A DIALECT DISTANCE METRIC BASED ON STRING AND TEMPORAL ALIGNMENT

Thomas Kisler and Uwe D. Reichel

Institute of Phonetics and Speech Processing, University of Munich
{kisler, reichelu}@phonetik.uni-muenchen.de

Abstract: The Levenshtein distance is an established metric to represent phonological distances between dialects. So far, this metric has usually been applied on manually transcribed word lists. In this study we introduce several extensions of the Levenshtein distance by incorporating probabilistic edit costs as well as temporal alignment costs. We tested all variants for compliance with the axioms that within-dialect utterance pairs are phonologically more similar than across-dialect ones. In contrast to former studies we are not applying the metrics on preselected, prototypical word lists but on real connected speech data which was automatically segmented and labeled. It turned out, that the transcription edit distances already performed well in reflecting the difference between within- and across-dialect comparisons, and that the adding of a temporal component rather weakens the performance of the metrics.

1 Introduction

Comparing dialects manually is, unfortunately, very time consuming and relies solely on the performance of the human expert. Therefore, automatic methods have been proposed to not only increase the speed of the analysis, but, furthermore, to have an objective metric that produces transparent and reproducible results.

The Levenshtein distance, i.e. the minimum number of edit operations to convert one transcription into another, is a widely used metric for dialect comparison [1]. Kessler [2] applies the Levenshtein distance to compare the distance between different Gaelic dialects, with material from sites in Ireland, Scotland and the Isle of Man, with satisfactory results in clustering those Gaelic dialects in their natural classes. He compares different metrics based on phone string comparison and feature string comparison, but finds that the approach based on phone strings has the best performance.

Nerbonne and Heeringa [3] apply the word length normalized Levenshtein distance on Dutch dialects, where they use feature vectors to be able to express the closeness of two phones, like "t" and "d", where the naive Levenshtein distance would still result in the same distance as for "a" and "d". To account for the fact that in normalization for the average word length, missing words would be penalized in an extensive way, they use the longer word for the normalization of the distance.

Heeringa, Kleiweg, Gooskens, et al [4], furthermore, analyze the effects of n-grams on the distance measure using both examples from Norwegian and German. They find that the *n*-grams only mildly improve the result. Furthermore, they confirm that the methods based on phones instead of feature vectors lead to better results.

The work just mentioned produces good results on the underlying data. As the distances are measured on a small sample set of words from a word list and uttered by mostly only a few prototypical speakers of a certain dialect area (exception is [4]), they might be biased by a)

the manual transcription, b) the speaker themselves or c) the selection of words. We, therefore, wanted to explore the power of the metrics on automatically segmented and labeled (S&L) data on a big set of speakers with connected speech, rather on a comparatively small word lists, where the subjects do not necessarily speak a strong dialect. Additionally those methods work only on the transcriptions. Through the availability of the speech recordings we wanted to incorporate the available temporal aspect. Therefore we propose an addition to the basic Levenshtein-distance based metrics in adding a temporal component. To account for phonological similarity of phones we don't choose expert-driven, theory-depended feature vectors, but a probabilistic cost function reflecting co-occurrence frequencies of phone pairs.

2 Data

For the evaluation of the proposed metric, the corpus of Regional Variants of German (RVG) was used [5]. The RVG corpus was recorded between 1995 and 2009 at the Institute of Phonetics and Speech Processing at the University of Munich and contains recordings of over 400 speakers from nine dialectal regions. We grouped this nine regions into four bigger dialectal groups. Normally, Germany is divided in three different supergroups which are Low German (Niederdeutsch), Central German (Mitteldeutsch) and Upper German (Oberdeutsch). Through the transition area in the West Central German (Westmitteldeutsch) area, called "rheinischer Fächer" and the homogeneous area of the East Low German (Ostmitteldeutsch) area [6], we decided to not group the Central German dialects together and use four groups (group names and abbreviations after the braces).

- Low Franconian (Niederfränkisch)
- West Low German (Westniederdeutsch)
- East Low German (Ostniederdeutsch)

} Low German (Niederdeutsch): *LG*

- West Central German (Westmitteldeutsch) } *WCG*
- East Central German (Ostmitteldeutsch) } *ECG*
- Alemannic (Alemannisch)
- Bavarian (Bayerisch-Österreichisch)
- East Franconian (Ostfränkisch)
- South Franconian (Südfränkisch)

} Upper German (Oberdeutsch): *UG*

From the corpus we selected the read speech part, that comprises a set of sentences read by multiple speakers. To rule out those speakers that are unlikely to have strong regional variation, we did not take people into the analysis that classified themselves as speakers of the standard German variety ("Hochdeutsch").

3 Method

3.1 Utterance distance metrics

The speakers were compared pairwise for each input text. Five distances $d_*(v,w,z)$ of the phonetic segmentations v and w of two speakers, who had to read the same text z have been calculated. Thereby, each phonetic segment v_i (analogously w_j) is defined by a label v_i^l and a time interval v_i^t:

- $v_i =< v_i^l, v_i^t >$,
- $w_j =< w_j^l, w_j^t >$.

All distances are based on transcription and/or temporal alignment, which is solved by dynamic programming [7], minimizing the overall alignment costs. As described in more detail in the next sections, the distances d_n and d_p are based on transcription alignment using a naive and a probabilistic cost function respectively. d_t is based on temporal alignment, and d_{nt} and d_{pt} combine both alignment types.

3.1.1 Naive edit distance d_n

The distance d_n is based on the cost function c_n that simply assigns cost 1 to each substitution, insertion, and deletion. The costs are thus defined as:

- **Substitution:** $c_n(v_i^l, w_j^l) = \begin{cases} 0 & : \text{ equal}(v_i^l, w_j^l) \\ 1 & : \text{ else.} \end{cases}$
- **Deletion:** $c_n(v_i^l, \_) = 1$
- **Insertion:** $c_n(\_, w_j^l) = 1$

To make d_n independent of utterance length, we normalized it to the mean length of v and w.

3.1.2 Probabilistic edit distance d_p

Following the PermA approach in [8] the probabilistic edit costs c_p are defined in terms of conditional probabilities reflecting symbol co-occurrences in the transcription data of our RVG sub corpus. The costs are defined as follows:

- **Substitution:** $c_p(v_i^l, w_j^l) = \begin{cases} 0 & : \text{ equal}(v_i^l, w_j^l) \\ 1 - P(w_j^l | v_i^l) & : \text{ else.} \end{cases}$
- **Deletion:** $c_p(v_i^l, \_) = 1 - P(\_ | v_i^l)$
- **Insertion:** $c_p(\_, w_j^l) = 1 - P(w_j^l | \_)$

Smoothing of the conditional probabilities is achieved by weighting the co-occurrence count increments within a triangular window of length 3 and size 1.

To avoid unsatisfying heuristics in cost assignment (see [9] for examples), insertion and deletion operations are basically considered as substitutions involving the _-symbol. Respective co-occurrences are counted by _-symbol padding and permutation in the shorter sequences in the alignment training pairs. See [8] for details.

To fulfill the symmetry requirement of a distance metric, d_p was calculated as the mean of $d_p(v,w)$ and $d_p(w,v)$ each normalized with respect to the length of v and w respectively.

3.1.3 Temporal distance d_t

Following [10] temporal costs are defined in terms of the *overlap ratio* of the segments' time intervals v_i^t and w_j^t:

$$c_t(v_i^t, w_j^t) \quad = \quad 1 - \frac{t_{ij}}{v_i^t + w_j^t - t_{ij}}$$

The overlap ratio is visualized in Figure 1. t_{ij} is the duration of the overlap of segments v_i and w_j. The denominator gives the total duration from the start of the first occurring segment to the

end of the second one. Thus the overlap ratio is 0 in case of no overlap yielding cost 1. In case of perfect overlap the ratio is 1 yielding cost 0. In our approach deletions and insertions are treated as calculating the overlap with segments of length 0, both resulting in an overlap ratio 0 (since $t_{ij} = 0$) and, therefore, cost 1.

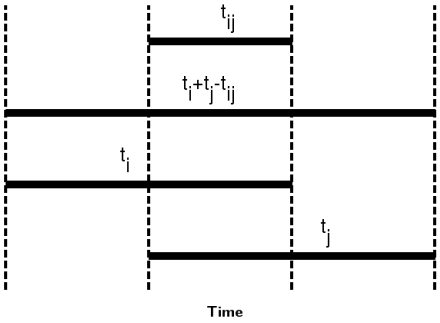

Figure 1 - Overlap ratio of time intervals t_i and t_j: $\frac{t_{ij}}{t_i+t_j-t_{ij}}$.

To remove the influence of speech pauses and utterance length on d_t, the temporal distance was calculated separately for each pair of corresponding words each time setting the temporal onset of both words to 0. As for the previous distance metrics word length normalization was carried out. Finally, the word-based normalized distances were averaged over the whole utterance.

3.1.4 Complex distances d_{nt} and d_{pt}

The distances d_{nt} and d_{pt} are based on the mean of the respective edit and the temporal cost component, the naive edit costs for d_{nt} and the probabilistic costs for d_{pt}.

$$
\begin{aligned}
c_{nt}(v_i, w_j) &= \frac{c_n(v_i^l, w_j^l) + c_t(v_i^t, w_j^t)}{2} \\
c_{pt}(v_i, w_j) &= \frac{c_p(v_i^l, w_j^l) + c_t(v_i^t, w_j^t)}{2}
\end{aligned}
$$

3.2 Dialect comparisons

By the alignment variants described above five distance values per speaker pair and text stimulus were derived.

For the following axioms we need two definitions:

- Within-dialect distances WD_x are all distances of speakers belonging to one dialect x, so $WD_x = \{d(x,x)\}$
- Across-dialect distances AD_x are all distances of pairs where one speaker belongs to dialect x and one speaker to dialect y, so $AD_x\{\forall y \neq x : d(x,y)\}$

We formulated the following three axioms in order to test whether or not the metrics can be used for dialect comparison:

Ax1 $WD_{all} < AD_{all}$

Ax2 $\forall x : WD_x < AD_x$ (x belonging to the nine dialects listed in section 2)

Ax3 $\forall x : WD_x < AD_x$ (x belonging to the four dialect groups listed in section 2)

For **Ax1**, we evaluated the appropriateness of our measures by comparing the overall within-dialect distances WD_{all} with the across-dialect distances AD_{all} of all dialects combined. The trivial expectation that within-dialect distances are smaller than across-dialect distances should be reflected by our distance metrics.

For testing the compliance to **Ax2** we used the available assignment to dialects. As mentioned before, there are nine dialects, so we made an overall of nine comparisons between the dialects, where for each dialect the WD_x were tested against the AD_x of a certain dialect. Again WD_x should be significantly smaller than AD_x.

For testing the compliance to **Ax3** we grouped the dialects together to bigger groups as mentioned in 2 and repeated the evaluation from **Ax2**.

4 Results

4.1 Overall within- vs. across-dialect distance

The test against **Ax1**, which is the intuitive understanding that within-dialect distances are smaller than across-dialect distances, we used a one-tailed Welch's t-test, which resulted in a highly significant distance difference in the expected direction for all 5 metrics ($t(\infty) > 8.5, p < 0.01$).

4.2 Per dialect within- vs. across-dialect distances

The per dialect within- vs. across distance comparison for all five metrics resulted in the pattern seen in table 1, where not only the significant differences are shown, but as well the tendencies towards $WD_x < AD_x$ or $WD_x > AD_x$. Especially the later case is an undesirable state and means that based on the available data the metric is not able to reflect **Ax2**.

Table 1 - Overview of the number of cases where the criteria for **Ax2** is met or not met that WD_x (within-dialect distances) is smaller than AD_x (across dialect distances).

	d_n (naive)	d_{nt} (naive + temporal)	d_p (prob)	d_{pt} (prob + temporal)	d_t (temporal)
Significant $WD_x < AD_x$	5	3	5	3	4
Insignificant $WD_x < AD_x$	1	2	1	2	2
Insignificant $WD_x > AD_x$	3	4	3	4	3

4.3 Per dialect group within- vs. across-group distances

The per dialect group pairwise within-group to across-group comparison resulted in highly significant results for all five metrics and all groups ($p < 0.01$), based on a one-tailed Welch's t-test, except for the comparison for *LG* with *WCG* with d_t (temporal costs). There the result is still weakly significant ($p < 0.1$).

5 Discussion

5.1 Overall within- vs. across-dialect distances

The within-dialect distances are significantly smaller than the across-dialect distances, which shows that all cost functions produce results that go along with the intuitive understanding that the difference of speech for speakers within a certain dialect are smaller, than for speakers across dialects. This compliance to **Ax1** we take as a first indicator for the examined distances to be appropriate metrics for dialect comparison.

5.2 Per dialect within- vs. across-dialect distances

The per dialect within- vs. across-dialect distance comparison revealed, that the intuitive understanding that WD_x is smaller than AD_x does not hold true for all dialects. The metrics d_n (naive Levenshtein) together with d_p (distance based on conditional probabilities) produce the best results. Adding the temporal component c_t as described above does not only not improve the results, but makes them worse. So for the temporal variant three cases are left where WD_x is smaller than AD_x. The metric d_t based on temporal costs alone works again better when not combined with other cost functions. That the naive binary Levenshtein distance on phones works better than more complicated cost functions, is in line with findings of [2, 4].
The possible problems with the data in general and the automated segmentation and labeling, are discussed in greater detail in 5.4.

5.3 Per dialect group within- vs. across-group distances

In contrary to the original dialect assignment the per dialect group comparison of the within- vs. across-distances shows the expected behavior, that WD_x of a group is smaller, than AD_x. The fact that it does work for bigger groups, but not with the initial assignment, causes us to assume there might be some dialectal regions which are more diverse than others or maybe the effects have been to small to be significant in the evaluation of the initial dialect assignment.

5.4 Problems and further work

The partial violation of **Ax2** might have four reasons.
First, the used data is not guaranteed to be very dialectally influenced, even though people from many different regions were recorded and we did not regard the subjects indicating their own dialect as standard German. It might be possible to further sort out people with only weak regional influence, but such a selection is not trivial. When done by a single expert, the data might be biased towards his opinions. So a set of experts or untrained personal might be necessary.
Second, we needed to use read speech data in order to be able to align utterance pairs. Read speech is very likely to be less influenced by regional variation, especially as the speakers where not asked to produce very dialectal speech, which eliminated more dialectal differences. Again, this could be solved by using data from speakers which are perceived as strong dialect speakers.
Third, unlike in the data used in previous studies, which used words that are supposed to have strong local variations, the text that has been read in the RVG corpus constitutes connected speech and, therefore, a large amount of words not differing between dialects. On the other hand, the inclusion of all and not just differing words should give a more realistic picture, since the focus is not only on differences between dialects, but also on similar dialect characteristics.
Fourth, we evaluated the metrics against data that was automatically segmented and labeled. The statistical model of MAUS [11], the tool used for the automatic S&L, for German is trained

on standard German speakers and biases the result by eliminating rare, but dialect-related phone sequences. An example for this is the word "fährt" (drives) in a sentence read both from a speaker of West Low German and one of Bavarian. The differences are easy to spot between West Low German "fE:6t" and Bavarian "fO6d" (expert transcriptions in German SAM-PA[1]). MAUS though labels both utterances with the "E:" (instead of "O" in the Bavarian utterance) and, therefore, eliminates the dialectal differences at this position. The phrase final "t" and "d", are detected correctly by MAUS. This problem could be solved by using manually segmented and labeled data or might be reduced by turning off the statistical language model and let MAUS make the alignment solely based on the acoustic signal.

Another point is the unbalanced number of speakers per dialect. There are dialects with only 14 speakers (South Franconian) and dialects with over a hundred (West Low German and Bavarian-Austrian), which has an affect on within-dialect variability.

One particular short-coming of the presented temporal metrics itself is, that insertions and deletions in the overlap ratio are punished very hard, which is inappropriate to reflect e.g. frequent phonological processes as Schwa deletions. This might be an explanation for their low performance.

6 Acknowledgments

The work of the authors has been carried out within the CLARIN-D project [12] (BMBF-funded).

References

[1] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[2] B. Kessler, "Computational dialectology in Irish Gaelic," 1995. [Online]. Available: http://arxiv.org/abs/cmp-lg/9503002

[3] J. Nerbonne and W. Heeringa, "Measuring dialect distance phonetically," in *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, 1997, pp. 11–18.

[4] W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne, "Evaluation of string distance algorithms for dialectology," in *Proceedings of the Workshop on Linguistic Distances*, ser. LD '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 51–62.

[5] S. Burger and F. Schiel, "RVG 1 - A Database for Regional Variants of Contemporary German," in *Proc. of the 1st Int. Conf. on Language Resources and Evaluation*, Granada, Spain, 1998, pp. 1083–1087.

[6] S. Barbour and P. Stevenson, *Variation im Deutschen - Soziolinguistische Perspektiven*. de Gruyter, 1998.

[7] R. Wagner and M. Fischer, "The string to string correction problem," *Journal of the Association for Computing Machinery*, vol. 21, no. 1, 1974.

[1] Speech Assessment Methods Phonetic Alphabet

[8] U. Reichel, "Perma and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, Portland, Oregon, 2012, p. paper no. 346.

[9] G. Kondrak, "Algorithms for Language Reconstruction," Ph.D. dissertation, University of Toronto, 2002.

[10] S. Paulo and L. Oliveira, "Automatic phonetic alignment and its confidence measures," in *Proc. 4th International Conference EsTAL*, Alicante, Spain, 2004, pp. 36–45.

[11] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. ICPhS*, San Francisco, 1999, pp. 607–610.

[12] "http://eu.clarin-d.de/index.php/en/," CLARIN-D web page.