Evaluation of Automatically Generated Transcriptions of Non-Native Pronunciations using a Phonetic Distance Measure

Stefan Schaden

D-45147 Essen, Germany
stefan.schaden@freenet.de

Abstract

The paper reports on the evaluation of a rule-based technique to model prototypical non-native pronunciation variants on the symbolic transcription level. This technique was developed to explore the possibility of an automatic generation of adapted pronunciation lexicons for different non-native speaker groups. The rule sets, which are currently available for nine language directions, are based on non-native speech data compiled specifically for this purpose. Since manual phonetic annotations are available for the speech data, the evaluation was performed on the transcription level by measuring the phonetic distance of the automatically generated pronunciations variants and actual pronunciations of non-native speakers. One of the central questions to be addressed by the evaluation is whether the rules have any predictive value: It has to be determined if and to what degree the rules are capable of generating realistic pronunciation variants for previously unseen speakers. Secondly, the rules should not only represent the pronunciations of individual speakers adequately; instead, they should be representative of speaker groups. The paper outlines the evaluation methodology and presents results for selected language directions.

1. Introduction

Foreign accents often pose a challenge not only for linguistics and second language acquisition research, but also for some applications of speech technology. The range of phonetic variation in this domain is particularly broad, since numerous speaker-related factors are involved as potential causes of pronunciation errors. As a result, the deviations from the canonical ('correct') target language pronunciation that we encounter in non-native speech are often difficult to predict and even more difficult to model on a technical level.

In Schaden (2003), a technique was described to model prototypical non-native pronunciation variants on the transcription level using sets of phonological rules. For each language direction (a combination of a native language L1 and a target language L2), there is one rule set that models the most characteristic non-native pronunciation errors and introduces these errors into canonical pronunciation dictionaries. The hand-crafted rules, available for nine language directions, are based on empirical non-native speech data (see Schaden & Jekosch, 2006, these proceedings) that was recorded and phonetically transcribed.

The present paper reports on the evaluation of this rule-based approach. The evaluation method will be described, as well as exemplary results that indicate the central benefits and problems of the rule-based technique. Key questions addressed by the evaluation are (1) the predictive value of the rules, i.e. if and to what degree the rules are capable of generating realistic pronunciation variants also for new speakers, and (2) cross-speaker representation, i.e. the question whether the rules can model regularly occurring errors not only for individual speakers, but also for speaker groups.

2. Evaluation Method

2.1 System-based Evaluation

A general decision that has to be taken before engaging in an evaluation of phonetic transcriptions is whether the evaluation addresses either (a) the linguistic adequacy of the rules (linguistic evaluation) or (b) their suitability for a particular speech system and its impact on the system performance (application-based evaluation).

Application-based evaluations generally proceed along the following lines: First, a subcomponent of a speech system (e.g. ASR, TTS) is identified as a potential candidate for improvements or adaptations. Secondly, this component is detached from the overall system and modified in particular aspects. Lastly, after re-integrating the modified subcomponent into the system, the impact of the modifications on the performance is rated by comparing it to the baseline system. Thus, the yardstick for the success of the modification is the overall system performance (according to commonly applied criteria such as word error rate).

In the field of phonetic transcriptions, an application-based evaluation was pursued e.g. by van Bael et al. (2003), who rated the quality of transcriptions according to their effects on ASR performance. This approach is reasonable whenever the target application is known in advance. However, the rule-based modeling technique we developed was designed with no particular speech system in mind. It was left open from the outset whether the target application will be ASR, TTS, or any other technical or non-technical application. In this case, an application-based evaluation is not preferable for various reasons.

For instance, if we decide to measure the effects of lexicon modifications on the overall performance of an ASR system, we must expect that the results will be influenced by specific interactions of the lexicon with other system components (in ASR systems, the interplay between acoustic models and lexicon is one of these influences). Therefore, the performance values obtained by such measurements do not necessarily indicate the quality of the transcriptions themselves; rather, they indicate whether the lexicon modifications are *suitable* for the particular system that is being tested. So whatever the specific effects of a potential interaction of system components are, we must expect that the results reflect the quality of the lexicon modifications only in an indirect way.

For these reasons, we employed a method to determine the performance of the rules without incorporating them into an application. The evaluation is performed on the very same level of linguistic representation on which the rule system itself operates, i.e. on the level of phonetic

transcriptions. This approach has the advantage that it indicates the transcription quality itself; however, it does not tell us about the impact of the modified transcriptions on the performance of any speech system.

2.2 Outline of the Method

The rule-system that generates the foreign-accented transcriptions does not aim to reproduce *all* pronunciation errors made by non-native speakers in an exact manner. Instead, it attempts to *approximate* the most characteristic errors in the best possible way. The rules are designed to generate multiple variants that reflect so-called *accent levels* (for details see Schaden, 2003). Accent levels model different L2 proficiency levels of speakers (ranging from near-native pronunciation to gross mispronunciations), each of which is associated with its characteristic pronunciation errors. The current model is based on four levels from 1 to 4, where higher integers indicate increasing deviations from the canonical form. The topmost level 4 is a heavily accented pronunciation that follows almost completely the grapheme-phoneme correspondences of the speaker's L1 (e.g. German [bɪʁmɪŋham] for English *Birmingham*).

Given this design of the rule system, it is the aim of the evaluation to compare the automatically generated variants to the pronunciations encountered in empirical speech data and to determine the degree to which they approximate this data. In order to quantify this relation, we employed a phonetic distance metric that operates on the transcription level (see section 3 below).

Obviously, this evaluation method relies on the availability of transcriptions of pronunciations of non-native speakers in sufficiently large numbers. We used a set of manual transcriptions that was prepared in the course of the research. Transcriptions were available for a vocabulary of 215 European city names from different countries as well as 50 short sentences from various L2s spoken by speakers of four native languages. As a benchmark for the rule-based transcriptions, the canonical L2 transcriptions were used. The performance of the rules is rated by their capability to achieve a better approximation to the empirical pronunciations than the canonical L2 form. This principle is depicted in the following figure:

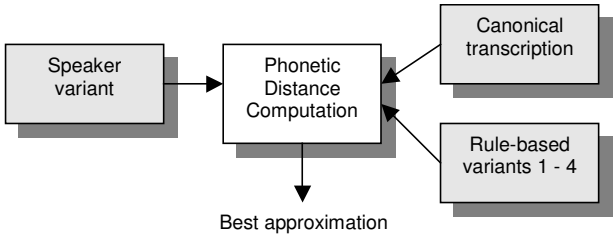

Figure 1: General principle of the evaluation

Since it is one of the objectives to determine the approximation of the rules for entire speaker groups, the procedures associated with this general model were applied to multiple speakers and to a test lexicon (see section 4.2). The overall performance is computed on the basis of average approximations derived from these values.

3. Computation of Phonetic Distance

3.1 Minimum Edit Distance

In recent years, the computation of phonetic distance on the level of phonetic transcription has yielded interesting and encouraging results in various subfields of linguistics, such as dialectology (Kessler, 1995; Heeringa, 2004), diachronic linguistics (Covington, 1998), diagnosis of articulation problems (Somers, 1999; Connolly, 1997), cross-language information retrieval (Oakes & Banerjee, 2003), but also for practical applications such as the automatic detection of confusable names in large corpora (Kondrak & Dorr, 2004).

In the present contribution, we apply the approaches and some of the specific techniques described in this research to transcriptions of non-native speech. We employed the *minimum edit distance* (also known as *Levenshtein distance*) as the base algorithm to compute the distance of pairs of sequences. Edit Distance is a dynamic programming algorithm that computes the distance between two symbol sequences A and B (not necessarily words) by identifying the minimum number of character substitutions (SUBS), insertions (INS) an deletions (DEL) required to transform string A into string B. For example, for the word pair *hunger* and *fingers*, we obtain a distance value of 3, since two substitutions and one insertion are needed at minimum for a transformation of *hunger* into *fingers* (substitutions $h \rightarrow f$, $u \rightarrow i$ and the insertion of final *s*).

The minimum edit distance has the advantage that the compared strings do not need to be of the same length (as opposed to metrics like Hamming distance and Manhattan distance, which require identical string lengths). Any differences in string lengths can be expressed in terms of an arbitrary number of character insertions or deletions, resulting in greater distance values. Given that non-native pronunciation variants are often characterised by insertions or deletions of speech sounds, this is an essential requirement for string comparisons in this domain.

The basic edit distance algorithm is reasonably well documented in the literature (for overviews, see e.g. chapter 5 of Heeringa, 2004; Kruskal, 1983; Nerbonne et al., 1996) and will therefore not be outlined in detail here. It should be noted, however, that there are varieties and different implementations of the algorithm that may result in different distance values. Among other things, these differences pertain to the penalties assigned to the types of edit operations. A variety of the algorithm assigns a penalty of 2 for SUBS, but only 1 to both INS and DEL. This is justified by the fact that a SUBS can be described as a sequence of INS and DEL (see also Kruskal, 1983). Using this variant, the distance for *hunger–fingers* would be 5 instead of 3. After experimenting with different alternatives, we applied an equal base penalty of 1 for all three operations SUBS, DEL, and INS.

3.2 Phonetic weights

The basic edit distance algorithm that was sketched so far has a major disadvantage for the present application: If the penalty values are constant for all segments that undergo the edit operations, potential similarity relations between segments are not taken into account. In the case of phonetic strings, this not an optimal solution, since it

should make a difference whether a segment is substituted by a phonetically similar segment (e.g. a replacement of [e] by [ɛ]) or by a sound of a completely different quality (e.g. replacing [e] by [p]). These relations should be incorporated in the overall distance metric in such a way that minor phonetic alternations in individual segments will yield lower distance values than significant phoneme shifts.

Therefore, we introduced a *segmental weight factor* to adapt the substitution penalties according the similarity of individual phonetic segments. In this way, lower distance values will be obtained if a non-native pronunciation variant is characterised only by moderate segmental deviations from the canonical form, whereas a heavily accented form will generally yield a higher distance value due to marked phoneme shifts. Different implementations of phonetic weight computation were applied in previous research, e.g. by Kondrak (2003) Heeringa (2004), Kessler (1995), Somers (1999), and Connolly (1997). We determined the similarity of phonetic segments on the basis of binary phonetic features: Each phoneme is represented as a set of features such as [obstruent], [low], [tense], each of which is specified either positively [+] or negatively [-]. Given this representation, the similarity of two phonemes can be computed by the number of shared features relative to the total number of features defined. We used a set of 25 phonetic features for this task.

3.2.1 Summary of the Distance Metric, Limitations

To sum up, the phonetic distance metric employed for the evaluation is so far made up of two components (although listed in a sequential order for convenience, their actual application is closely integrated in practice):

(1) *Edit Distance* to compute the minimum number of substitutions, insertions and deletions needed to transform string A into string B (basic sequence distance).

(2) *Phonetic segment similarity* to assign a weight factor to substitution operations in order to account for the similarity of individual phonetic segments.

The values obtained in this way were normalised to a range from 0 to 1. This extra step is required in order to obtain values that are independent of the absolute number of edit operations. Otherwise, it would not be possible to compare distances between word pairs of different lengths, since longer string pairs may generally contain more edit operations than shorter ones. Finally, the value was transformed into a similarity value[1], where 1 denotes the maximum similarity (= identity) of two strings, and 0 is the maximum distance (a value which is never reached in practice).

The distance/similarity metric employed for the evaluation is still limited in various respects. For instance, the issue of *feature prominence* has not been fully taken into account. This additional refinement allows for the fact that some features such as [vocalic] should be rated higher than features like [aspirated]) since they include a major shift of phoneme classes, whereas the latter represents only a minor phonetic detail. This differentiation can be incorporated by assigning another weight factor to each feature at the point where segment similarity is computed (see e.g. Kondrak, 2003). Although a basic feature prominence function that distinguishes major phoneme classes from other features is implemented in our algorithm, it has not yet been exploited by using an elaborate system of feature weights.

Another potential enhancement of the metric is *context-sensitivity* of edit operations: For instance, a deletion of [p] in a [pf] cluster is likely to result in a less prominent sound change than the deletion of [p] in an intervocalic position like [epa]). Therefore, different weights should be assigned to insertion and deletion operations according to the quality of the adjacent phonemes (direct neighbours).

4. Test Design

4.1 Speakers

The overall evaluation was conducted for six language directions and the corresponding rule sets. Due to limited space, this paper will only present results for the two language directions L1 Italian/L2 German and L1 Italian/L2 English. However, some important general tendencies to be observed in this data hold for other language directions as well. The selected language directions are particularly suited for illustration purposes since the group of native Italian speakers in our data collection is quite heterogeneous with respect to speakers' age and L2 proficiency levels. As the cross-speaker coverage of the rules is a main focus of interest of the evaluation, this was viewed as a benefit.

The reference data (i.e. speech and phonetic transcriptions) includes pronunciations of 43 city names and 10 short sentences of each the target languages English and German by a total number of 16 native speakers of Italian. The pronunciations of the city names are available in two different production modes. In the first mode, the speakers read the names as isolated words from prompt sheets. In the second mode, the speakers were presented with the correct L2 pronunciations as acoustic prompts and had to repeat it. The two production modes were treated separately in the evaluation since there are marked pronunciation differences between the reading and repeating modes that may have a considerable effect on the results.

4.2 Test Procedure

A computation of phonetic distance outlined above forms the basis of the evaluation. This algorithm was integrated in a standard test procedure in which we computed the phonetic distances between the empirical pronunciations produced by non-native speakers and (a) the canonical transcriptions and (b) each of the four rule-based variants. A central question to be answered by this procedure is whether or not the rule-based variants achieve a higher degree of approximation to empirical pronunciations than the canonical L2 transcription on a cross-speaker basis. With four accent levels generated by the rules plus the canonical form, there are five distance values altogether. The lowest distance among these values will qualify as the *best approximation*.

Since we were ultimately interested in the performance of the rules with respect to entire speaker groups and a particular vocabulary (rather than individual speakers and lexicon items), this computation was integrated into a procedure that identifies the average phonetic distances for (i) a specific speaker group and (ii) a test vocabulary of *N* entries. More precisely, we wanted to identify the relative

[1] The transformation is: *similarity (a, b) = 1 – distance (a, b)*

number of best approximations achieved by each of the five transcription variants. Formally, this test procedure can be written as follows (in pseudo-code):

$W_1 .. W_i$ lexicon entries, $W \in$ *test lexicon*
$S_1 .. S_n$ speakers, $S \in$ *speaker group*
$V_0(W)$ canonical transcription of W
$V_1 .. V_4(W)$ rule-based variants for W

for each $W \in$ *test lexicon* **do**
  for each $S \in$ *speaker group* **do**
    for $V_0 .. V_4(W)$ **do**
      D := COMPUTE_DISTANCE ($S(W), V_i(W)$)
    DMin := GETMINIMUM ($D(V_1) .. D(V_4)$)

  if $DMin < D(S_i, V_0)$ **then**
    rule-based ++
  else
    canonical ++
return [*rule-based*, *canonical*]

The values *rule-based* and *canonical* accumulated in this procedure indicate the absolute number of instances the rule-based and canonical variants achieved the best approximation. We can further distinguish the proportions achieved by each of four rule-based variants in order to obtain a more detailed picture of the performance of individual accent levels. The values computed in this procedure (transformed into relative frequencies %) are the central figures in the discussion of results in the next section.

5. Results

For the purpose of evaluation, the speakers were split up into two groups. Group A is a group of six Italian exchange students (aged 20-30) who lived in German at the time of the recordings. Their L2 proficiency levels for German and English are relatively high (see below). In contrast, Group B includes 10 native Italians recorded in their home country. All speakers in the latter group are aged 40-65 yrs. and achieve lower proficiency levels for both L2s English and German.

Proficiency levels were rated by integers on a scale ranging from 0 (no L2 knowledge at all) to 5 (native speaker). Thus, a non-native speaker with excellent L2 proficiency could achieve a maximum rating of 4.

Group / **L2**	**A** **(6 speakers)**	**B** **(10 speakers)**
English	2.7	1.7
German	2.9	0.7

Table 1: Average proficiency levels in groups A and B

Note that although the rule-sets that generate the accent pronunciations were mostly based on empirical data from the same speech data collection, there is no overlap between the 'training set' used for the rule creation and the evaluation test set. This is an important requirement with respect to the question whether the rules are capable of generating plausible variants for previously unseen speakers.

The total number of spoken words (tokens) covered in the evaluation is approx. 6.200 for group A and approx. 10.000 for group B.

5.1 Speaker Group A

The following tables 2 and 3 show the results for speaker group A for both target languages English and German respectively. The results are listed separately for three production modes and prompt types: (1) city names read from a prompt sheet, (2) city names elicited by acoustic prompts in which the correct L2 pronunciation was presented, and (3) sentences read from a prompt sheet. The values represent the relative number (%) of best approximations to the speakers' pronunciations for each transcription variant (canonical, rule-based), as detailed above in 4.1. For the rule-based variants, the proportion of each accent level is provided. The sum of the proportions of all accent levels is given as an indication of the overall coverage of the rule-based variants.

Task / **Transcription**	***names read***	***names repeated***	***sentences read***
Canonical	**29.8%**	**50.0%**	**55.9%**
Rule-based			
Level 1	33.3%	31.0%	31.9%
Level 2	22.9%	8.1%	8.2%
Level 3	8.9%	9.3%	1.8%
Level 4	5.0%	1.6%	2.1%
All levels 1-4	**70.2%**	**50.0%**	**44.1%**

Table 2: Results for speaker group A (6 speakers). L1 Italian/L2 English

Task / **Transcription**	***names read***	***names repeated***	***sentences read***
Canonical	**29.4%**	**41.2%**	**57.6%**
Rule-based			
Level 1	19.8%	28.4%	13.7%
Level 2	25.2%	17.9%	18.0%
Level 3	22.5%	11.3%	9.7%
Level 4	3.1%	1.2%	1.0%
All levels 1-4	**70.6%**	**58.8%**	**42.4%**

Table 3: Results for speaker group A (6 speakers). L1 Italian/L2 German

For both language directions, significant differences can be observed in the results for each of the three types of speech (a) city names read, (b) city names repeated and (c) sentences read. The same tendency occurs in all other language directions not presented here. This is caused by several factors: The notable difference between the *city names read* vs. *repeated* tasks reflects the fact that the speakers' approximation of the correct L2 form is generally much better if the pronunciation was elicited by an acoustically presented reference form. Read speech, in contrast, often includes additional pronunciation errors (such as spelling pronunciation errors). A second general tendency is the relatively low performance of the rule-based variants for the sentences. Two factors are involved here: First, the rules were explicitly optimised for names rather than for 'regular' lexicalised vocabulary. Therefore,

the relatively poor results outside the domain of names are not surprising. Secondly, the pronunciation of lexicalised L2 vocabulary is often closer to the correct L2 norm than the pronunciation of foreign names. For these reasons, the results for the canonical transcriptions are usually better for both repeated names and sentences.

Looking at the performance of the rule-based variants, it can be observed that there is generally no single variant among the four accent levels that achieves a significantly higher approximation than the canonical forms. Particularly for the L2 German, the number of best approximations is evenly dispersed over three accent levels, with a general tendency of decreasing approximations for higher accent levels. This bias towards lower accent levels, as well as the fact that the highest accent level 4 is virtually not represented in the results, reflects the speaker distribution in group B reasonably well (as stated above, L2 proficiency levels in group A are relatively high). Since accent level 4 models a heavily accented pronunciation that strongly deviates from the canonical L2 form, this result is in line with the model of accent levels. Still we can observe the general tendency for this speaker group that none of the single rule-based variants obtains a higher approximation than the canonical form (one exception being the read city names task for the L2 English).

5.2 Speaker Group B

Since group B includes speakers with rather low L2 proficiency levels, we could ideally expect an increased proportion of best approximations among the rule-based variants, particularly for the higher accent levels. However, this hypothesis was only partially confirmed. The following table shows the results for group B with the target language English:

Task / Transcription	*names read*	*names repeated*	*sentences read*
Canonical	**29.5%**	**45.8%**	**54.8%**
Rule-based			
Level 1	27.4%	27.9%	31.6%
Level 2	23.5%	7.2%	6.8%
Level 3	11.4%	15.8%	4.6%
Level 4	8.1%	3.3%	2.2%
All levels 1-4	**70.5%**	**54.2%**	**45.2%**

Table 4: Results speaker group B (10 speakers). L1 Italian/L2 English

Although the average proficiency level of 1.7 is notably lower than in group A (avrg. level = 2.7), the results and general tendencies are very similar to the first speaker group. Again, there is a proportion of approx. 30% of best approximations for the canonical forms, as well as a bias towards lower accent levels among the rule-based variants.

For the L2 German, in contrast, the hypothesis that the proportion of higher accent levels will increase is well supported, as shown in the following table:

Task / Transcription	*names read*	*names repeated*	*sentences read*
Canonical	**11.7%**	**27.3%**	**40.3%**
Rule-based			
Level 1	13.3%	20.3%	7.9%
Level 2	33.3%	31.7%	27.7%
Level 3	29.8%	17.7%	21.6%
Level 4	11.9%	3.0%	2.5%
All levels 1-4	**88.3%**	**72.7%**	**59.7%**

Table 5: Results speaker group B (10 speakers). L1 Italian/L2 German

Here, the canonical transcriptions achieve the best approximation for merely 12% of all utterances, while at the same time there is a significant shift towards the rule-based variants, especially at higher accent levels. With a total proportion of about 60% for the city names reading task, the levels 2 and 3 approximate the pronunciation variants in this speaker group particularly well.

The most plausible explanation of the discrepancy between the results for the L2s English and German is a difference in the proficiency levels for these target languages within speaker group B. Whereas the average proficiency level for English is 1.7, it is as low as 0.7 for the L2 German.

With respect to the plausibility of the rule-based variants as well as the implementation of accent levels, we can draw a preliminary conclusion from these results. While the model of accent levels seems generally well suited to cover a variety of potential pronunciation variants within a group of non-native speakers, its implementation needs refinement. More precisely, it requires a better representation of speakers with high proficiency levels (weak accent), while at the same time, the topmost accent level 4 seems to be dispensable in most cases, since it rarely achieves a good approximation. The latter finding, however, was not unexpected: Level 4 has been added to an earlier model of only 3 levels in order to generate pronunciations in which speakers apply their L1 pronunciation rules to the target language almost without any modifications (as in the previously mentioned German [bɪʁmɪŋham] for English *Birmingham*). However, this rarely happens in practice.

6. Summary and Discussion

It is probably inappropriate to interpret the results in a one-dimensional way, stating that either the rule-based or the canonical transcriptions outperform the other. It will ultimately depend on factors such as speaker groups, type of vocabulary, and type of speech whether the rule-based transcriptions achieve a better match to non-native pronunciations. Nonetheless it is possible to identify some general tendencies:

- For speakers with relatively high L2 proficiency levels, only the rules for lower accent levels 1 and 2 achieve a good approximation. However, the proportion of best approximations among these rule-based variants is often similar to the proportion achieved of the canonical transcriptions, so the performance gain is rather low.

- For speakers with very low L2 proficiency levels, the approximation of the canonical transcriptions drops significantly in favour of the rule-based variants. At the same time, the proportion of best approximations among of the higher accent levels increases.
- As an additional indication of the degree of approximation of individual transcriptions, we also counted the number of perfect matches (complete identity of transcriptions) in the evaluation procedure. There is a significant advantage of the rule-based variants with respect to this value. On average, their proportion of perfect matches is 2 to 3 times higher than for the canonical transcriptions.
- In most cases there is no single rule-based variant that outperforms the canonical transcriptions with respect to their approximation to actual non-native pronunciations. However, if we accumulate the values for all accent levels to one total value, the rule-based variants achieve a much higher approximation.

The last observation is in line with the model of accent levels that underlies the rule-based generation technique. Since it seems inappropriate and insufficient to model non-native pronunciations by adding only one single accented pronunciation variant to each lexicon item, the model was introduced as an attempt to split up the spectrum of potential pronunciation variants within a non-native speaker group into discrete, prototypical variants. Accordingly, it was not anticipated to find the best approximations to non-native speech data in only one variant. More generally spoken, the dispersion of best approximations over multiple accent levels should by no means be viewed as a shortfall of the rule-based technique; rather, it is one of its central features, reflecting the broad scope of inter-individual variation to be expected in non-native speech.

For further-reaching interpretations of the results, however, we need to return to application-oriented questions discussed in the beginning of this paper. Whether or not the model of multiple rule-based pronunciations may be exploited for system improvements will ultimately depend on the type of application in which the rules can be integrated. For example, the phonetic lexicons of speech recognizers may in principle profit from alternative pronunciations (provided that the transcriptions add pronunciation variation which is not already sufficiently covered by the acoustic models). However, too many alternative pronunciations can increase the confusability of lexicon entries in the recognition process, so that positive effects may be neutralised. Another potential application of the rules is speech synthesis. If the rules are being employed for the generation of 'foreign-accented' speech output, the requirements are clearly different than for ASR. A system designer can profit from the accent levels by choosing in advance the desired accent strength to be generated. However, in contrast to ASR, there will rarely be a need to apply all accent levels simultaneously.

Examples like these show that the criteria for judging the suitability of the rules are always closely intertwined with the potential application context. The evaluation described in this paper cannot predict the merit of the rule-based pronunciations in specific applications. It can only indicate the degree to which the modified transcriptions match empirical linguistic data.

Acknowledgements

This study was funded by the *Deutsche Forschungsgemeinschaft* (DFG) within a research project carried out at the Institute of Communication Acoustics, University of Bochum, Germany. The author would like to thank all former colleagues for their support, particularly Prof. Ute Jekosch (now at Technical University of Dresden).

7. References

Connolly, J. H. (1997): "Quantifying target-realization differences". *Clinical Linguistics & Phonetics 11*, 267-298.

Covington, M. A. (1998): "Alignment of multiple languages for historical comparison". *Proceedings 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, 275-280.

Heeringa, W. J. (2004): *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Ph.D. Thesis. Groningen: University Library Groningen.

Kessler, B. (1995): "Computational dialectology in Irish Gaelic". *Proceedings 7th Conference of the European Chapter of the Association of Computational Linguistics (EACL 95*), Dublin, 60-66.

Kondrak, G. (2003): "Phonetic alignment and similarity". *Computers and the Humanities* 37**,** 273–291.

Kondrak, G.; Dorr, B. J. (2004): "Identification of confusable drug names: A new approach and evaluation methodology". *Proceedings 20th International Conference on Computational Linguistics (COLING 2004).* Geneva, Switzerland, 952-958.

Kruskal, J. (1983): "An overview of sequence comparison". In: Sankoff, D. & Kruskal, J. (eds.): *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading/Mass.: Addison-Wesley, 1–44.

Nerbonne, J.; Heeringa, W.; Hout, E. van den; Kooi, P. van der; Otten, S.; Vis, W. van de (1996): "Phonetic distance between Dutch dialects". In: Durieux, G., Daelemans, W., & Gillis, S. (eds.) *CLIN VI, Papers from the sixth CLIN Meeting,* University of Antwerp, 185–202.

Oakes, M. P.; Banerjee, S. (2003): "Regular sound changes for cross-language information retrieval". *Working Notes for CLEF 2003 Workshop.* Trondheim, Norway, 137-142.

Schaden, S. (2003): "Generating non-native pronunciation lexicons by phonological rules". *Proceedings 15th International Conference of Phonetic Sciences (ICPhS 2003)*, Barcelona, 2545–2548.

Schaden, S.; Jekosch, U. (2006): "Casselberveetovallarga and other unpronounceable places: The CrossTowns Corpus". *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy.

Somers, H. (1999): "Aligning phonetic segments for children's articulation assessment". *Computational Linguistics* 25 (2), 267-275.

van Bael, C.; Binnenpoorte, D.; Strik, H.; van den Heuvel, H. (2003): "Validation of phonetic transcriptions based on recognition performance". *Proceedings Eurospeech 2003*, Geneva, 1545-1548.