was then reduced to smaller sets by applying transformation rules that map or merge phone symbols. The optimal trade-off was found measuring the phone error rates of automatically learnt grapheme-to-phone rules and by a perceptual evaluation of 27 representative synthesized sentences. Further, we describe a method to semi-automatically enlarge the lexical resources for the target language variety using a lexicon base for Standard Austrian German.

Keywords: Speech synthesis, language varieties, phonetic encoding, graphem-to-phone, pronunciation lexicon.

1 Introduction

Data driven methods for speech synthesis, such as unit selection speech synthesis, or more recently HMM-based methods, induced a shift in perspective on various levels of speech processing. One of these levels is phonetic coding which is used as the prime lexical resource. But it has not the status of an independent, linguistically motivated system, anymore. Rather should the resources, specifically the set of symbols used therein, be adapted towards the data itself, reflecting the need to reconcile several conflicting tradeoffs that have to be handled in an optimized way.

The task in speech synthesis is to produce an acoustic output (speech signal) from a string of symbols defined as phonetic (or phonological) units. Such strings are retrieved from a lexicon or derived by grapheme-to-phoneme conversion which can be rule based or based on statistical methods. The set of symbols has usually been taken as given by definition, but as soon as one tries to transcribe actual speech from a certain language variety, the applicability of such a

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 207–216, 2010.
© Springer-Verlag Berlin Heidelberg 2010

set will easily come under scrutiny. A step further, when we attempt to cover varieties further away from a given standard variety (which is generally defining the coding in the linguistic resources), this problem becomes evident very quickly. Within a data-driven approach, however, whatever fine-grained differences there might be, most of the (phonetic) subtleties are covered by the data itself. Here, the task is to retrieve the optimal sequence of sound segments or models more or less directly from the data. An additional dimension emerges when we want to derive the phonetic encoding of language varieties from a standard resource.

The speech synthesis system we developed is based on the Festival Multisyn unit selection synthesis system [1]. Regarding the symbolic encoding used in the pronunciation dictionary there are three tasks with diverging constraints:

1. Automatic segmentation of speech data requires the models to be as distinct and coherent as possible (→ rich phonetic transcription, many symbols), but prefers many instances of phones for building the models combinations (→ condensed phonetic transcription, few symbols).
2. Unit selection requires that target segments are unambiguously retrievable (→ rich phonetic transcription), but also requires high coverage of segment combinations, i.e. the sparsity problem (→ condensed phonetic transcription).
3. Graphem-to-phone conversion methods (for unknown word handling) prefer less classes in the output, thus having a smaller potential to make errors (→ condensed phonetic transcription).

Here we are focusing on unit selection synthesis where the optimization of the phone set is vital. In HMM based synthesis there is already a built-in optimization of the phone set by means of context clustering. Only phones, which are relevant according to the data are used in the clustering [2].

It would of course be desirable to have methods for automatically deriving a phone set from a corpus of recordings [3]. Since these methods are not robust enough, yet, we believe that our approach, using multiple phone sets to segment and synthesize speech and evaluating them through subjective listening tests, is justified.

In [4] we already described the methods how to model language varieties (Viennese dialects/sociolects) using a common language resource (Standard Austrian German). Within that project we gained the insight that it is not sufficient to simply define some alternative set of phone symbols and certain rules or methods to obtain the appropriate phonetic transcriptions from the standard resources. The problems are lexical and morphological differences, ambiguous mappings of phones and finally, the target set of phones itself can be disputed, especially in the light of a certain degree of variability regarding the phonetic realization of various phones in a given language variety (including the possibility that speakers do not strictly adhere to only one variety.)

Therefore we decided in a first step to encode a preliminary sample of lexical entries for the Viennese varieties in a phonologically rich form. This means that beyond a mere analysis on a (disputable) phonemic level, we also encoded systematic phonetic or contextually motivated differences, such as intervocalic lenition, final-devoicing of plosives, etc. as well as distinctions with uncertain status,

such as open/close mid-vowels. In a next step we designed a set of phonologically motivated rules that operate on these codings, applying various mappings or merges in order to obtain smaller sets of symbols. Five of these sets were used to build synthetic voices which in turn were used to evaluate the qualities of each of these sets.

In the next session we describe the linguistic background of the language varieties (Viennese vs. Standard Austrian German) with special focus on problems occurring during voice building. In section 3 we present the different sets of phonetic symbols and the transformational rules with which we obtain them. Section 4.2 describes the test we performed with these sets in connection with voice building and the results of an evaluation on the resulting unit selection voices, and in section 4.1 we show how these results correlate with the performance of the relevant phone symbol sets within grapheme-to-phone conversion tasks. Finally, we give an outline of the architecture of the methods we use to obtain a large-scale lexicon for each of the language varieties.

2 Linguistic Background

Modeling language varieties for speech synthesis is a challenging task from an engineering viewpoint, but also from a phonological and phonetic perspective there are several questions that demand clarification: i) what is the set of phones in a certain variety, ii) are there clear correspondences between this set of phones and the standard variety, iii) can these correspondences be formulated in terms of phonologically motivated transformations, and iv) how consistent is the variety actually used?

Starting with the last question it turns out that speakers regularly oscillate between various Viennese varieties [5]. This may have to do with the fact that Viennese varieties are rather sociolects than dialects, hence associated with social groups rather than regions. (It seems that regional varieties associated with certain districts in Vienna, as described in the literature, have been lost some time ago.) However, speakers may want to signal a certain amount of affinity to a social group by using a specific language variety or at least displaying certain phonetic aspects of this variety. Nevertheless, from an engineering point of view sociolects and dialects behave alike, they are varieties of a certain language, either defined socially or regionally. Other triggers for one or the other variety are lexical: certain words or word forms (e.g., preterite) do not exist in Viennese, a speaker is automatically forced to perform a certain shift in speaking style. During recording we tried to exercise as much control as possible on these factors.

Regarding the first three questions the answers are positive, with certain provisos: i) some phones still exhibit a high degree of variation (mid vowels, lenis plosives/spirants) such that uncertainties remain, ii) various correspondences between phones contain ambiguities. (Examples: [a] → [a], [ɒ]; [a͡ɛ] → [æː], [a]), and iii) certain transformations have unclear status. The most prominent example for such a transformation is final devoicing, which is clearly operative in Standard Austrian German, but much less obvious in Viennese. It seems that the domain

it applies to is the prosodic phrase, not the morpheme/word domain, and for sure it does not apply when a clitic pronoun starting with a vowel follows the consonant.

The strategy to deal with all these factors is to start with a basic lexicon that uses a symbol inventory designed upon phonological considerations. Referring to the last example, final devoicing is coded by a diacritic, intervocalic stops that may or may not have a phonetic realization as spirants are coded specifically as such etc. Of course, for the purpose of defining a symbolic base for a unit selection algorithm, this set is too rich and may lead to sparse data or even inappropriate classifications during voice building and unit selection. However, while building upon such a resource one can think of reintegrating the lexicon with a certain sets of rules in order to obtain phonetic representations of lexical entries and symbol sets that are more sound in number and hopefully more appropriate towards the data. Hope alone is not enough, therefore we designed several sets of transformational rules and a series of tests to be discussed in the remainder of this paper in order to assess the quality of the overall output in relation to the rule sets.

3 Phone Sets

Table 1 contains the description of the transformational rules that were used for defining the relevant phone sets. Most of them merge certain classes of phones, often sensitive to the phonological contexts, only two of them split complex phones (diphthongs, vowel-ɐ combinations) into smaller phone units.

Since it is impossible to test the effect of a single transformation rule applied to the set of phones in isolation, we designed an array of rule sets where each rule has some phonological motivation and chose 5 of these sets for further evaluation.

Table 1. Description of rules defining phone sets

Rule	description
merge_eschwa	merge [ə] with [ɛ]
merge_a_aschwa	merge [aɐ] with [a]
merge_a_aschwa_l	merge [aɐ] with [aː], add length
split_Vaschwa	split V-ɐ diphthongs into separate phones
split_diphthong	split all diphthongs into separate phones
rem_V_nasal	merge nasal vowels with non-nasal
neut_mid_v	merge tense mid vowels w. lax: [e] → [ɛ]
findev	merge final lenis w. fortis (fin. devoicing)
rem_findev	merge final lenis w. lenis (no fin. dev.)
merge_spirants	merge spirants with lenis, nasal or [v]
despirantize	merge spirants w. lenis plosives: [β] → [b]
rem_syllabic	merge syllabic consonants w. non-syllabic
rem_nons_gem	merge long consonants w. short, exc. [s]
rem_length	merge all long phones with short: [aː] → [a]

Table 2. Definition of phone sets by rules

Rule	P1	P4	P6	P7	P9
merge_eschwa	√	×	×	×	×
merge_a_aschwa	√	√	×	×	×
merge_a_aschwa_l	×	×	×	×	√
split_Vaschwa	×	×	×	√	×
split_diphthong	×	×	×	√	×
rem_V_nasal	×	×	√	√	×
neut_mid_v	×	×	×	×	√
findev	√	×	×	×	×
rem_findev	×	√	√	√	√
merge_spirants	×	×	√	√	×
despirantize	×	×	×	×	√
rem_syllabic	×	×	√	√	√
rem_nons_gem	√	√	×	×	√
rem_length	×	×	√	√	×
Number of phones	75	76	47	39	66

Table 2 shows the definitions for the different phone sets where √ stands for applying the respective rule, whereas × means that it is not operative. In the last row the number of symbols within the resulting phone set is shown, but notice that this number applies only to data obtained from Viennese dialect sources (see section 4.1); when data from the 'transformed' Austrian Standard is included, the numbers increase by 3-4 symbols.

The transformation rules only affect the set of symbols, not the lexical representations themselves. The rules do not give ambiguous outputs, so it is possible and easy to generate the corresponding lexical sources. In the following we describe various tests designed in order to evaluate the quality of the overall output with each of the variants of symbolic encoding.

4 Evaluation

4.1 Evaluation of Phone Sets for Automatic G2P Rules

The quality of automatic grapheme-to-phone conversion depends on the coherence or the mapping between graphemic symbols and phone symbols. Therefore we decided to use G2P methods as indirect evidence for the coherence of a given set of phones. For the initial recordings of Viennese dialect/sociolect, we used texts for which an orthography exists that reflects the phonological properties of Viennese dialect at least to a certain degree. These texts were used to automatically learn G2P rules from them. Since the 4 groups of texts obey different standards regarding orthography, we also created different lexica for each of the groups. It has to be mentioned that although the text sources listed below belong to different text genres (poetry, comics, plain text, songs lyrics), they were treated the same way while recording the speech data: each of the texts was split

into a set of isolated sentences. The speakers had to read each of these sentences as a separate item, thus minimizing the chance of co-textual influences.

- **artmann.lex:** Isolated sentences from poems by H. C. Artmann [6] ("med ana schwoazzn dintn"). Orthography very close to the actual pronunciation of the dialect and very consistent. (1614 words)
- **asterix.lex:** Sentences from the comic "Asterix" in the Viennese translation by H. C. Artmann. Orthography less coherent due to mimicking of other varieties including the standard variety by orthographic means. (1194 words)
- **wean.lex:** Sentences from various sources, containing typical Viennese words and phrases. Orthography rather inconsistent. (1473 words)
- **ostbahn.lex:** Sentences from songs by Dr. Kurt Ostbahn. Orthography oscillates between orientation towards pronunciation and standard orthography. (391 words)

We were concerned that the standard Festival G2P rule learner is no longer state-of-the-art or inadequate for such small lexica. There was a "Letter-to-Phoneme Conversion Challenge" planned for 2006, but due to illness of the host it was never completed. However, preliminary results suggested [7] that Marelie Davel's and Etienne Barnard's called "Default & Refine" [8] works better for small lexica than "Pronunciation by Analogy" [9]. Marelie Davel kindly provided us her implementation of D&R.

The evaluation described in this section was motivated by the following questions: How consistent are the four sub-lexica? How much consistency do we lose by combining them? And how do the different phonetic encoding schemes fare with the G2P methods?

Since the sizes of our sub-lexica are different and we wanted comparable phone error rates, we evaluated the G2P rules by repeated random sub-sampling validation instead of k-fold cross-validation. This method randomly splits the dataset into training and validation data. For each such split, the classifier is retrained with the training data and validated on the remaining data. The results from each split can then be averaged. The advantage of this method (over k-fold cross validation) is that the proportion of the training / validation split is not dependent on the number of folds, i.e. if we want to compare the consistency of two sub-lexica, we can choose the amount of training data to be equal. The disadvantage of this method is that some observations may never be selected in the validation subsample, whereas others may be selected more than once. However, when the number of repetitions is large enough, our estimates of phone error rates should be reliable enough.

Our assessment of consistency was confirmed: artmann.lex is most consistent, followed by asterix.lex, and wean.lex. Figure 1 shows that for phone set P9 and a subset of 1200 words taken from artmann.lex only for training, the phone error rate in the held-out data is about 17%. Figure 1 also illustrates that artmann.lex is not consistent with asterix.lex. When 1200 words of asterix.lex were used for training, the phone error rate for the held-out data was around 19%, and testing with artmann.lex resulted in about 26% phone error rate.

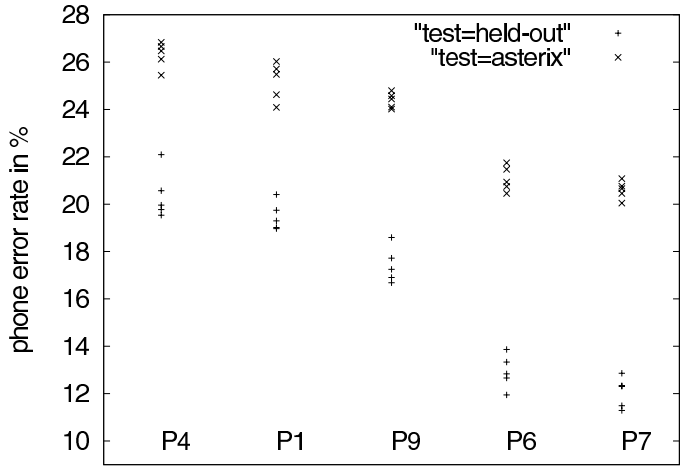

Fig. 1. Phone error rates, shown for 5 random splits of the artmann.lex: 1200 words were used for training, the remaining ones for testing. Each of the 5 phone sets was tested with the held-out data, and also with the entire asterix.lex.

Regarding the coherence of the tested phone sets, the results are a bit disappointing: the smaller the set of phones, the better the performance of the G2P component. One might be surprised that number is the only effective parameter in this experiment. Since the respective phone sets display different context sensitive splits and merges of phone symbols it could be possible that due to a better mapping to orthography one set with a larger number of phones outperforms the others. This is not the case. The conclusion we can draw from this finding is that the claim that a lower number of symbols enhances the performance of G2P methods is correct. This does, however, not tell much about the performance of unit selection speech synthesis, which will be the topic of the next section.

4.2 Evaluation of Phone Sets for Synthesis

For this evaluation we had 8 listeners that had to make pairwise comparisons between 27 prompts synthesized with the respective phone sets (270 comparisons in total). The unit selection voices are built from recordings of our male Viennese speaker. Since we primarily wanted to assess the segmental quality of the synthetic voices (or better: the different phone sets underlying them) relative to the perceived authenticity of the dialect, the subjects definitely had to be acquainted with Viennese dialect, but not to be native speakers of this dialect. The synthesized soundfiles are encoded in 16 bit, 16 kHz sampling rate and were presented to the subjects over a web-based application, the actual setting was that the subjects listened to the synthesized sentences with headphones.

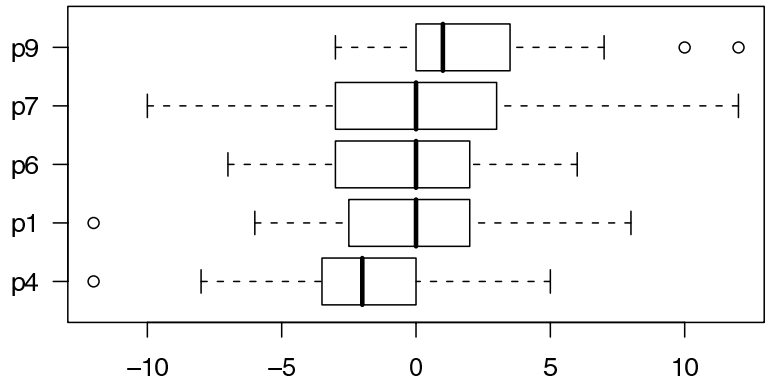

Fig. 2. Box-plots of pairwise comparison score for voice samples generated with different phone sets

Differences between P9 and P1, P9 and P4, and P9 and P6 turned out to be significant ($p < 0.05$) according to a Mann-Whitney-Wilcoxon test. In this evaluation, P9 scored as the best phone set.

Figure 2 shows the results of the pairwise comparisons for the different phone sets. The data for one voice i using a certain phone set consists of scores $s_j = w_{ij} - l_{ij}$, where $j \neq i$ and w_{ij} and l_{ij} are the numbers of comparisons won and lost, respectively, of voice i against voice j per listener.

The good performance of the P9 set can be partly explained by taking into consideration the actual diphone coverage of the 27 test sentences for the different phone sets. This would explain the relative superiority of P9 (only 4 diphones missing) over P6 (26 missing) and P7 (21 missing), but such a line of reasoning would leave unexplained the relatively bad performance of P1 (10 missing) and P4 (12 missing). What is important to note is that missing diphones do not produce gaps in the output, but invoke backoff rules individually defined for each phone set which are much more complex than the backoff rules used in the standard Festival multisyn system (always replacing a missing vowel with schwa). So what we actually evaluated were not the phone sets in isolation, but the phone sets together with their associated backoff rules, taking as implicit parameters segmental quality of the speech synthesis output and dialectal authenticity.

Although the significance of a listening test with only 27 test sentences can be disputed, we take it as the most indicative test for the overall quality of the system. Interestingly, the phone set that turned out as the best among the alternatives (P9) is the one with the most balanced number of phones. P6 and P7 gain their lower numbers mainly by merging nasal vowels with non-nasals and by abandoning length contrasts. P1 and P4 have more phone symbols because they retain spirants and syllabicity of nasals, which can be retrieved by the phonological context. It would be desirable to assess the question which types of processes increase or decrease the overall performance, but due to the large number of possible combinations it seems impossible to investigate the behavior of just one process/transformation rule in isolation.

Regarding the tasks with their intrinsic constraints presented in the introduction, it may be possible to optimize them separately. This would require an objective measure for transcription accuracy for task 1 (automatic segmentation) and an objective or subjective measure for constraint 2 (unit-selection). However, these measures are partly hard to obtain and a combination of the performance for each of the constraints is needed anyway, we decided to do a joint optimization through the subjective listening test described above.

5 Excursus: Conversion Rules for Dialects

To extend the available data for building dialect voices we recorded a large amount of speech data where the speaker had to read a text presented in standard German orthography. The speaker was instructed to "translate" the text into Viennese "on the fly", i.e., to use Viennese pronunciation of the words whenever possible. We tried to control the factors that would force the speaker to switch to the standard variety.

The problem was that while there exist lexical entries for all the words contained in the sentences, the phonetic coding corresponds to the Standard Austrian German variety. Therefore we defined a set of rules, similar to the rules used to obtain the different phonetic encodings, to transform these phonetic strings into Viennese dialect. These rules, however, produce multiple variants for many words: either the rule has ambiguous output per se or it is not predictable whether the rule applies or not.

However, for the automatic phone segmentation of the recorded speech data we generated lattices using all pronunciation variants from the transformed lexicon. During the segmentation process one variant is selected as the best fitting phonetic transcription given the acoustic data. With this kind of feedback loop it is possible to eliminate variants that do not exist or are wrongly predicted.

These transformation rules are primarily intended to obtain a phone segmentation of the recordings. The texts were selected for diphone coverage (with lexical stress, word, and syllable boundaries) by using the standard pronunciation, because at that stage we could not employ a validated Viennese pronunciation lexicon covering these texts. This compromise was based on the assumption that a good coverage of units and combinations of units in Standard Austrian German would coincide with a good coverage in the Viennese dialect/sociolect.

6 Conclusion

By meeting the challenge of creating synthetic voices of language varieties (in our case Viennese sociolect/dialect) by using resources developed for the standard variety, we faced the interesting fact that the optimal phonetic encoding is by no means straightforward. Since the task is neither to obtain maximal phonetic accuracy nor to develop a perfect phonological representation, the optimal encoding has to be decided upon by aspects of engineering, in particular G2P

conversion, automatic segmentation, and, most important, unit selection synthesis. Based on our phone set evaluations in Section 4.1 and Section 4.2 one phone set turned out to be the best one for encoding Viennese sociolect/dialect. Still, there was an explicit trade-off: by choosing a set with an average number of phones we accept a higher phone error rate for the G2P rules, but get a better synthesis quality according to the subjective evaluation.

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.

References

1. Clark, R., Richmond, K., King, S.: Multisyn voices from ARCTIC data for the Blizzard challenge. In: Proc. Interspeech (2007)
2. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. of Eurospeech, September 1999, pp. 2347–2350 (1999)
3. Aylett, M.P., King, S.: Single speaker segmentation and inventory selection using dynamic time warping, self organization, and joint multigram mapping. In: 6th ISCA Speech Synthesis Workshop, Bonn, Germany (2007)
4. Neubarth, F., Pucher, M., Kranzler, C.: Modeling Austrian dialect varieties for TTS. In: Proc. Interspeech 2008, Brisbane, Australia (2008)
5. Moosmüller, S.: Soziophonologische Variation im gegenwärtigen Wiener Deutsch. Franz Steiner Verlag, Stuttgart (1987)
6. Artmann, H.C.: Sämtliche Gedichte. Jung und Jung, Salzburg und Wien (2003)
7. Damper, R.: Personal communication (June 2008)
8. Davel, M., Barnard, E.: Pronunciation prediction with Default & Refine. Computer Speech and Language 22(4) (2008)
9. Damper, R., Stanbridge, C., Marchard, Y.: A Pronunciation-by-Analogy Module for the Festival Text-to-Speech Synthesiser. In: SSW4 (2001)