Inducing a measure of phonetic similarity from pronunciation variation

Martijn Wieling, Eliza Margaretha and John Nerbonne
University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands

Abstract

Structuralists famously observed that language is "un systême où tout se tient" (Meillet, 1903, p. 407), insisting that the system of relations of linguistic units was more important than their concrete content. This study attempts to derive content from relations, in particular phonetic (acoustic) content from the distribution of alternative pronunciations used in different geographical varieties. It proceeds from data documenting language variation, examining six dialect atlases each containing the phonetic transcriptions of the same sets of words at hundreds of different sites. We obtain the sound segment correspondences via an alignment procedure, and then apply an information-theoretic measure, pointwise mutual information, assigning smaller segment distances to sound segment pairs which correspond relatively frequently. We iterate alignment and information-theoretic distance assignment until both remain stable, and we evaluate the quality of the resulting phonetic distances by comparing them to acoustic vowel distances. Wieling et al. (in press) evaluated this method on the basis of Dutch and German dialect data, and here we provide more general support for the method by applying it to several other dialect datasets (i.e. Gabon Bantu,

U.S. English, Tuscan and Bulgarian). We find relatively strong significant correlations between the induced phonetic distances and the acoustic distances, illustrating the usefulness of the method in deriving valid phonetic distances from distributions of dialectal variation.
Keywords: phonetic distance, pointwise mutual information, acoustic vowel distance, confusion matrix, variation matrix

---

1. Introduction

As Laver (1994, p. 391) points out, there is no generally accepted procedure to determine phonetic similarity, nor even specific standards: "Issues of phonetic similarity, though underlying many of the key concepts in phonetics, are hence often left tacit."

It is clear that there has nonetheless been a great deal of work on related topics in phonetics and laboratory phonology. In phonetics, Almeida and Braun (1986) developed a measure of segment distance in order to gauge the fidelity of phonetic transcriptions. It was used, e.g., to evaluate intra- and intertranscriber differences. Cucchiarini (1993) refined this work and Heeringa (2004) also experimented with Almeida & Braun's segment distance measure in dialectometry.

In laboratory phonology, Pierrehumbert (1993) experimented with a simple feature-overlap definition of similarity to which Broe (1996) added an information-theoretic refinement discounting redundant features. Frisch (1996) recast these definitions in terms of natural classes, rather than features, and Frisch et al. (2004) demonstrate that the Arabic syllable is best described as involving a gradient constraint against similar consonants in initial and

final position, the so-called 'Obligatory Contour Principle'. Bailey and Hahn (2005) measure the degree to which Frisch's (1996) definitions predict the frequency of perceptual confusions in confusion matrices, obtaining fair levels of strength ($0.17 \leq r^2 \leq 0.42$).

In general, the work from phonetics and (laboratory) phonology has experimented with theoretically inspired definitions of similarity as a means of explaining phonotactic constraints or potential confusions. Bailey and Hahn (2005) contrasted theoretically inspired definitions of phonetic similarity to empirical measures based on confusion matrices. A confusion matrix (Miller and Nicely, 1955) normally records the outcome of a behavioral experiment. It is a square matrix in which the rows represent sounds (or symbols) presented to subjects and the columns the sounds perceived. Each cell (r, c) records the number of times the signal in row r was perceived as the signal in column c. So cell (ɔ,o) records how often [ɔ] was perceived as [o], and the diagonal then represents the non-confused, correctly perceived signals.

As opposed to confusion matrices which record variants in speech perception, we introduce VARIATION MATRICES which record (dialectal) variants in speech production. In our case the variation matrix is initiated not with a behavioral experiment, but rather using distributional data available in dialect atlases. Based on alignments of dialectal pronunciations for a large set of words, we obtain the frequency with which sound segments align. Continuing with the example above, cell (ɔ,o) in a variation matrix thus represents the number of times [ɔ] was used in the pronunciation of one variety, whereas [o] was used at the corresponding position in the pronunciation of another variety. We will use these variation matrices to directly extract information

about sound segment similarity in a data-driven manner (as opposed to proceeding from a theoretical notion, see above). Specifically, we employ the information-theoretic pointwise mutual information (PMI) measure of association strength to determine the final sound segment distances.[1] Studies involving (data similar to) confusion matrices have often applied MDS as well (Fox, 1983), just as we will here.

The automatically derived sound segment distances are evaluated by comparing them to independent acoustic characterizations. Since there is a consensus that formant frequencies characterize vowels quite well, we compare in particular the phonetic segment distances of vowels generated by our method to vowel distances in formant space. As we do not know how to measure acoustic differences between consonants, we cannot evaluate these, but we do examine them.

The PMI-based procedure we use to automatically derive phonetic segment distances was originally proposed by Wieling et al. (2009), who evaluated the induced distances by using them in an alignment procedure. The results were evaluated against a gold standard (Wieling et al., 2009) and were found superior to versions using a binary segment distance (0 or 1). A slightly improved version of the induction procedure (Wieling and Nerbonne, 2011b) played a supporting role in work aimed at removing the effects of inconsistent transcription practices. Wieling et al. (in press) applied the procedure to transcriptions from Dutch and German dialect atlases and compared the results to acoustic vowel distances. In addition to investigating whether these

[1] Ohala (1997) calls for an information-theoretic perspective on confusion matrices, but he is particularly interested in non-symmetric aspects of the matrices.

results generalize to other dialect datasets, we will provide a more elaborate discussion of the motivation and the consequences of the work.

Besides the phonetic perspective illustrated above, there are several other research areas for which improved sound segment distances are valuable. In dialectometry, obtaining pronunciation distances between different words (e.g., using the Levenshtein distance algorithm; see Section 3) is of central importance as these are used to compare dialectal pronunciations between different varieties (Nerbonne and Heeringa, 2009). Using sensitive sound segment distances, instead of the standard binary segment distance, will likely improve pronunciation distances between individual words. Improving segment distances also improves alignments (see Wieling et al., 2009) and will likely improve the ability in (automatically) identifying the sound correspondences which historical linguistics relies on (Hock and Joseph, 1996, Ch.4, 16).

Sequence alignment and sequence distance are central concepts in several areas of computer science (Sankoff and Kruskal, 1999; Gusfield, 1999), and the Levenshtein distance and its many descendants are used frequently, not only for phonetic transcriptions, but also for comparing computer files, macromolecules and even bird song (Tougaard and Eriksen, 2006). Kernighan et al. (1990) induced segment distances from teletype data in order to better predict the intended word when faced with a letter sequence that did not appear in their lexicon.

Phonetic similarity also plays a role when discussing the comprehensibility of foreigners' speech and how heavy their accents are (Piske et al., 2001; Flege et al., 1995), when assessing the success of foreign language instruc-

tion, or when discussing the quality of speech synthesizers (van Heuven and van Bezooijen, 1995). Sanders and Chin (2009) measure the intelligibility of the speech of cochlear implant bearers using a measure of phonetic similarity. Kondrak and Dorr (2006) apply a measure of pronunciation distance to identify potentially confusing drug names. And, although we will not attempt to make the argument in detail, we note that the many appeals to "natural" phonetic and phonological processes also seem to appeal to a notion of similarity, at least in the sense that the result of applying a natural process to a given sound is expected to sound somewhat like the original, albeit to varying degrees.

Finally, we note that it was a major structuralist tenet that linguistics should attend to the relations (distributions) among linguistic entities more than to their substance proper (Meillet, 1903, p. 407). For example, a structuralist attends more to phonemic distinctions, to sounds which fall in the relation "potentially capable of distinguishing lexical meaning" than to the details of how the sounds are pronounced, but also to sounds that fall in the complementary distribution relation (not found in the same phonetic environment) or the free variation relation (found in the same phonetic environment, but without an effect on lexical meaning).

In the present case we attend to sounds which participate in the relation "potentially used as a dialect variant" (across different speakers) and we do not privilege either phonemic or sub-phonemic variation. Some structuralists might well draw the line at considering variation outside a tightly defined variety, and in that sense we are perhaps not merely developing structuralist ideas. Other structuralists nonetheless recognized that the speech of "the

whole community" was the proper concern of linguistics, in spite of the fact that "every person uses speech forms in a unique way" (Bloomfield, 1933, p.75). They did not advocate attention to the idiolects of speakers in "completely homogeneous speech communities" (Chomsky, 1965, p.3).

In suggesting a renewed focus on phonetic and phonological relations, i.e. distributions, we are aware that phonetics — and to some extent phonology (Cole, 2010) — has largely and successfully ignored the advice to concentrate on relations, in favor of examining the articulatory, acoustic and auditory basis of sounds, and we do not presume to question the wisdom of that development. It nonetheless remains scientifically interesting to see how much information is present in (cross-speaker) distributions. As we note above, the sort of distribution we examine below is perhaps of a different sort than the ones many structuralists had in mind, but its key property is that it is derived from a large number of alternative pronunciations.

2. Material

2.1. Dialect pronunciations

In this study we derive phonetic segment distances for several datasets. In addition to the results on a Dutch and German dataset (reported by Wieling et al., in press), we also report results on four additional dialect datasets (i.e. U.S. English, Gabon Bantu, Bulgarian and Tuscan). In order to focus on segmental distances we ignore suprasegmentals, and in order to limit the number of distinct phonetic sounds in each dataset, we ignore diacritics. To obtain a reliable set of vowel distances, we also exclude vowels having a frequency lower than one percent of the maximum vowel frequency in each

dataset.

The Dutch dialect dataset was included in the study of Wieling et al. (in press) and contains phonetic transcriptions of 562 words in 613 locations in the Netherlands and Flanders. The words were selected by Wieling et al. (2007) from the Goeman-Taeldeman-Van-Reenen-Project (GTRP; Goeman and Taeldeman, 1996) in order to conduct an aggregate analysis of dialectal pronunciation variation in the Netherlands and Flanders. The Dutch dataset differentiates 18 vowels (excluding the low-frequency vowels): /a, ɑ, ɒ, ʌ, æ, e, ɛ, i, ɪ, y, o, ɔ, u, ʊ, ɵ, œ, ø, ə/.

The German dataset, also included in the study of Wieling et al. (in press), contains phonetic transcriptions of 201 words in 186 locations obtained from the *Phonetischer Atlas der Bundesrepublik Deutschland* (Göschel, 1992). Nerbonne and Siedle (2005) provide a detailed overview as well as a dialectometric analysis of this dataset. The German dataset differentiates 21 vowels (excluding the low-frequency vowels):/a, ɑ, ɒ, ʌ, ɐ, æ, e, ɛ, i, ɪ, y, ʏ, o, ɔ, u, ʊ, ɯ, ɵ, œ, ø, ə/.

The U.S. English dataset contains phonetic transcriptions of 153 concepts in 483 locations (1162 informants) collected from the *Linguistic Atlas of the Middle and South Atlantic States* (Kretzschmar, 1994). We obtained the simplified phonetic data from `http://www.let.rug.nl/~kleiweg/lamsas/download/`, which in turn was created from data available at `http://us.english.uga.edu/lamsas/`. The U.S. English dataset differentiates 17 vowels (excluding the low-frequency vowels): /i, ɪ, e, ɛ, u, ʊ, æ, a, ɑ, ɒ, ɜ, ɞ, o, ɔ, ʌ, ɐ, ə/.

The Bantu dataset consists of phonetic transcriptions of 160 words in 53

locations and is equal to the subset of the *Atlas Linguistique du Gabon* analyzed and discussed in detail by Alewijnse et al. (2007). The Bantu dataset is distinctive, because varieties of several different languages (e.g., Fang and Tsogo) are included. In contrast to the Dutch, German and U.S. English datasets, the Bantu dataset differentiates only eight vowels (excluding the low-frequency vowels): /e, ɛ, i, o, ɔ, u, a, ə/.

The Bulgarian dataset consists of phonetic transcriptions of 152 words in 197 locations equally distributed over Bulgaria. The dataset was analyzed and discussed in detail by Prokić et al. (2009). Like the Bantu dataset, the Bulgarian dataset is characterized by a relatively small number of vowels (10): /i, e, ɛ, u, ʊ, a, ɑ, o, ɤ, ə/.

The Tuscan dataset, finally, consists of 444 words in 213 locations. In every location on average 10 informants were interviewed. This dataset was analyzed and discussed by Montemagni et al. (in press) and is a subset of the *Atlante Lessicale Toscane* (Giacomelli et al., 2000). As this dataset was compiled with a view to identifying lexical variation (note that we focused on a single lexical form per word), transcriptions are quite crude and consequently only a limited number of vowels were included. The Tuscan dataset thus only differentiates eight vowels (excluding the low-frequency vowels): /i, e, ɛ, u, o, ɔ, a, ə/.

2.2. Acoustic vowel measurements

For every dialect dataset, we obtained formant measurements of the first two formants, F1 and F2. The sources of the Dutch and German formant frequency measurements were identical to those used by Wieling et al. (in press), but we will repeat them here for completeness.

For Dutch, we obtained average vowel formant frequency (Hertz) measurements of 50 male (Pols et al., 1973) and 25 female (van Nierop et al., 1973) speakers of standard Dutch. The formant frequency information was obtained from the initial (stable) part of the stressed vowel waveform and was based on 10 sampling points (i.e. 10 periods generated as a continuous periodic waveform and input to the wave analyzer). In contrast to Wieling et al. (in press), we also included the vowels generally pronounced as diphthongs in standard Dutch (i.e. /e/, /o/, and /ø/) yielding measurements for twelve vowels: /i, ɪ, y, ʏ, e, ɛ, a, ɑ, o, ɔ, u, ø/. We averaged the mean frequencies of men and women in order to obtain a single set of frequencies.

For German, we used average vowel formant frequency measurements of 69 male and 58 female standard German speakers (Sendlmeier and Seebode, 2006) for 14 vowels (stressed, except for the schwa): /i, ɪ, y, ʏ, e, ɛ, a, o, ɔ, u, ʊ, ʌ, ɘ, ə/. We averaged the mean frequencies of men and women in order to obtain a single set of frequencies. Unfortunately, no information was provided about where in the course of the vowel the measurements were taken and how many time points were sampled.

For U.S. English, we used average vowel formant frequency measurements of 45 men and 48 women speaking standard U.S. English (Hillenbrand et al., 1995). The formant frequency information was obtained from the initial (stable) part of the vowel waveform and was based on 7 sampling points. We included acoustic measurements of 11 stressed vowels: /i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ/ and we averaged the mean frequencies of men and women in order to obtain a single set of frequencies.

The Bantu dataset consisted of different languages, but we were only

able to find vowel formant measurements for the Fang language (Nurse and Philippson, 2003, p. 22). We included acoustic measurements of 8 vowels: /i, e, ɛ, ə, a, ɔ, o, u/. Every measurement was based on six pronunciations of the vowel by a single speaker. Unfortunately, no information was provided about where in the course of the vowel the measurements were taken, if the vowels were stressed or not, or how many time points were sampled.

For Bulgarian, we used the formant frequency measurements of a single Bulgarian male speaker (a radio commentator speaking standard Bulgarian) reported by Lehiste and Popov (1970) for 6 vowels: /i, e, ə, a, o, u/. Every measurement was based on 18 pronunciations of the stressed vowel by a single speaker. Unfortunately, no information was provided about where in the course of the vowel the measurements were taken and how many time points were sampled.

For Tuscan, we averaged the formant frequency measurements for two Tuscan dialects (the Pisan and Florentine varieties) reported by Calamai (2003). The formant frequency information was obtained from the (stable) vowel waveform and was based on 3 sampling points. For both dialects, recordings of two male speakers for 7 stressed vowels (pronounced multiple times) were used: /a, ɛ, e, i, ɔ, o, u/.

3. Methods

3.1. Obtaining sound distances based on dialect pronunciations

The automatic procedure we use to determine the segment distances is identical to the approach of Wieling et al. (in press). The procedure first aligns all dialect pronunciations of the same word using the Levenshtein dis-

tance algorithm (minimizing the number of insertions, deletions and substitutions to transform one string into the other; Levenshtein, 1965) employing a binary same-different distinction between the sound segments. To enforce linguistically sensible alignments, the Levenshtein algorithm we employ does not align vowels with consonants.

As an example, consider the application of the Levenshtein distance algorithm to two different dialectal pronunciations of the Dutch word *auto's*, 'cars':

ɑʊtos	delete ɑ	1
ʊtos	subst. ʊ/o	1
otos	insert h	1
othos		
		3

The corresponding alignment is:

ɑ	ʊ	t		o	s
	o	t	h	o	s
1	1		1		

The alignment above clearly illustrates how corresponding segments are identified. Note, however, that the Levenshtein algorithm using binary segment distances will also generate the following alternative alignment (having the same cost):

ɑ	ʊ	t		o	s
o		t	h	o	s
1	1		1		

Based on these initial alignments, the algorithm collects non-identical correspondences such as the [ɑ]:[o] and [ʊ]:∅ (a deletion of [ʊ]) in a large segment × segment variation matrix. For example, the ([ɑ],[o]) cell of the table records how often the [ɑ] aligned with the [o]. These counts are subsequently used in the pointwise mutual information (PMI; Church and Hanks, 1990) formula to determine the association strength between every pair of (non-identical) sound segments:

$$\mathrm{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x)\, p(y)} \right)$$

Where:

- $p(x, y)$ is calculated as the number of times $[x]$ and $[y]$ correspond in aligned pronunciations, divided by the total number of non-identical aligned segments (i.e. the relative occurrence of the aligned sound segments x and y in the whole dataset).

- $p(x)$ and $p(y)$ are calculated as the number of times sound segment x (or y) occurs in non-identical segment correspondences, divided by the total number of individual segments occurring in non-identical segment pairs. Note that dividing by this term normalizes $p(x, y)$ with respect to the probability of x and y being statistically independent.

If x and y correspond more frequently than would be expected by chance, then the PMI value will be positive; otherwise it will be negative. Higher PMI values thus signify more similar sound segments. To convert these similarity values to *positive* distances, we subtract the PMI value from zero and add

the maximum PMI value. The PMI-based segment distance (i.e. PMI distance) between identical segments is always set to zero, as from an alignment perspective no cost accrues to aligning identical sounds.

The procedure above thus assigns low distances to sound segments which correspond relatively frequently in alignments (i.e. more frequently than would be expected on the basis of the individual sound segment's frequency) and high distances to sound segments which correspond relatively infrequently. These sound segment distances are then used in a second iteration of the Levenshtein algorithm (instead of the binary distinctions) to obtain new alignments (and a new variation matrix).

The procedure of obtaining sound segment distances on the basis of the alignments and updating the alignments on the basis of the new sound segment distances is repeated until the alignments (and consequently sound segment distances) remain constant (on average after about 5 iterations). How well these final obtained segment distances correspond with acoustic sound distances is discussed in Section 4. Below we see the effect of the procedure on the example alignment (there is no alternative alignment anymore):

ɑ		ʊ	t		o	s
		o	t	h	o	s
0.035		0.019		0.027		

To appreciate how the present attention to relations is several magnitudes more encompassing than earlier structuralist work, we note that the procedure always involves a large number of correspondences. A word has 4 or 5 segments on average, so an aligned pronunciation pair yields about 5 correspondences. We work with word lists of minimally 152 and maximally 562

words, meaning we obtain 760 to 2810 correspondences per pair of sites. As our datasets contain data from between 53 and 613 sites, there are between 1378 and 187,578 site pairs. Consequently, we collect between 1×10^6 and 5×10^8 correspondences per dataset.

3.2. Calculating acoustic distances

Similar to Wieling et al. (in press) the acoustic distances between vowels are calculated on the basis of the Euclidean distances of the average formant frequencies (in Bark, to correct for our non-linear perception of formant frequency; Traunmüller, 1990). Unfortunately, as we mainly obtained the average formant frequencies from published research, we were not able to apply speaker-based normalization (e.g., Lobanov, 1971).

We employ the acoustic distances to validate the corpus-based PMI procedure, but while the induced segmental distances are based on an entire language area, the acoustic differences have normally been measured using pronunciations according to the standard variety. One might object that we should compare with the acoustics of each of the varieties we examine, but we note that we induce distances from IPA (or other) transcriptions which are used consistently across an entire language area. We therefore take it that we can use the acoustic pronunciations of the relevant IPA vowels according to the standard variety as validation material.

4. Results

For all datasets, Table 1 shows the correlation between the acoustic and PMI distances. We assessed the significance of the correlation coefficients by using the Mantel test (Mantel, 1967), as our sound distances are not

	Pearson's r	Explained variance (r^2)	Significance
Dutch	0.672	45.2%	$p < 0.01$
Dutch w/o Frisian	0.686	47.1%	$p < 0.01$
German	0.633	40.1%	$p < 0.01$
German w/o /ə/	0.785	61.6%	$p < 0.01$
U.S. English	0.608	37.0%	$p < 0.01$
Bantu	0.642	41.2%	$p < 0.01$
Bulgarian	0.677	45.8%	$p < 0.01$
Tuscan	0.758	57.5%	$p < 0.01$

Table 1: Correlations between the acoustic and PMI distances for all datasets. Significance was assessed using the Mantel test (Mantel, 1967). The correlation for the German dataset was also reported by Wieling et al. (in press). The correlation for the Dutch (including Frisian) dataset differs slightly from the value of 0.657 reported by Wieling et al. (in press), as we did not exclude diphthongs in the present study.

completely independent. It is clear that the acoustic and PMI distances match reasonably well, judging by the correlation coefficients ranging from 0.61 to 0.76 (including all vowels).

As we obtain a matrix of vowel distances, we can use multidimensional scaling (MDS; Togerson, 1952) to position each vowel at the optimal position relative to all other vowels in a two-dimensional plane. Figure 1(a) visualizes the relative positions of the Dutch vowels based on their acoustic distances (since these are determined on the basis of the first two formants, the complete variance is always visualized in two dimensions), while Figure 1(b) shows the relative placement of the Dutch vowels on the basis of their PMI distances (the latter figure was reprinted from Wieling et al., in press). Sim-

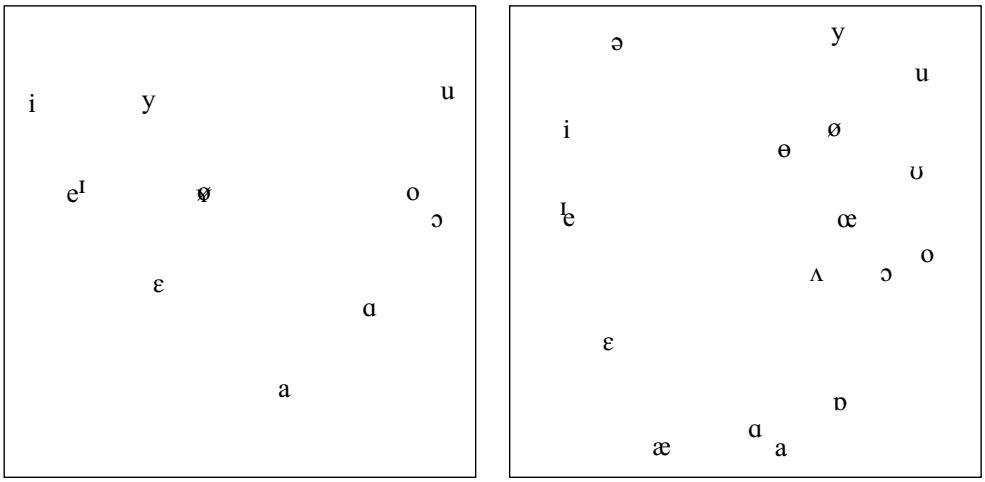
ilarly, Figures 2 to 6 show the relative positions of the vowels based on the acoustic distances (a) as well as the PMI distances (b) for German (both reprinted from Wieling et al., in press), U.S. English, Bantu, Bulgarian and Tuscan. As the MDS calculations did not allow for missing distances, some sounds may be missing from the PMI distance visualizations. When the PMI method did not yield a distance between a pair of sounds (i.e. the two sounds did not align), we excluded one of these sounds from the MDS procedure.[2] Of course, all distances were included when calculating the correlation between the acoustic and PMI distances (shown in Table 1).

The visualizations on the basis of the acoustic distances are all highly similar to the IPA vowel chart. The visualizations on the basis of the PMI-derived distances show more differences with the IPA vowel chart and will be discussed for every figure separately.

In examining the MDS visualizations of the vowels, one should keep in mind that they are visualizations of the relative distances of the vowels to each other — and not simply visualizations of vowels in any absolute coordinate system. So questions regarding the relative position of a certain vowel compared to other vowels can be answered, while those about the absolute position of a vowel (e.g., in the top-right) cannot.

The visualization of the Dutch PMI distances in Figure 1(b) captures 76% of the variation and was discussed briefly by Wieling et al. (in press). Here we offer a more thorough discussion of these results. The visualization reveals quite sensible positions of the [i], [u], [a] and similar sounds,

[2]We excluded the sound which maximized the number of sounds displayed in the MDS visualization.



(a) Acoustic distance visualization (b) PMI distance visualization

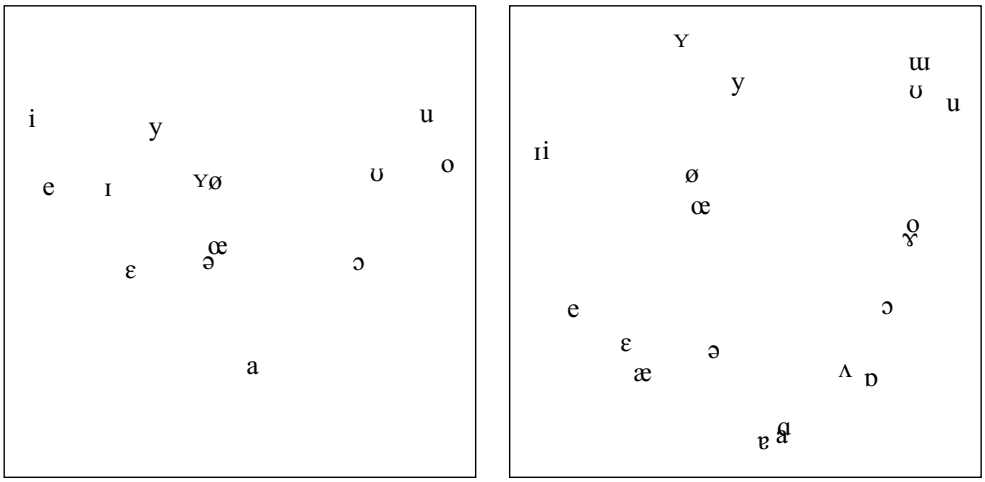
Figure 1: Relative positions of Dutch vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 76% of the variation in the original distances. The right figure was reprinted with permission from Wieling et al. (in press).

especially taking into account that the distances are based *purely* on how frequently the sounds align in dialect data. Unfortunately, the position of the [ə] (schwa) in Figure 1(b) deviates to a great extent from the position on the basis of the acoustic distances. Investigating the underlying alignments revealed that the schwa was frequently deleted, which resulted in relatively high distances between the schwa and the other vowels (which were not deleted as frequently) compared to the other distances. Consequently, excluding the schwa improved the ability to visualize the distances between the vowels adequately in two dimensions: the explained variance of the MDS visualization (not shown) increased from 76% to 85%. A second striking deviation for the Dutch dataset is the position of the front rounded vow-

els, which are surprisingly back (i.e. [y], [ø] and [œ]). Unfortunately, we do not have an immediate explanation for this, but it is likely that this reflects the frequency with which [u] and [y], etc. correspond, which may ultimately suggest a systematic limitation to the technique (i.e. sensitivity to umlaut).[3]

We initially excluded the Frisian dialects from the Dutch dataset as Frisian is recognized as a different language politically and is generally recognized as historically less closely related to Dutch than (for example) English. In addition, Frisian and Dutch dialects have some sound correspondences consisting of rather dissimilar sounds (Wieling and Nerbonne, 2011a), such as [o]:[ɛ] (e.g., *bomen*, 'trees': [bomə] vs. [bjɛmən]) and [a]:[ɪ] (e.g., *kamers*, 'rooms': [kamə[…]

[3]One referee suggested that we might look at the McGurk-effect (McGurk and MacDonald, 1976) for an explanation of why the front rounded vowels group more closely with the back vowels. We find this intriguing, but we see no opportunity to develop the idea here.

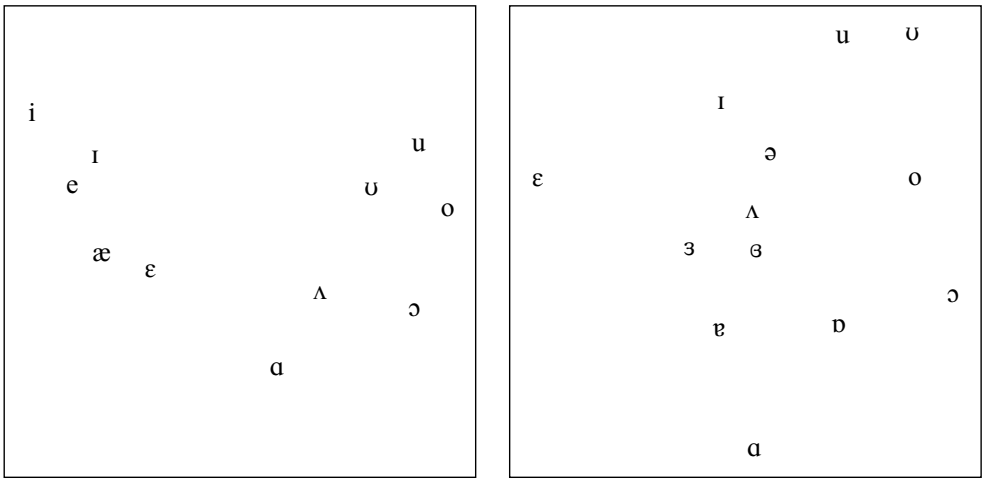


(a) Acoustic distance visualization (b) PMI distance visualization

Figure 2: Relative positions of German vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 70% of the variation in the original distances. Reprinted with permission from Wieling et al. (in press).

and similar sounds. While the position of the schwa was more sensible than in the Dutch visualization shown in Figure 1(b), it was the most frequently deleted sound. Consequently, excluding the schwa increased the explained variance of the visualization from 70% to 83% and also resulted in a higher correlation between the acoustic and PMI distances (see Table 1).

The positions of the vowels based on the U.S. English PMI distances in Figure 3(b) (capturing 65% of the variation) are much more chaotic than the Dutch and German visualizations. If we ignore the [ɛ], the positions of the [ɪ], [ɒ] and [u] seem reasonable, however. The deviating position of the [ɛ] was likely caused by its relatively large distance (i.e. infrequent alignment) from [o] and [u]. Note that [i], [e], [a] and [æ] were excluded from the MDS



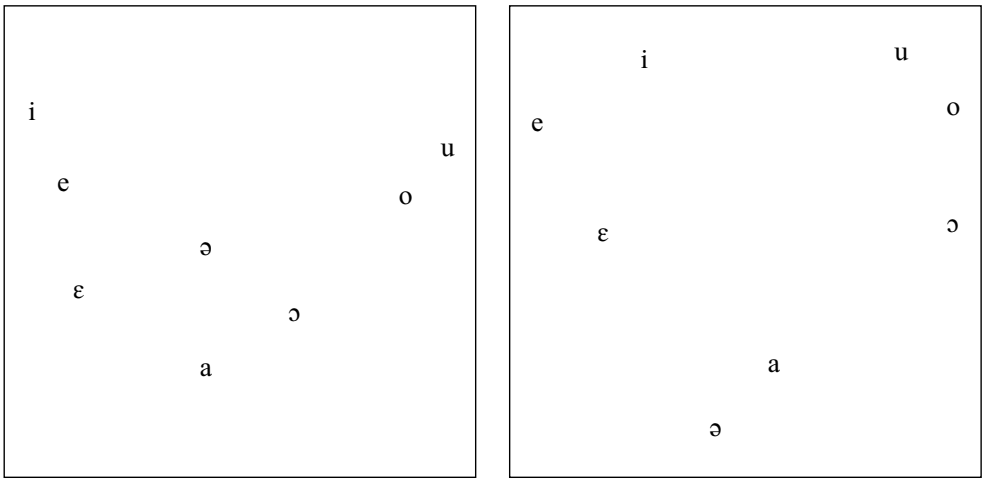
(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3: Relative positions of U.S. English vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 65% of the variation in the original distances.

visualization, as these sounds did not align with all other vowels (and no missing distances were allowed in the MDS procedure).

We turn now to the Bantu data. Similar to Dutch and German, the visualization of the Bantu PMI distances (capturing 90% of the variation) in Figure 4(b) reveals reasonable positions of the [i], [u] and [a]. The most striking deviation is the position of the schwa, caused by its low distance from the [a] and greater distance from [i] and [u].

Similar to the U.S. English visualization, the visualization of the Bulgarian data in Figure 5(b) (capturing 86% of the variation) reveals a deviating position of the [ɛ], likely caused by its relatively large distance from [o] and [u]. Note that the [ɔ] was excluded from the MDS visualization, as this sound



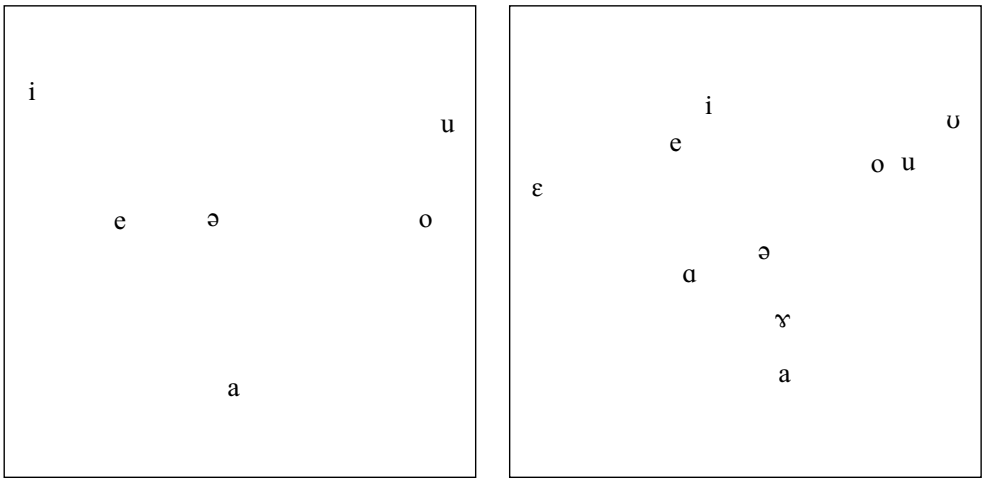
(a) Acoustic distance visualization (b) PMI distance visualization

Figure 4: Relative positions of Bantu vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 90% of the variation in the original distances.

did not align with all other vowels (and no missing distances were allowed in the MDS procedure).

The visualization of the Tuscan PMI distances in Figure 6(b) captures 97% of the variation and shows a reasonably good placement of all sounds. Of course, this is not so surprising as there are only five sounds included in the visualization (i.e. the [ə], [ɔ] and [ɛ] were excluded as these sounds did not align with all other sounds and the MDS procedure did not allow missing distances).

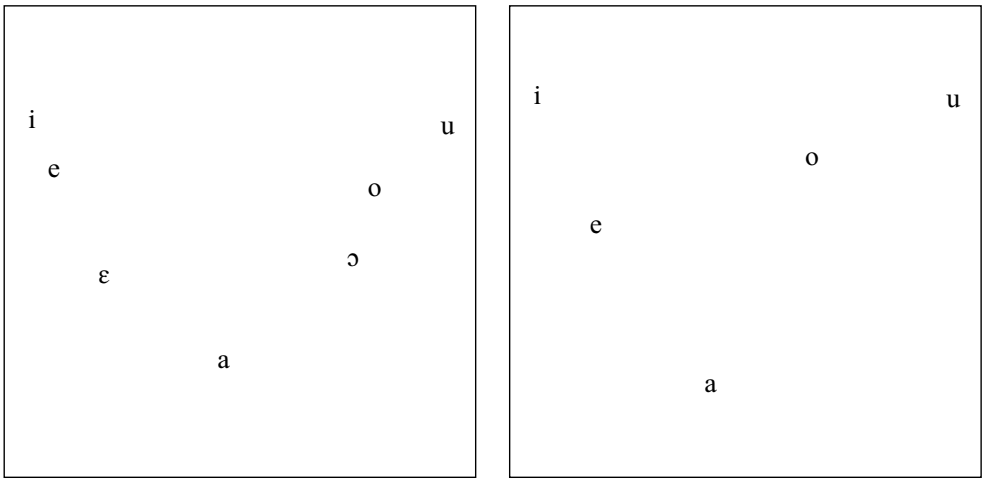
As there are no acoustic distance measurements for consonants, we were not able to evaluate the quality of the automatically generated consonant distances explicitly. To illustrate that the consonant distances also seem quite sensible, Figure 7 shows the MDS visualization of several Dutch consonants



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 5: Relative positions of Bulgarian vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 86% of the variation in the original distances.

(50% of the variance is captured in the visualization). Note that consonants having a frequency lower than one percent of the maximum consonant frequency were excluded, as well as consonants which did not align with all other consonants (no missing distances are allowed in the MDS procedure). Figure 7 clearly shows sensible groupings of the velar consonants [x], [χ], [ɣ], [ɡ], [ŋ] in the upper-left, the rhotics [ʁ], [r], [ɾ] in the upper-right, the alveopalatal consonants [j], [s], [n], [t], [d] in the center, the laterals [l], [ɬ] to the right and the bilabial and labiodental consonants [ʋ], [w], [b], [p], [v] at the bottom. In contrast, the position of the [z] close to the velars is not easy to explain. The visualization of these consonantal distances seem to indicate that place and manner characteristics dominate over voicing.



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 6: Relative positions of Tuscan vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 97% of the variation in the original distances.

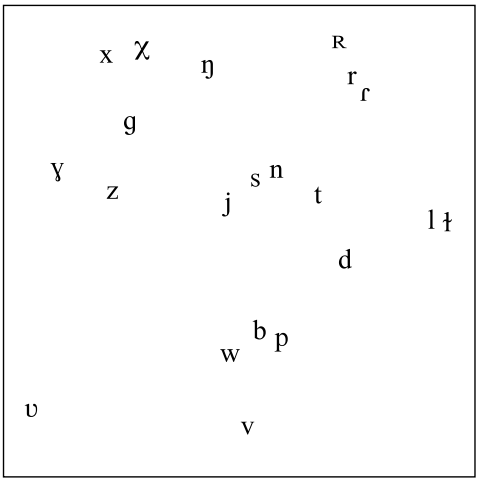


Figure 7: Relative positions of Dutch consonants based on their PMI distances. The visualization captures 50% of the variation in the original distances.

5. Discussion and conclusion

In this paper, we have introduced variation matrices and showed that their structure reflects phonetic distance. So one contribution has been to add to the range of phenomena that fall within the purview of phonetics.

We have gone on to show that the degree to which variation matrices reflect phonetic distance may be calculated by exploiting the dependency of alignment algorithms on segment distances. We used acoustic distances in formant space as an evaluation of the claim that we could derive information about phonetic distances from cross-speaker distributions of variation. The level of correlation between the automatically determined phonetic distances and acoustic distances was similar in six independent dialect datasets and ranged between 0.61 and 0.76, a good indication that the relation between being similar in pronunciation and functioning as an alternative pronunciation is not accidental or trivial.

Of course, one might argue that these results are perfectly in line with what one would expect. Indeed, it is more likely that dialectal pronunciations will be similar, rather than completely different and, consequently, similar sounds will align more frequently than dissimilar sounds. However, we would like to emphasize that this study has quantified how much information is implicit in (cross-speaker) distributions, something which has largely been lacking. Whether or not one is surprised at how much information is found in these distributions undoubtedly depends on one's theoretical convictions, but the present paper has quantified this.

In line with this, the MDS visualizations of the automatically obtained segment distances were never completely identical to the visualizations based

on the acoustic data. In some cases, this was caused by the frequency with which a particular sound segment (e.g., the schwa in Dutch and German) was deleted in the alignments (which consequently affected the other distances), but in other cases acoustically similar sounds simply aligned infrequently. So, while there is a clear connection between acoustic distances and the information about phonetic distances present in the distribution of alternative pronunciations, it is by no means perfect. It would be interesting to see if there is some kind of structure in these deviations. Unfortunately, we do not yet have a clear approach toward investigating this.

Clearly, we first need phonetic events in order to study their distributions. In this sense, our study has demonstrated how to *detect* phonetic relations from (cross-speaker) distributions, but we concede that it would be overeager to imagine that these distributions "cause" the phonetics. We would like to note, however, that there have been demonstrations that distributions within acoustic space do influence children's learning of categories (Maye et al., 2002). There is room in linguistic theory to imagine that distributions in fact do influence phonetics.

We emphasize that we tested our inductive procedure against the ground truth of acoustics, and that we restricted our test to comparisons of vowels only because there is phonetic consensus about the characterization of vowels in a way that supports a measure of distance. While we did not investigate the automatically generated consonantal distances in this paper extensively (as these cannot be validated easily), a visual inspection of Dutch consonantal distances (see Figure 7) suggests that the method also yields satisfying results for consonants.

It is promising that the good performance of the PMI method with respect to the alignment quality (Wieling et al., 2009) is also supported by a relatively strong correlation between the PMI distances and the acoustic distances. Obtaining improved alignments and assessing pronunciation and sound segment distances more accurately is valuable in dialectometry (see, e.g., Wieling et al., 2011, who attempt to predict PMI-based dialectal word pronunciation distances on the basis of several word-related and sociolinguistic factors) and also in historical linguistics where the identification of regular sound correspondences is important.

Of course, we have not tried to demonstrate that improved segment distance measures lead to genuine improvements in all the various areas discussed in the introduction, including not only historical linguistics and dialectometry, but also second-language learning (foreign accents), spelling correction, and the study of speech disorders. We note merely that there is broad interest in measures of phonetic segment similarity, the focused issue to which we contribute. We are well aware that potential and genuine improvements are two very different matters.

Finally, we suggest that the results be viewed as vindicating the structuralists' postulate that the sound *system* of a language is of central importance, as this is reflected in the relations among variant pronunciations. We have shown that distributions (of alternative dialectal pronunciations) contain enough information to gauge content (i.e. phonetic similarity) to some extent. The only phonetic content made available to the algorithm was the distinction between vowels and consonants, and yet the algorithm could assign a phonetic distance to all pairs of vowel segments in a way that correlates

fairly well with acoustic similarity. We know of no work in the strict structuralist tradition that attempted to analyze corpora of 10^8 segment pairs, nor of attempts to analyze entire tables reflecting pronunciation relations. We nonetheless find it appropriate to emphasize that our focus in this paper is very much in the structuralist tradition of understanding the systems by studying relations within it.

As noted in the introduction, the modern study of pronunciation often emphasizes the need to go beyond the distributions of sounds, and therefore the need to interpret pronunciation physically. From this perspective, it is interesting that this study has shown that we are able to characterize the phonetic distance between segments (in a data-driven manner) fairly well on the basis of the distribution of the segment's pronunciation variants among closely related varieties.

Acknowledgments

We thank Peter Kleiweg for implementing the PMI procedure in the L04 package which was used to generate the vowel distance visualizations. Mark Liberman discussed the ideas in this paper with us generously. Finally, we thank the three anonymous reviewers and the editor for their extensive and helpful comments.

References

Alewijnse, B., Nerbonne, J., Van der Veen, L., Manni, F., 2007. A computational analysis of Gabon varieties. In: Osenova, P. (Ed.), Proceedings of the RANLP Workshop on Computational Phonology. pp. 3–12.

Almeida, A., Braun, A., 1986. 'Richtig' und 'Falsch' in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. Zeitschrift für Dialektologie und Linguistik LIII (2), 158–172.

Bailey, T. M., Hahn, U., 2005. Phoneme similarity and confusability. Journal of Memory and Language 52 (3), 339–362.

Bloomfield, L., 1933. Language. Holt, Rhinehart and Winston, New York.

Broe, M., 1996. A generalized information-theoretic measure for systems of phonological classification and recognition. In: Computational Phonology in Speech Technology: Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology. Association for Computational Linguistics, Santa Cruz, pp. 17–24.

Calamai, S., 2003. Vocali fiorentine e vocali pisane a confronto. Quaderni del Laboratorio di Linguistica, Scuola Normale Superiore di Pisa 3, 40–71.

Chomsky, N. A., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge.

Church, K. W., Hanks, P., 1990. Word association norms, mutual information, and lexicography. Computational Linguistics 16 (1), 22–29.

Cole, J., 2010. Editor's note. Laboratory Phonolgy 1 (1), 1–2.

Cucchiarini, C., 1993. Phonetic transcription: A methodological and empirical study. Ph.D. thesis, Katholieke Universiteit Nijmegen.

Flege, J., Munro, M., MacKay, I., 1995. Factors affecting strength of perceived foreign accent in a second language. Journal of the Acoustical Society of America 97 (5), 3125–3134.

Fox, R. A., 1983. Perceptual structure of monophthongs and diphthongs in english. Language and Speech 26 (1), 21–60.

Frisch, S., 1996. Similarity and frequency in phonology. Ph.D. thesis, Northwestern University.

Frisch, S. A., Pierrehumbert, J. B., Broe, M. B., 2004. Similarity Avoidance and the OCP. Natural Language & Linguistic Theory 22 (1), 179–228.

Giacomelli, G., Agostiniani, L., Bellucci, P., Giannelli, L., Montemagni, S., Nesi, A., Paoli, M., Picchi, E., Salani, T. P. (Eds.), 2000. Atlante Lessicale Toscano. Lexis Progetti Editoriali, Roma.

Goeman, T., Taeldeman, J., 1996. Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. Taal en Tongval 48, 38–59.

Göschel, J., 1992. Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas". Wissenschaftlicher Bericht, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Gusfield, D., 1999. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge Univesity Press, Cambridge.

Heeringa, W., 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance. Ph.D. thesis, Rijksuniversiteit Groningen.

Hillenbrand, J., Getty, L., Clark, M., Wheeler, K., 1995. Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America 97 (5), 3099–3111.

Hock, H. H., Joseph, B. D., 1996. Language history, language change, and language relationship: An introduction to historical and comparative linguistics. Walter de Gruyter, Berlin.

Kernighan, M., Church, K., Gale, W., 1990. A spelling-correction program based on the noisy channel model. In: Kahlgren, H. (Ed.), Proc. of COLING '90. Helsinki, pp. 205–210.

Kondrak, G., Dorr, B., 2006. Automatic identification of confusable drug names. Artificial Intelligence in Medicine 36 (1), 29–42.

Kretzschmar, W. A. (Ed.), 1994. Handbook of the Linguistic Atlas of the Middle and South Atlantic States. The University of Chicago Press, Chicago.

Laver, J., 1994. Principles of Phonetics. Cambridge Univeristy Press, Cambridge.

Lehiste, I., Popov, K., 1970. Akustische Analyse bulgarischer Silbenkerne. Phonetica 21, 40–48.

Levenshtein, V., 1965. Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR 163, 845–848.

Lobanov, B., 1971. Classification of Russian vowels spoken by different speakers. The Journal of the Acoustical Society of America 49, 606.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27, 209–220.

Maye, J., Werker, J., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82, B101–B111.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Meillet, A., 1903. Introduction l'étude comparative des langues indoeuropéennes. Librairie Hachette et Cie, Paris.

Miller, G. A., Nicely, P. E., 1955. An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America 27, 338–352.

Montemagni, S., Wieling, M., de Jonge, B., Nerbonne, J., in press. Patterns of language variation and underlying linguistic features: a new dialectometric approach. In: Proceedings of the SILFI-2010 conference.

Nerbonne, J., Heeringa, W., 2009. Measuring dialect differences. In: Schmidt, J. E., Auer, P. (Eds.), Theories and Methods. Language and Space. Mouton De Gruyter, Berlin, pp. 550–567.

Nerbonne, J., Siedle, C., 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. Zeitschrift für Dialektologie und Linguistik 72, 129–147.

Nurse, D., Philippson, G., 2003. The Bantu languages. Routledge, London.

Ohala, J. J., 1997. Comparison of speech sounds: Distance vs. cost metrics. In: Speech Production and Language. In honor of Osamu Fujimura. Mouton de Gruyter, Berlin, pp. 261–270.

Pierrehumbert, J. B., 1993. Dissimilarity in the arabic verbal roots. In: Proceedings of the North East Linguistics Society. Vol. 23. GLSA, Amherst, MA, pp. 367–381.

Piske, T., MacKay, I. R., Flege, J. E., 2001. Factors affecting degree of foreign accent in an l2: A review. Journal of Phonetics 29 (2), 191–215.

Pols, L., Tromp, H., Plomp, R., 1973. Frequency analysis of dutch vowels from 50 male speakers. The Journal of the Acoustical Society of America 43, 1093–1101.

Prokić, J., Nerbonne, J., Zhobov, V., Osenova, P., Simov, K., Zastrow, T., Hinrichs, E., 2009. The computational analysis of Bulgarian dialect pronunciation. Serdica Journal of Computing 3, 269–298.

Sanders, N., Chin, S. B., 2009. Phonological distance measures. Journal of Quantitative Linguistics 16 (1), 96–114.

Sankoff, D., Kruskal, J. (Eds.), 1999. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. CSLI, Stanford, [1]1983, with a foreword by John Nerbonne.

Sendlmeier, W., Seebode, J., 2006. Formantkarten des deutschen

Vokalsystems. TU Berlin, http://www.kgw.tu-berlin.de/forschung/Formantkarten (accessed: November 1, 2010).

Togerson, W., 1952. Multidimensional scaling. I. Theory and method. Psychometrika 17, 401–419.

Tougaard, J., Eriksen, N., 2006. Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure. Behaviour 143 (2), 239–252.

Traunmüller, H., 1990. Analytical expressions for the tonotopic sensory scale. The Journal of the Acoustical Society of America 88, 97–100.

van Heuven, V. J., van Bezooijen, R., 1995. Quality evaluation of synthesized speech. In: Paliwal, K. (Ed.), Speech coding and synthesis. Elsevier Science, Amsterdam, pp. 707–738.

van Nierop, D., Pols, L., Plomp, R., 1973. Frequency analysis of Dutch vowels from 25 female speakers. Acoustica 29, 110–118.

Wieling, M., Heeringa, W., Nerbonne, J., 2007. An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data. Taal en Tongval 59, 84–116.

Wieling, M., Margaretha, E., Nerbonne, J., in press. Inducing phonetic distances from dialect variation. Computational Linguistics in the Netherlands Journal 1.

Wieling, M., Nerbonne, J., 2011a. Bipartite spectral graph partitioning for

clustering dialect varieties and detecting their linguistic features. Computer Speech & Language 25 (3), 700–715.

Wieling, M., Nerbonne, J., 2011b. Measuring linguistic variation commensurably. Dialectologia Special Issue II: Production, Perception and Attitude, 141–162.

Wieling, M., Nerbonne, J., Baayen, R. H., 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. PLoS ONE 6 (9), e23613.

Wieling, M., Prokić, J., Nerbonne, J., 2009. Evaluating the pairwise alignment of pronunciations. In: Borin, L., Lendvai, P. (Eds.), Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. pp. 26–34.