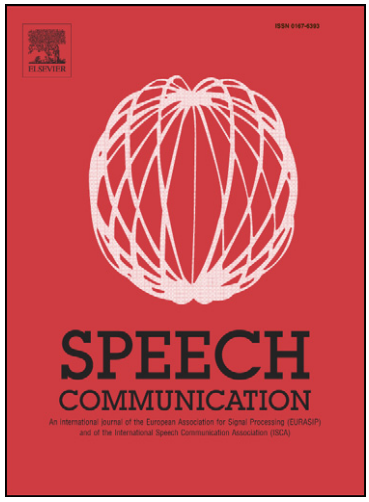
Measuring Norwegian Dialect Distances using Acoustic Features

Wilbert Heeringa, Keith Johnson, Charlotte Gooskens



Measuring Norwegian Dialect Distances using Acoustic Features

Wilbert Heeringa, Keith Johnson, Charlotte Gooskens

PII: S0167-6393(08)00125-8
DOI: 10.1016/j.specom.2008.07.006
Reference: SPECOM 1741

To appear in: *Speech Communication*

Received Date: 12 May 2006
Revised Date: 23 July 2008
Accepted Date: 23 July 2008

Please cite this article as: Heeringa, W., Johnson, K., Gooskens, C., Measuring Norwegian Dialect Distances using Acoustic Features, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.07.006

Measuring Norwegian Dialect Distances using Acoustic Features ⋆

Wilbert Heeringa

University of Groningen, Department of Information Science

Keith Johnson

UC Berkeley, Department of Linguistics

Charlotte Gooskens

University of Groningen, Scandinavian Department

Abstract

Levenshtein distance has become a popular tool for measuring linguistic dialect distances, and has been applied to Irish Gaelic, Dutch, German and other dialect groups. The method, in the current state of the art, depends upon phonetic transcriptions, even when acoustic differences are used the number of segments in the transcriptions is used for speech rate normalization.

The goal of this paper is to find a *fully* acoustic measure which approximates the quality of semi-acoustic measures that rely on tagged speech. We use a set of 15 Norwegian dialect recordings and test the hypothesis that the use of the acoustic signal only, without transcriptions, is sufficient for obtaining results which largely agree with both traditional Norwegian dialectology and the perception of the speakers themselves.

We use formant trajectories and consider both the Hertz and the Bark scale. We experiment with an approach in which z-scores per frame are used instead of the original frequency values. Besides formant tracks, we also consider zero crossing rates: the number of times per interval that the amplitude waveform crosses the zero line. The zero crossing rate is sensitive to the difference between voiced and unvoiced speech sections.

When using the fully acoustic measure on the basis of the combined representation with normalized frequency values, we obtained results comparable with the results obtained with the semi-acoustic measure. We applied cluster analysis and multidimensional scaling to distances obtained with this method and found results which largely agree with both the results of traditional Norwegian dialectology and with the perception of the speakers. When scaling to three dimensions, we found the first dimension responsible for gender differences. However, when leaving out this dimension, dialect specific information is lost as well.

Key words: acoustics, acoustic features, dialectology, dialectometry, dialect, phonetics, phonology

1 Introduction

Computational dialectometry has been proven to be useful for finding dialect relationships and identifying dialect areas. The first to develop a method of measuring dialect distances was Jean Séguy, assisted and inspired by Henri Guiter (Chambers and Trudgill, 1998). Strongly related to the methodology of Séguy is the work of Goebl, although the basis of Goebl's work was developed mainly independently of Séguy (Goebl, 1982, 1993). In the methodology of both Séguy and Goebl two items to be compared (lexically, phonetically, syntactically or at other levels) are the same or different. Distinctions are binary.

In 1995 Kessler used the *Levenshtein distance* for finding linguistic distances between Irish Gaelic dialect varieties, and in 1996 the same algorithm was applied to Dutch dialect varieties by Nerbonne et al. The Levenshtein distance is a sensitive measure with which distances between strings (in this case transcriptions of word pronunciations) are calculated. This means that distinctions between pronunciations of a particular word are gradual rather than just binary. Gooskens and Heeringa (2004) showed that linguistic distances between 15 Norwegian varieties measured with Levenshtein distance correlate significantly with perceptual distances measured between the same 15 Norwegian varieties ($r = 0.67$).

⋆ We thank Jørn Almberg for his permission to use the recordings and transcriptions of 'The North Wind and the Sun' and Sabine Rosenhart for help with cutting the word samples. We thank Arnold Dalen for his help in finding a reliable dialect map and for classifying each of the 15 dialect varieties in the right dialect group in accordance with this traditional dialect map. We thank Peter Kleiweg for letting us use the programs which he developed for the graphic representation of the maps, dendrograms and multidimensional scaling plots in the present article. We thank John Nerbonne for reviewing the English and giving useful comments. We thank Therese Leinonen for usefull discussion about the characteristics of Norwegian dialects. We are grateful to the two anonymous reviewers for their valuable remarks. This research was carried out within the framework of a talentgrant project, which is supported by a fellowship (number S 30-624) from the Netherlands Organisation of Scientific Research (NWO).

Email addresses: wilbert.heeringa@meertens.knaw.nl (Wilbert Heeringa), keithjohnson@berkeley.edu (Keith Johnson), c.s.gooskens@rug.nl (Charlotte Gooskens).

Although the introduction of Levenshtein distance in the field of dialectometry was a significant improvement for comparing dialects phonetically, the results still depend on the quality of the phonetic transcriptions, which may vary greatly, depending on the skills, or idiosyncratic habits, of the transcriber. When several transcribers are involved, the data may reveal 'dialect' differences that are actually merely differences between transcribers. For example, Heeringa (2005), found the Frisian dialect area to be divided in a northern and southern part, which reflected the work areas of the two transcribers. The effect of different transcribers on the transcriptions was also be seen in an analysis of the whole Dutch dialect area (see Heeringa (2004), pp. 235–266).

A first attempt to measure dialect distances directly on the basis of the acoustic signal was made by Heeringa and Gooskens (2003). But some information from the transcriptions was still used. The number of segments in a pronunciation was used for the purpose of speech rate normalization. We will refer to this methodology as the semi-acoustic approach.

The goal of this paper is to go one step further and to find a fully acoustic measure which approximates the quality of the semi-acoustic measure of Heeringa and Gooskens (2003). We test the hypothesis that varieties of Norwegian can be classified on the basis of acoustic features only, without the filter of a given listener (i.e. the transcriber). The classification scheme obtained in this way correlates significantly with both the traditional dialectology criteria, which classifies the dialects according to a number of relevant linguistic features (phonological, lexical, etc.) and the results of a perceptual classification experiment. We will experiment with different representations of the acoustic signal to investigate which representation gives the best results.

The basis of the research presented in this paper is a database which contains recordings of Norwegian dialect varieties compiled by Jørn Almberg and Kristian Skarbø. [1] The database comprises recordings of translations of the fable 'The North Wind and the Sun'. In this paper we will compare our results to the results of a perception experiment reported by Charlotte Gooskens. When the perception experiment was carried out, recordings of only 15 varieties were available. Therefore we use the same 15 varieties. Today more than 50 recordings are available, giving much better possibilities to pick a representative selection of varieties.

Section 2 gives a brief overview of the main linguistic phenomena that play a role in traditional Norwegian dialectology. In Section 3 we describe the perception experiment. In Section 4 we describe our acoustic model and its parameters. In Section 5 we validate the results of our methodology. Section 6 shows results. The results will be compared to both the results of traditional Norwe-

[1] Department of Linguistics, University of Trondheim. The recordings are available at `http://www.ling.hf.ntnu.no/nos/`.

gian dialectology and the results of the perception experiment. In Section 7 some conclusions will be drawn.

2 Traditional Norwegian dialectology

In traditional Norwegian dialectology the dialect map of Skjekkeland (1997) is an authoritative map. It divides the Norwegian dialect area in two main groups, Vestnorsk (northern and southwestern varieties) and Austnorsk (southeastern varieties). Vestnorsk is divided in Nordnorsk (north) and Vestlandsk (southwest). Nordnorsk, Vestlandsk and Austnorsk in turn are divided in three smaller groups, giving a total of nine groups. In our set of 15 dialect varieties, six groups are represented. On the map in Figure 1 the six groups as represented by the 15 varieties are shown.

Skjekkeland's map is based on 24 single linguistic features. In this section we classify the 15 varieties on the basis of especially those features which are represented in our data, i.e. the transcriptions of the Norwegian translations of 'the North Wind and the Sun'. The text consists of 58 different words. Since we have a translation of the text for each of the 15 varieties, we have a translation of each of the 58 words in (nearly) all dialects. Due to the free translation of some phrases for certain varieties a few of the expected words were missing. If the same word appears more than once in a text, we consider only the first occurrence.

2.1 Features

In this section we give an overview of the features which have contributed to Skjekkeland's dialect classification and which are represented in our data. In some cases we discuss features which are closely related to the ones mentioned by Skjekkeland.

Apocopation of verb endings. Skjekkeland shows that the endings of infinitive verbs and weak feminine nouns might have different pronunciations or they have been apocopated. In our data we found three infinitive verbs which show a clear distinction between forms with final vowel and apocopated forms: *kunne* 'could', *gjelde* 'count' and *innrømme* 'admit'. The same distinction was found for third person singular of three verbs given in the paste tense: *skulle* 'would', *blåste* 'blew' and *måtte* 'had to'. The distributions of the six verbs are shown in Table 1. The table shows that we find more apocopated verbs when

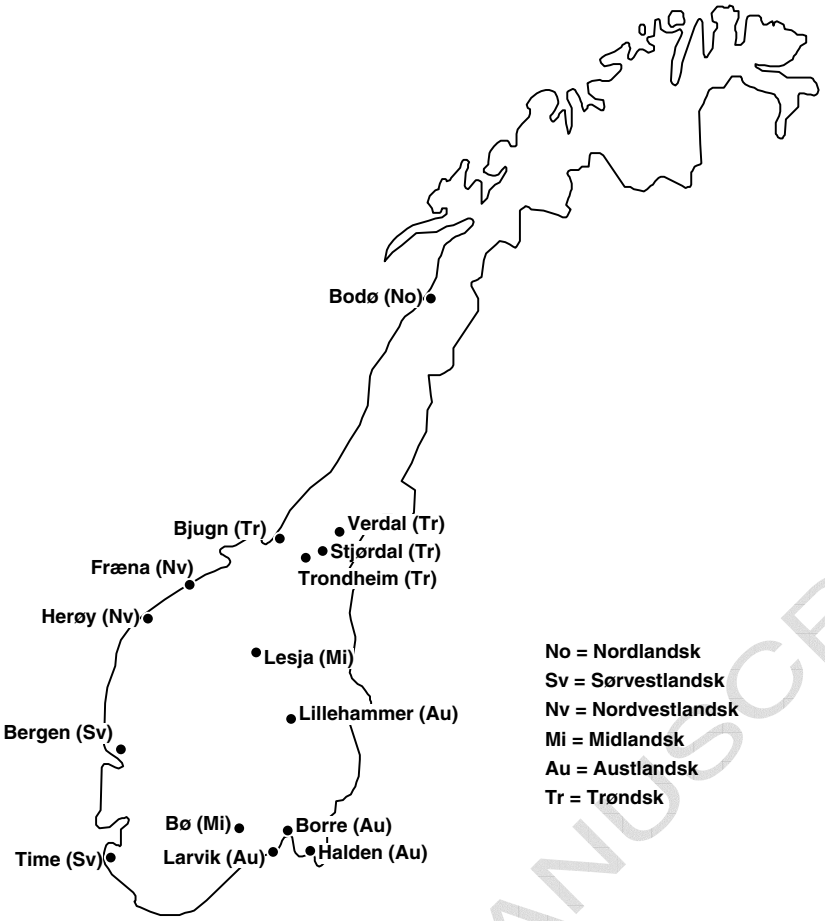


Fig. 1. Map of Norway showing the 15 dialect varieties in the present investigation. Skjekkeland (1997) distinguishes nine Norwegian dialect groups. Six groups are represented by our set of 15 varieties. The abbreviation after the name of each location indicates the dialect group to which the variety belongs. The same abbreviations are used in the other figures in this paper. Skjekkeland (1997) also gives a more global division in which the Norwegian dialect area is divided in *Vestnorsk* (covering No, Sv and Nv) and *Austnorsk* (covering Mi, Au and Tr).

we go further to the north. In the Trøndsk varieties of Bjugn, Stjørdal, Trondheim and Verdal apocopated verbs are found only. Each of the six columns of the table may be considered as a separate feature which we use for classifying the 15 varieties.

Toneme 1 intonations. Pitch and intonation contours are known to be significant dialect markers in Norwegian (Christiansen, 1954; Fintoft and Mjaavatn, 1980). Skjekkeland distinghuishes four different realizations of toneme 1. This toneme is described by Kristoffersen (2000). According to Skjekkeland, the Trøndsk varieties and the variety of Lesja are one group, the Nordvestlandsk varieties together with the Sørvestlandsk variety of Bergen are one

	kun-ne	gjel-de	inn-røme	skul-le	blås-te	måt-te
Bergen	no	no	no	no	no	no
Bjugn	yes	yes	yes	yes	yes	yes
Bodø	no	?	yes	yes	no	no
Boe	no	no	?	yes	yes	no
Borre	no	?	no	no	no	no
Fræna	no	?	no	?	yes	no
Halden	no	no	no	no	no	no
Herøy	no	?	?	?	yes	no
Larvik	no	no	no	yes	no	no
Lesja	yes	no	no	yes	yes	no
Lillehammer	no	no	no	no	no	no
Stjørdal	yes	?	yes	yes	yes	yes
Time	no	no	no	yes	no	no
Trondheim	yes	?	yes	?	yes	yes
Verdal	yes	yes	yes	yes	yes	yes

Table 1
For six verbs it is shown in which varieties they are apocopated ('yes') and in which variaties they are not ('no'). If the desired lexeme was not found in the text, a '?' is put in the cell. Note that in the Trøndsk varieties of Bjugn, Stjørdal, Trondheim and Verdal apocopated verbs are found only.

group, the Austlandsk varieties together with the Midlandsk variety of Bø are one group, and the Nordlandsk variety of Bodø and the Sørvestlandsk variety of Time are in the same group. Toneme 1 intonations are frequently found among the data. In the transcriptions of the data the exact realizations are not noted, therefore we adopted them from Skjekkeland. For each of the 58 words of the text 'the North Wind and the Sun' we checked whether they have a toneme 1 intonation in most varieties. We found that in each of 26 words the majority of the 15 varieties have toneme 1. Therefore we counted this feature 26 times when classifying the varieties.

Palatalization of alveolar consonants. This feature is represented in 12 words. In 11 words the [n] is palatalized to a [ɲ], and in one word the [s] is palatalized to a [ʃ]. The words suggest a division between the northern and southern varieties. Five words suggest a division between Bodø, Bjugn, Verdal, Stjørdal, Fræna, Herøy and Lesja in the north and the other varieties

(including Trondheim) in the south.

Pronunciation of the /r/. In our data set this feature is represented in 13 words. In the two Sørvestlandsk varieties of Bergen and Time the velar approximant [ɰ] is pronounced while in other varieties usually the alveolar tap [ɾ] is used.

Voicing of voiceless plosives. Skjekkeland mentions that the plosives [p], [t] and [k] are (partly) voiced in some varieties. In our data this feature is represented in the word *tok* 'took'. In the varieties of Time and Bø the final consonant is the voiced [g].

Assimilation of /d/ before an alveolar sonorant. Skjekkeland shows that the /ld/ has been assimilated to /ll/ in many Norwegian dialect varieties. This feature is not represented in our data. However we found three words in which the assimilation of /nd/ to /nn/ or /n/ was found, namely in *nordavinden* 'the north wind', *gåande* or *gående* 'going' and *kunne* 'could'. The /nd/ pronunciation is retained in the variety of Herøy, all other varieties have /nn/ or /n/.

Change of initial /hv/. Initial /hv/ may have been changed into /v/ or /gv/ or /kv/. The variation is reflected in the Norwegian translations of the interrogative word 'who' which may be either *hvem* (pronounced with initial [ʋ]) or *kven* (pronounced with initial [kʰ]). The varieties of Trondheim, Bergen, Lillehammer, Larvik, Borre and Halden have initial [ʋ], the other ones have [kʰ].

Endings of definite nouns. Skjekkeland shows the variation in the pronunciations of endings of datives of masculine or neutral strong singular definite nouns. This phenomenon is not found in our data, but we found a related phenomenon. We found two objects: *mannen* 'the man' and *frakken* 'the coat'. While the pronunciation of *mannen* usually ends on [ɲː] or [nː], it ends on [ɲːɪ] in the variety of Herøy. The pronunciation of *frakken* usually ends on [ən], but in the variety of Fræna it ends on [ɪŋ], and in the variety of Herøy it ends on [ɪ]. This phenomenon distinguishes the Nordvestlandsk varieties from the other ones.

Lexical variation. Skjekkeland showns the lexical forms for the subject form of the first person singular *jeg* 'I' and the lexical forms for the subject form of the first person plural *vi* 'we'. The text 'the North Wind and the Sun' does not contain these two pronouns. However, 17 words vary lexically. The main pattern suggested by the lexical variation is a north versus south division which is the same as the division suggested by the palatalization feature discussed above.

Conjugation of verbs. In our data especially the conjugation of the past tense of the third person singular is represented, namely by the following words: *blåste* 'blew', *trakk* 'drew', *skein* or *skinte* 'shone' and *måtte* 'must'. In general the southern varieties and the Nordlandsk variety of Bø have a form ending on [ə] or [tə], but that ending is never found for the Trøndsk varieties.

2.2 Classification

In Section 2.1 we found ten different phenomena, mentioned by Skjekkeland, and represented in our data. Some phenomena were represented in more than one word. We consider each occurence of a phenomenon in a particular word as a feature (cf. Table 1). In this way we identified 85 features.

On the basis of the 85 features we calculated distances between the varieties, using a simple metric introduced by Jean Séguy (cf. Séguy (1973)). The distance between two varieties is equal to the number of features on which they disagree. Varieties which agree on all features have a distance of 0, and those who differ on all 85 features have a distance of 85.

In order to visualize the relationship between the dialect varieties, we performed a cluster analysis using *group average* (see Jain and Dubes (1988)) on the basis of these distances. The resulting tree is shown in Figure 2. In addition a multidimensional scaling analysis was carried out. In our research we used MDS routines as implemented in the statistical R package.[2] The resulting plot is shown in Figure 3.

Both the tree and the plot agree with the classification of Skjekkeland for the greater part. But different from Skjekkeland's division, the varieties of Lesja and Bø are not found in one Midlandsk group. There may be two explanations. First, not all phenomena mentioned by Skjekkeland are frequently represented in the data, which may indicate that the text 'the North Wind and the Sun' is not a representative text. However, if the text *is* representative, the missing

[2] The program R is a free public domain program and available via `http://www.r-project.org/`.

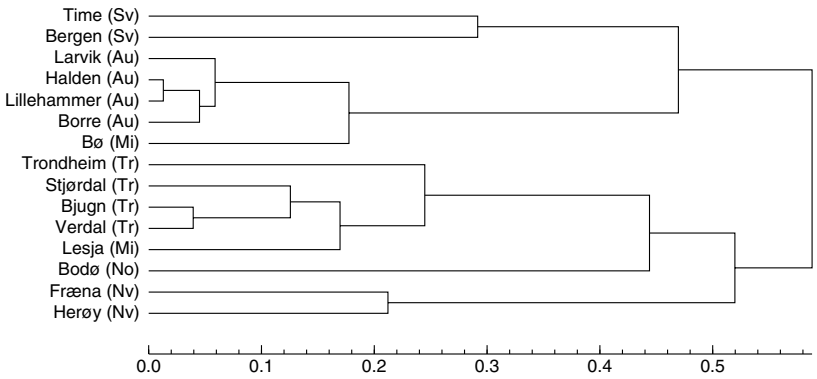


Fig. 2. The tree shows the classification of 15 Norwegian varieties on the basis of phenomena mentioned by Skjekkeland (1997) and represented in the dialect data. The tree structure explains 82% of the variance.

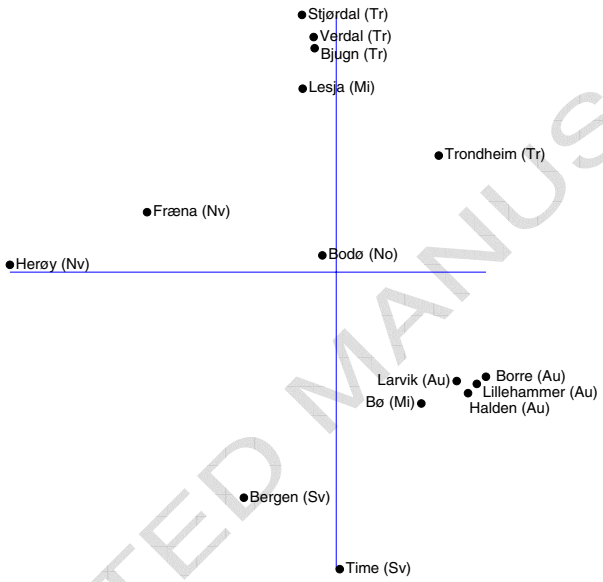


Fig. 3. The plot shows the classification of 15 Norwegian varieties on the basis of phenomena mentioned by Skjekkeland (1997) and represented in the dialect data. The two dimensions explain 89% of the variance.

phenomena may be less important than suggested by Skjekkeland. Second, the distribution of the linguistic features shown by Skjekkeland may have changed. The oldest map dates from 1969, the newest one from 1988.

Trondheim is clustered with the Trøndsk varieties, but not so closely. This variety shares palatalization with the Austlandsk varieties, but has the same realization of toneme 1 as in the Trøndsk varieties.

3 Perceptual distance measurements

In this section we briefly describe the perception experiment and show some results. A detailed description is given by Gooskens and Heeringa (2004).

3.1 Experiment

In order to obtain distances between 15 Norwegian varieties as perceived by Norwegian listeners, a recording of a translation of the fable 'The North Wind and the Sun' in each of the 15 varieties was presented to Norwegian listeners in a listening experiment. The listeners were 15 groups of high school pupils, one from each of the places where the 15 varieties are spoken. All pupils were familiar with their own dialect variety and had lived most of their lives in the place in question (on average 16.7 years). Each group consisted of 16 to 27 listeners. The mean age of the listeners was 17.8 years, 52 percent were female and 48 percent male.

The texts of the 15 varieties were presented in a randomized order. Every session was preceded by an example recording. While listening to the dialect varieties the listeners were asked to judge each of the 15 varieties on a scale from 1 (similar to native dialect variety) to 10 (not similar to native dialect variety). This means that each group of listeners judged the linguistic distances between their own variety and the 15 varieties, including their own variety. In this way we get a matrix with 15×15 distances. There are two mean distances between each pair of varieties and these need not be the same. For example the distance which the listeners from Bergen perceived between their own variety and the variety of Trondheim is different from the distance as perceived by the listeners from Trondheim to Bergen. In the analyses in Section 5 we will use both measurements. In that way asymmetries are fully taken into account.

When classifying dialect varieties (see Section 3.2) the two mean distances between each pair of varieties are averaged. We are aware of the fact that this value may not fully reflect the perception of the speakers. But the classification procedures we use require that the relationship between two varieties is expressed as one single value.

3.2 Results

On the basis of the matrices of the mean judgments, we classified the 15 varieties by performing cluster analysis and multidimensional scaling. The results are found in Figures 4 and 5 respectively.

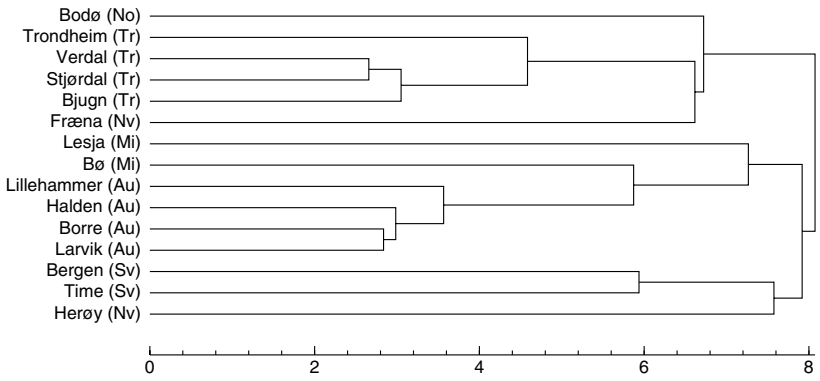


Fig. 4. Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian varieties. The tree structure explains 91% of the variance.

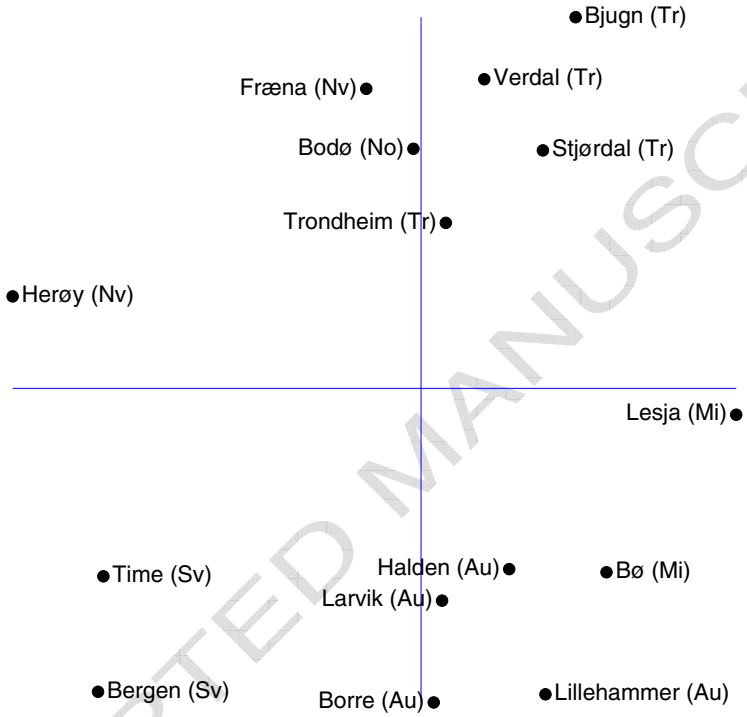


Fig. 5. Multidimensional scaling of the results derived from the 15 × 15 matrix of perceptual distances. The vertical axis represents the first dimension, and the horizontal axis the second dimension. The two dimensions explain 67% of the variance.

The dendrogram and the multidimensional scaling plot agree with each other and partly agree with the the classification we found at the end of Section 2. We clearly find the Trøndsk and Austlandsk varieties as groups. We do not find the Nordvestlandsk varieties grouped together. Fræna is grouped with the Trøndsk varieties which might be explained by the fact that this variety shares palatalization of alveolars with the Trøndsk varieties. Herøy also shares palatalization, but is distinguished from Fræna by retaining the cluster /nd/ which has been assimilated to /nn/ or /n/ in all other varieties. This, how-

ever, does not explain why Herøy is clustered together with the Sørvestlandsk varieties in the dendrogram.

Just as in the traditional tree (see Figure 2) the variety of Trondheim appears as an outlier in the group of Trøndsk varieties. As mentioned in Section 2.2 this variety shares palatalization with the Austlandsk varieties, which makes this variety distant to the other Trøndsk varieties.

Lesja has been grouped together with Bø and the Austlandsk varieties. In the classification results in Section 2 we found Lesja to be grouped among the Trøndsk varieties, which may be explained by the fact that this variety shares the toneme 1 realization and palatalization with these varieties. However, some relationship to the southeastern varieties is found in the pronunciation of the central vowel in *gåande* or *gående* 'going' which is the diphthong [oːɑ] in most varieties, but a monophthong in both Lesja and the southeastern varieties: [ɔ] (Lesja), [ɔː] (Lillehammer and Borre) and [oː] (Halden). Some relationship is also found in the word *varmt* 'warm'. Most varieties have final /rmt/, but the Midlandsk varieties (including Lesja) and Austlandsk varieties have final /nt/. Skjekkeland (1997) does not give maps which show the geographic distribution of these two phenomena, but they may have played a role in the perception of the subjects.

4 Acoustic distance measurements

As a basis for our acoustic measurements we used the 15 recordings of the fable 'The North Wind and the Sun' corresponding with the 15 Norwegian varieties. The recordings of the varieties of Larvik, Bø, Herøy and Bodø are pronounced by male speakers, the other recordings are pronounced by female speakers. We return to this issue at the end of Section 4.1.1. Since our acoustic distance metric is word-based, each text is split in separate word samples. In Section 2 we wrote that each text usually consists of 58 different words, but due to the free translation of some phrases for certain varieties a few of the expected words were missing. For all 15 varieties each of the (nearly) 58 words were cut from the text, so we usually got 58 word samples per variety. If the same word appears more than once in a text, we selected only the first occurrence.

The methodology we use for the comparison of word samples is strongly related to the methodology presented by Heeringa and Gooskens (2003). We describe the methodology we used below. In cases where our model differs from the model of Heeringa and Gooskens (2003), we will make a remark about that.

4.1 Representations

In Section 2.1 we found ten linguistic phenomena which are known to be important distinct markers in Norwegian dialectology on the one hand and represented in our text 'the North Wind and the Sun' on the other hand. We need to choose an acoustic representation in which the variation due to all of these phenomena is clearly represented. On the other hand, we want to minimize or – if possible – to eliminate the influence of variation in voice quality.

Heeringa and Gooskens (2003) examined three representations: Barkfilter spectrograms, cochleagrams and formant tracks. Highest correlations with the perceptual distances were obtained when using the formant track representation. The acoustically based results were also correlated with measurements based on the phonetic transcriptions. Again the formant track representation had the highest correlation, even significantly higher than the correlations of the two other acoustic representations. The authors write that "this outcome may indicate that the influence of voice characteristics is less strong when distances are measured on the basis of formants, rather than on the basis of the Barkfilter or cochleagrams." Therefore we will consider the formant track representation in this paper.

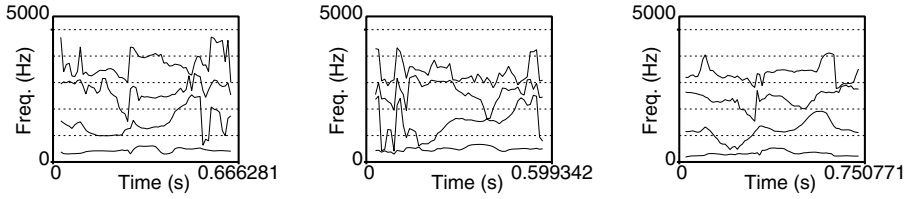
In addition we also consider zero crossing rates. The zero crossing rate is sensitive to the difference between voiced and unvoiced speech sections. High zero crossing rates indicate noise, i.e. frication and low values are found in periodic, i.e. sonorant parts of speech (Frankel et al., 2000).

In Sections 4.1.1 and 4.1.2 the two acoustic representations are described in more detail. In Sections 4.1.3 we will answer the question to what extent the representations represent variation due to the ten phenomena mentioned in Section 2.1.

4.1.1 Formant tracks

Formants are measured in PRAAT[3] with Burg's algorithm. This algorithm may initially find formants at very low or high frequencies. However we used the version in PRAAT which removes formants below 50 Hz and formants above the maximum formant frequency minus 50 Hz. Burg's algorithm is much more reliable than the Split Levinson algorithm which always finds the requested number of formants in every frame, even if they do not exist.

[3] The program PRAAT is a free program and available via `http://www.fon.hum.uva.nl/praat/` .



Bjugn	Halden	Larvik
[²nuːɽɑˌʊjɲˑ]	[²nuːrɑˌʊinː]	[²nuɽɑˌʊinˑɳ]

Fig. 6. Formant track representations of three Norwegian pronunciations of *nordavinden* 'the northwind'.

The number of formants may vary over time in a word and per word. In the PRAAT program, we maintained the default value for the maximum number of formants which may be found: 5. Next, we found the minimum number of formants examining all points in time of all words which are taken into consideration. After that, on the basis of this minimum number of formants the word samples were compared. In all of our material at least three formants could be found in each time slice. Therefore, the comparison of word samples here is based on (the first) three formant tracks.

When finding formants using the computer program PRAAT, the time step was set to 0.01 seconds with an analysis window of 0.025 seconds. The ceiling of the formant search range was set to 5250.[4] Pre-emphasis starts at 50 Hz, which flattens the spectrum. This aids the algorithm to recognize local peaks as formants rather than the global spectral slope.

In Figure 6 we show visualizations of three Norwegian pronunciations of the word *nordavinden* 'the northwind' using formant tracks. The pronunciations of the varieties of Bjugn, Halden and Larvik are given.

The PRAAT program gives formant frequencies in Hertz. We also consider frequencies in Bark, which may be a more faithful scale perceptually. For this purpose we used the formula of Traunmüller (1990) as suggested in standard works about phonetics (Rietveld and Van Heuven, 1997, e.g.):

$$Bark = \frac{26.81 \times Hertz}{1960 + Hertz} - 0.53 \tag{1}$$

Furthermore we experimented with an approach in which z-scores instead of either the Hertz or the Bark frequency values are used. Per frame we calculated the mean and the standard deviation. Next within frame f a z-score is

[4] The authors of the PRAAT program advise to set the ceiling to 5000 Hz for males, and to 5500 Hz for females. We obtained the best results by using the average for both males and females.

calculated for each frequency f_i:

$$f_i = \frac{f_i - f_{mean}}{f_{standard\ deviation}} \tag{2}$$

The idea behind frequency normalization within a frame is that the relative positions of the F1, F2 and F3 to each other are more important than the absolute values of the three formants. Since pitch influences formants, the disadvantage of this approach is that toneme variation is no longer represented by the formants. The advantage is that speaker-dependent intonation variation is normalized. The (average) pitch per speaker may be different, especially as the result of gender differences. In addition pitch contours will differ when speakers put accents on words differently. With our normalization procedure the influence of pitch (contour) variation is neutralized.

4.1.2 Zero crossing rates

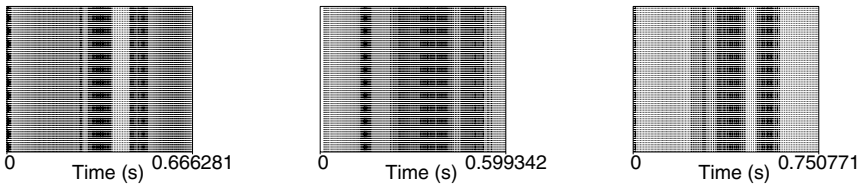
The number of times per interval that the amplitude waveform crosses the zero line is called the zero crossing rate. Zero crossing risers are the points in time when the waveform changes from negative to positive, and fallers represent the times when the amplitude goes down from positive to negative.

PRAAT offers a function which gives us the points in time of the risers or fallers or both risers and fallers. We used the default setting: risers only. However, when using fallers or risers and fallers, nearly the same results are obtained. We converted the zero crossing times to zero crossing rates using a time step of 0.01 seconds, the same as used in the formant analysis. The analysis window was set to a different size: 0.05 seconds. A larger analysis window gives more fluent estimations, but the size of our analysis window is just smaller than the shortest word sample.

In Figure 7 the zero crossing distributions are shown for three Norwegian pronunciations of the word *nordavinden* 'the northwind'. Again the pronunciations of the varieties of Bjugn, Halden and Larvik are given.

4.1.3 The representation of linguistic phenomena

In Section 2.1 we gave an overview of ten features which have contributed to Skjekkeland's dialect classification and which are represented in our data. In the Sections 4.1.1 and 4.1.2 we discussed two acoustic representations: formant tracks and zero crossings. Will variation due to the ten features be processed when using these two acoustic representations? We will answer these questions by briefly discussing the features again.



Bjugn	Halden	Larvik
[²nuːɽɑˌʊjɲˑ]	[²nuːɾɑˌʊinː]	[²nuɽɑˌʊinˑɳ]

Fig. 7. Zero crossing distributions of three Norwegian pronunciations of *nordavinden* 'the northwind'. White vertical lines across the black horizontal lines represent the times of the zero crossing risers.

Apocopation of verb endings. In terms of the Levenshtein algorithm apocopation is the deletion of one or more segments. It will be processed by both the formant track representation and the zero crossing representation. The cost of a deletion (and an insertion) will be lower as the formant frequencies and the number zero crossings of the segment are lower.

Toneme 1 intonations. Heeringa and Gooskens (2003) did not process pitch and intonation contours, although they are known to be significant dialect markers in Norwegian. The authors were in a dilemma. On the one hand the pitch represents dialect variation, on the other hand it represents speaker variation – especially gender differences – which is not relevant in the classification of dialects. Therefore they used monotonized versions of the samples. In this paper we will experiment with both original non-manipulated samples and normalized samples. We already described the normalization procedure at the end of Section 4.1.1. When using original samples, intonation variation is processed to the extent in which pitch differences influence formant tracks. Useful future work may be to include pitch as a third acoustic feature. When normalized samples are used, the influence of intonation is eliminated. The normalization procedure is described in Section 4.1.1.

Palatalization of alveolar consonants. Formants are especially useful for the identification of vowels as can be seen in the IPA quadrilateral, where height corresponds with the first formant and advancement with the second formant (Rietveld and Van Heuven, 1997, p. 133). Formant frequency differences might also reflect variation in stable voiced consonants, i.e. all voiced consonants except for the plosives. This can be seen in the spectrograms, shown as 'The ABC's of Visible Speech' by Potter et al. (1947, p. 54–56). We may be confident that the non-palatal [n] and the palatal [ɲ] will be distinguished by different formant tracks.

Pronunciation of the /r/. In the data we use we find a distinction between the velar approximant [ɰ] and the alveolar tap [ɾ] in our data. The [ɰ] will be represented well by formant tracks. For the [ɾ] this might be less clear. Potter et al. (1947, p. 54–56) extensively discuss the way in which the /r/ influences vowel patterns. The effect of the /r/ is seen in the formant tracks of surrounding vowels. We suppose this also holds for the [ɾ].

Voicing of voiceless plosives. Here the zero crossing rates representation is useful, since this representation distinguishes between voiced and unvoiced speech sections.

Assimilation of /d/ before an alveolar sonorant. In Section 2.1 we found that the consonant cluster [nd] changed into [nn] or [n] in most dialects (related to map 10). The plosive [d] could not be represented quite well by formant tracks, but similar as for the /r/ variation, the effect of the absence or presence of the [d] might be found in the surrounding segments, in our case the preceding [n] or the following vowel.

Change of initial /hv/. The word *kven* may have initial [ʋ] and [k^h]. The [ʋ] is a stable voiced consonant which has a clear formant structure, and the [k^h] is a plosive, affecting the following vowel. Therefore, we expect that the two different segments will be distinguished by the formant representation. Besides, the [ʋ] is voiced and the [k^h] is voiceless. The distinction will be measured when we use the zero crossing representation.

Endings of definite nouns. In Section 2.1 we wrote that we found a similar phenomenon in *mannen* and *frakken*. The pronunciation of *mannen* usually ends on [ɲː] or [nː], but in one case its ends on [ɲːɪ]. The Levenshtein algorithm will process this difference as the insertion or deletion of the [ɪ]. The pronunciation of *frakken* usually ends on [ən], but in one case it ends on [ɪŋ]. The difference between [ə] and [ɪ] will be well processed with the formant track representation. In another case the ending is [ɪ]. The deletion of the [ɲ] will be accurately processed by the Levensthein algorithm.

Lexical variation. In Section 2.1 we found that lexical variation is represented in a relatively large number of items. Differing lexemems have different word structures. Word structure differences will be processed when using the formant track representation (see Figure 8). But looking at Figure 9 we expect that zero crossings will provide additional information. The figure shows

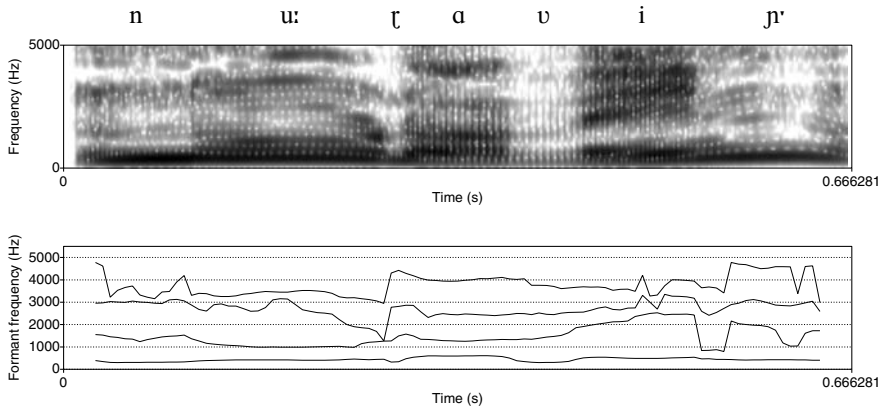


Fig. 8. Spectrogram and formant tracks of the pronunciation of *nordavinden* 'the northwind' in the variety of Bjugn, which is pronounced as [^{2}nuːɽɑˌʊiɲˈ].

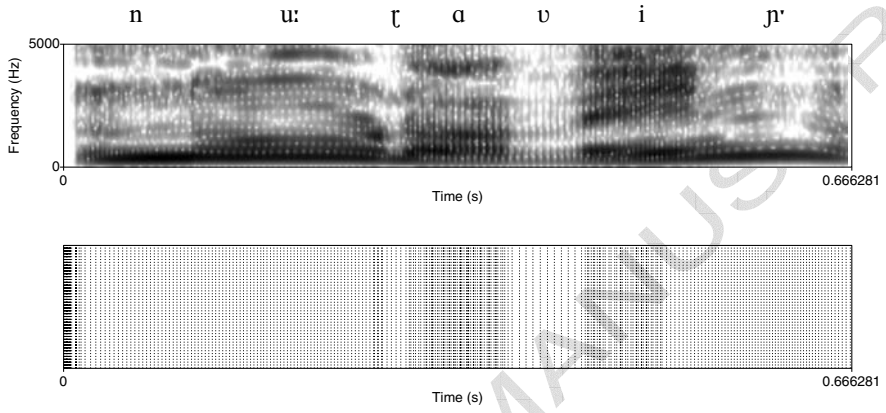


Fig. 9. Spectrogram and zero crossing distribution of the pronunciation of *nordavinden* 'the northwind' in the variety of Bjugn, which is pronounced as [^{2}nuːɽɑˌʊiɲˈ]. In the latter picture white vertical lines across the black horizontal lines represent the times of the zero crossing risers.

that especially the [ʊ], the [i] and the [ɲ] are clearly distinguished by the zero crossing representation.

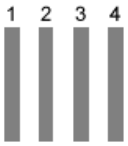
Conjugation of verbs. The past tense of the third person singular is in some dialect varieties realized by adding the ending [tə]. If this is not the case the central vowel is changed in most of the verbs. The addition of the postfix [tə] will be counted as two insertions by the Levenshtein algorithm. The change of the central vowel is measured when we used the formant track representation.

4.2 Speech rate normalization

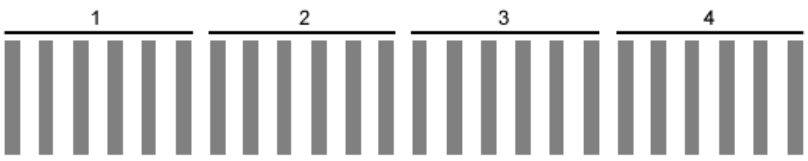
Different samples sizes may reflect dialect variation, but can also be the result of different speech rates. Therefore we had to normalize over speech rate. Heeringa and Gooskens (2003) normalized over the number of segments of a sample according to the transcription. We describe this transcription based approach in more detail in Section 4.2.1. Since our goal was to develop a fully transcription-independent methodology, we also consider another normalization procedure where the samples of a word pair are stretched so that they get the same number of frames. That transcription-independent approach is discussed in Section 4.2.2.

4.2.1 Transcription based

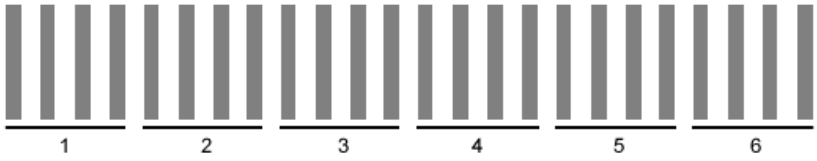
Assume that the acoustic representation of a word sample consists of l frames. If the number of segments of this word pronunciation according to the phonetic transcription is m, and we want to represent each segment by n frames, then we represent the complete word sample by $m \times n$ frames. Changing the representation of l frames into a representation of $m \times n$ frames is realized in two steps. First we duplicate each of the l frames $m \times n$ times. This gives $l \times m \times n$ frames in total. Second we regard the $l \times m \times n$ frames as $m \times n$ groups, each consisting of l frames, and fuse the frames in each group to one frame by averaging them. The result is a representation of $m \times n$ frames. We illustrate this by an example. Assume we have a word sample of $l = 4$ frames:



If this word pronunciation is transcribed as a sequence of $m = 2$ segments, and we want to represent each segment by $n = 3$ frames, then we represent the complete word sample by $2 \times 3 = 6$ frames. We change the representation of 4 frames into a representation of 6 frames. For this purpose first we duplicate each of the 4 frames 6 times. This gives 24 frames in total:



Second we treat the 24 frames as 6 groups, each consisting of 4 frames:

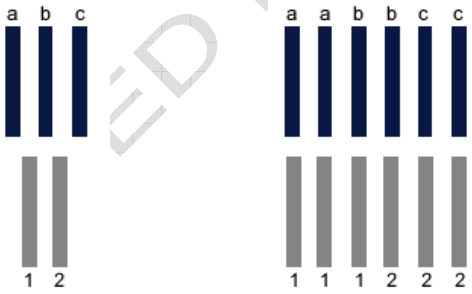


We fuse the frames in each group to one frame by averaging them. The result is a representation of 6 frames:

In our research we chose $n = 20$, i.e. 20 frames per segment. A higher value gives nearly the same results, but the computing time increases greatly.

4.2.2 Transcription-independent

When comparing one segment of m frames with another segment of n frames, each of the m frames is duplicated n times, and each of the n frames is duplicated m times. So both segments get a length of $m \times n$. Below two segments are schematically visualized, one with 3 frames (black bars) and one with 2 elements (grey bars). Now both get a length of 6 when each of the 3 frames are duplicated 2 times, and each of the 2 frames are duplicated 3 times.



4.3 Comparison of frames

Formant tracks When using the formant track representation, a sample is represented as a series of frames, each frame having three formant frequency values. When comparing a frame of a word pronunciation of one dialect variety with the corresponding frame of the corresponding word pronunciation of

another variety, the distance is calculated as:

$$d(f_1, f_2) = \sum_{i=1}^{n} \mid f_{1_i} - f_{2_i} \mid \tag{3}$$

where $n = 3$.

The distance measure we used is known as the *Manhattan* distance. Heeringa and Gooskens (2003) used the *Euclidean* distance: the square root of the sum of the squared differences. Since we found the best results with the Manhattan distance, this measure will be used throughout this paper.

A frame in one sample does not always correspond with another frame in the second sample. Frames can be inserted or deleted (see Section 4.4). In these cases frames are compared to a 'silence frame'. A 'silence formant frame' is defined as a frame for which all frequencies are equal to 0 Hertz or -0.53 Bark (see Section 4.1.1). This means that in absolute silence there are no vibrations. When using z-scores instead of the original Hertz or Bark values, the values are still set to 0.

Zero crossing rates When using zero crossing rates, frames consist of only one value. The distance between two frames is equal to the absolute difference of the two zero crossing rates. The value in a 'silence zero crossing rate frame' is set to 0: there are no zero crossings during silence.

Combined representation When combining formant frame distances with the corresponding zero crossing rate distances, the two distances are multiplied:

$$d(f_1, f_2) = \left(\sum_{i=1}^{n} \mid formant_{1_i} - formant_{2_i} \mid \right) \times \mid zero_1 - zero_2 \mid \tag{4}$$

where $n = 3$.

4.4 Levenshtein distance

In order to measure the degree to which two pronunciations differ, we use Levenshtein distance. As mentioned in Section 1 the algorithm can be used on the basis of phonetic transcriptions, but in this paper we show its application to acoustic representations, similar as was done by Heeringa and Gooskens (2003).

The Levenshtein algorithm computes the distance between two words. The pronunciation of a word in the first variety is compared with the pronunciation of the same word in the second. The algorithm determines how one pronunciation is changed into the other by inserting, deleting or substituting elements. Weights are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost, e.g. 1. A detailed description is given by Kruskal (1999). We illustrate the algorithm by an example in which transcriptions of two word pronunciations are compared to each other. Assume *gåande* or *gående* 'going' is pronounced as [goːɑns] in the variety of Bø and as [gɔːnə] in the variety of Lillehammer. Changing one pronunciation into the other can be done as follows (ignoring suprasegmentals and diacritics):

goːɑns	substitute o/ɔ	1
gɔːɑns	delete ɑ	1
gɔːns	insert ə	1
gɔːnəs	delete s	1
gɔːnə		
		4

In fact many sequence operations map [goːɑns] to [gɔːnə]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. Comparing pronunciations in this way, the distance between longer words will generally be greater than the distance between shorter words. The longer the words, the greater the chance for differences with respect to the corresponding word in another variety. Because this does not accord with the idea that words are linguistic units, the sum of the operations is divided by the length of the longest alignment which gives the minimum cost. The longest alignment has the greatest number of matches. In our example we get the following alignment:

g	oː	ɑ	n		s
g	ɔː		n	ə	
	1	1		1	1

In this paper, Levenshtein distance was applied to acoustic samples instead of phonetic transcriptions. Instead of phonetic segments, acoustic frames were aligned. In our example all operations have a weight of 1. However, when comparing acoustic samples, substitutions, insertions and deletions have gradual weights, calculated in the way as described in Section 4.3. Levenshtein distance was used in the same way by Heeringa and Gooskens (2003).

Using 58 words the distance between two varieties is equal to the average of 58 Levenshtein distances. When comparing two varieties on the basis of k word pairs, it may appear that for one or more of the pairs for one or both varieties, no sample is available. This can be the result of the fact that no translation is given. In these cases, the word pair was ignored.

All distances between the 15 varieties were arranged in a 15×15 matrix.

5 Validation

In this section we validate our computational results with the results of the perception experiment. For this purpose we correlate the computational distances with the perceptual distances (Section 5.1). We distinguish two types of measurements: semi-acoustic word sample-based measurements (Section 5.2) and (fully) acoustic word sample-based measurements (Section 5.3). We end with some conclusions (Section 5.4).

5.1 Correlation

In order to correlate the different computational measurements to the results of the perception experiment, the computational 15×15 matrices are correlated with the perceptual 15×15 matrix. When correlating we exclude the distances of varieties with respect to themselves, i.e. the distance of Bergen to Bergen, of Bjugn to Bjugn etc. These distances are found on the diagonal in the distance matrix, containing the cells $(1, 1), (2, 2), \ldots (n, n)$. In computational matrices these values are always 0, in the perceptual matrix they vary, usually being higher than the minimum score. This may be the result of the fact that for example the variety of the speaker of Bergen is different from the variety of the listeners in the same location. Since this causes uni-directional distortion for the diagonal distances (they only can be too high, not too low), we exclude them when calculating the correlation coefficient.

For finding the correlation coefficient, we used the Pearson's correlation coefficient (Sneath and Sokal, 1973, pp. 137–140). For finding the significance of a correlation coefficient we used the Mantel test. In classical tests the assumption is made that the objects which are correlated are independent. However, values in distance matrices are usually correlated in some way, and not independent (Bonnet and Van de Peer, 2002). A widely used method to account for distance correlations is the Mantel test (Mantel, 1967). As significance level we choose $\alpha = 0.05$. With the Mantel test it is also possible to determine whether one correlation coefficient is significantly higher than another.

5.2 Semi-acoustic word measurements

Table 2 shows correlation coefficients between perceptual distances and different semi-acoustic distance measurements. The measurements are made on the basis of words. The speech rate is normalized by counting the number of segments in the transcriptions (see Section 4.2.1). Correlations are given for the complete data set of 15 varieties. Since the mean vocal tract dimensions of males differ from those of females, gender differences may influence our results. Therefore we also show correlations on the basis of a subset of 11 varieties. The recordings of these varieties are pronounced by female speakers. The varieties of Bø, Bodø, Herøy and Larvik are pronounced by male speakers and excluded in the smaller set.

In the table we find three acoustic representations: formant tracks, zero crossing rates, and a combined representation where both formant tracks and zero crossing rates are used (see Sections 4.1 and 4.3). When formant tracks are used, we consider the Hertz scale and the Bark scale (see Section 4.1.1). Besides measurements on the basis of the original Hertz and Bark frequencies, also measurements are given where the frequencies are normalized per frame (see Section 4.1.1).

For each of the measurements we checked whether the 58 words are a sufficient basis for reliable analyses and calculated Cronbach's α values for each of them. Most of them were equal to or higher than the threshold of 0.70, but seven of them were lower, varying from 0.62 to 0.66. For these measurements the correlations are given in normal type setting, the other ones are printed in bold.

In the table we find that both the formant track representation and the combined representation have mostly higher correlations than the zero crossing rate representation. The combined representation gives an improvement only in comparison with the formant track representation when normalized frequency values are used. Considering the frequency scale we find that the Bark scale gives the best results, but when frequencies are normalized the Hertz scale gives the best results. Frequency normalization improves results when the combined representation is used, but not when using formant tracks only. [5]

[5] In Section 2 we did not find vowel variation to be a significant linguistic feature represented frequently in our data. But since formant tracks especially represent vowel variation very well, we felt tempted to perform an analysis on the basis of 34 vowels only. The results showed the same tendencies, but the use of the Bark scale also improves the results when frequencies are normalized. The correlation coefficients were much lower, varying from 0.21 to 0.35 (15 varieties) and from 0.16 to 0.33 (11 varieties). Since these results do not add much, they are not discussed further.

formant bundles	zero crossings	original frequencies		normalized frequencies	
		15 dial.	11 dial.	15 dial.	11 dial.
Hertz	no	**0.50**	**0.58**	**0.50**	**0.55**
Bark	no	**0.53**	**0.58**	**0.48**	**0.55**
Hertz	yes	0.42	0.53	**0.51**	**0.62**
Bark	yes	0.45	0.55	**0.50**	**0.60**
no	yes	**0.38**	0.52	**0.38**	0.52

Table 2
Correlation between perceptual distances and different semi-acoustic distance measurements based on words. Correlations are given for 15×15 and 11×11 matrices excluding the diagonals. All correlations are significant with P-value < 0.001. Correlations in bold are based on measurements with Cronbach's $\alpha \geq 0.70$.

Considering the three factors (representation, frequency scale and normalization) we did not find significant differences in most cases. For 15 varieties the highest correlation is obtained when using the formant track representation and original Bark frequencies. For 11 varieties the highest correlation coefficient is found when the combined representation is used and normalized Hertz frequencies are used. Using Bark frequencies gives the second best correlation.

5.3 Acoustic word measurements

Table 3 shows correlation coefficients between perceptual distances and different acoustic distance measurements. The measurements are made on the basis of entire words. When two word samples are compared, they are stretched so that they get the same number of frames before the distance between them is calculated (see Section 4.2.2). As for the semi-acoustic word-based comparisons, correlations are given for all 15 varieties and for the 11 varieties, represented by female speakers.

As in Section 5.2 for each of the measurements we checked whether the 58 words are a sufficient basis for reliable analyses and calculated Cronbach's α values for each of them. We found most of them to be equal to or higher than the threshold of 0.70, but eight of them were lower, varying from 0.55 to 0.69. For these measurements the correlations are given in normal type setting, the other ones are printed in bold.

Looking at the table, we find that both the zero crossing rate representation and the combined representation have higher correlation coefficients than the

formant bundles	zero crossings	original frequencies		normalized frequencies	
		15 dial.	11 dial.	15 dial.	11 dial.
Hertz	no	**0.27**	**0.38**	**0.34**	**0.39**
Bark	no	**0.28**	**0.38**	**0.31**	**0.38**
Hertz	yes	0.41	0.52	**0.48**	**0.58**
Bark	yes	0.42	0.53	**0.48**	**0.57**
no	yes	0.39	0.52	0.39	0.52

Table 3
Correlations between perceptual distances and different acoustic distance measurements based on words. Correlations are given for 15×15 and 11×11 matrices excluding the diagonals. All correlations are significant with P-value < 0.001. Correlations in bold are based on measurements with Cronbach's $\alpha \geq 0.70$.

formant track representation. The combined representation gives better results than the zero crossing rate representation. Considering the frequency scale, again our findings accords with those of the semi-acoustic measurements: the Bark scale gives the best results when original frequencies are used, but when frequencies are normalized the Hertz scale gives the best results. Frequency normalization improves results for both the formant track representations and the combined representations. For the semi-acoustic measurements we found improvements only for the combined representations. The highest correlation is obtained when using the combined representation and normalized Hertz frequencies, followed by the version with the Bark frequencies. For 15 varieties we found the correlations of the two measures to be significantly higher than those using the formant track representation and non-normalized frequencies (0.48 versus 0.27, P-value=0.01, 0.48 versus 0.28, P-value=0.01). For 11 varieties they are nearly significantly higher (0.58 versus 0.38, P-value=0.06, 0.57 versus 0.38, P-value=0.08). The two versions also had the highest correlations for the semi-acoustic measurements based on the 11 varieties, represented by female speakers.

5.4 Conclusions

Representation The semi-acoustic measurements match the perceptual results better than the (fully) acoustic measurements in all cases execpt one. For the zero crossing rates we did not find a clear difference: 0.38 versus 0.39 (15 varieties) and 0.52 versus 0.52 (11 varieties). This gives us the impression that zero crossing rate measurements are quite robust in the sense that they

are speech rate normalization procedure-independent. In Section 4.1 we suggested that zero crossing distributions represent the segmental structure to some extent. This may explain why combined measurements are better than formant track measurements for all cases of the acoustic measurements, but for only half of the cases of the semi-acoustic measurements. In case of the semi-acoustic measurements, segmental information is already read from the transcriptions, and therefore, the segmental information of the zero crossing distribution may partly be superfluous.

Frequency scale Looking at the word-based measurements we find the tendency that the Bark scale gives higher correlations than the Hertz scale when the original, non-normalized frequency values are used. When normalized frequency values are used, in a few cases the Hertz measurements are a bit higher than the Bark measurements. Therefore the use of the Bark scale is only useful when non-normalized frequency values are used.

Frequency normalization The use of normalized frequency values yields an improvement for the combined representation only when measurements are obtained on the basis of semi-acoustic word sample measurements. For acoustic word sample measurements normalization leads to improvement for both the formant track representation and the combined representation. At the end of Section 4.1.1 we wrote that the idea behind frequency normalization within a frame is that the relative positions of the F1, F2 and F3 to each other are more important than the absolute values of the three formants. Since the pitch influences formants, our normalization procedure eliminates toneme variation which is represented in the formant tracks. However, the gain is that speaker-dependent intonation variation is eliminated as well, which probably explains why results get improved when using this normalization procedure.

Our choice For all measurements, both semi-acoustic word-based and acoustic word-based, we found that the same two measurements outperform the other ones, namely the versions using the combined representation and normalized frequency values. To decide about the frequency scale we look at the very small difference for the 11 varieties, represented by female speakers, where the Hertz scale just gives higher results. So we choose the version which uses the Hertz scale.

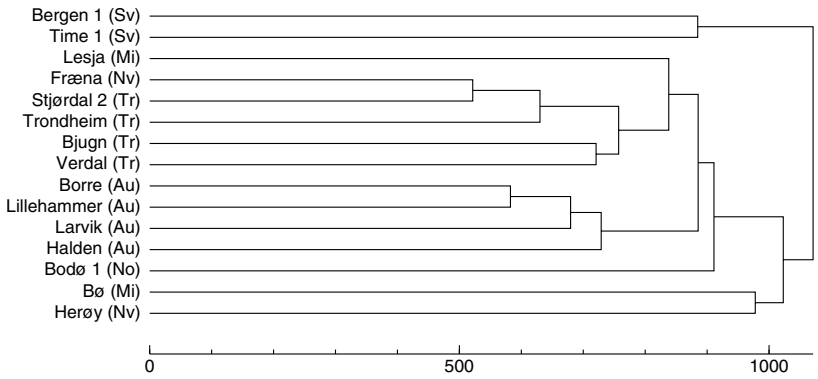


Fig. 10. Dendrogram obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The tree structure explains 45.1% of the variance.

6 Results

In Section 5 we validated several computational methods by comparing their results to the results of the perception experiment described in Section 3. In this paper we are especially looking for a comparison method which does not use any information from transcriptions. Among the category of transcription-independent comparison methods the best one appears to be the version with the combined representation (formant tracks and zero crossing rates) and normalized Hertz frequencies. For all 15 varieties, its results correlate with $r = 0.48$ to perception, and for 11 varieties we found $r = 0.58$. On the basis of the distances obtained with this method, the varieties are classified. As in Section 3.2 for the perceptual distances, we perform cluster analysis and multidimensional scaling. We will compare our results to both the traditional results and the results of the perception experiment.

6.1 Comparison of classifications

In Figure 10 the dendrogram is shown. When we compare this dendrogram with the ones in Figures 2 and 4, we find in all three figures that all the Trøndsk, the Austlandsk an the Sørvestlandsk varieties are clustered together. The dendrograms disagree about the the Midlandsk varieties and the Nordvestlandsk varieties. The disagreements about Bø, Herøy and Bodø may have to do with the fact that their recordings were pronounced by male speakers. However the position of Larvik, which recording was also pronounced by a male, is not so deviant in comparison with the perceptual dendrogram. Some lexical similarity between Bø and Herøy is found in the word *trakk* 'drew'. These two varieties have something like [dɾuːg], while all others have something like [tɾɑkː].

We also applied multidimensional scaling. The resulting plot is shown in Figure 11. This two dimensional plot explains 74.3% of the variance of the original computational distances. When comparing this plot, the traditional plot in Figure 3 and the perceptual plot in Figure 5 with each other, we find that all of them show the same pattern globally: Sørvestlandsk varieties and the variety of Herøy are on the left, the Trøndsk varieties on top and the Austlandsk varieties in the lower right corner. The groups in the perceptual plot are more sharply distinguished than in the computational plot, but the groups in the traditional plot are more sharply distinguished than in the perceptual plot. When we enter into more detail, we find that Lesja belongs to the Trøndsk varieties in the traditional plot, but in the two other plots Lesja is much closer to the Austlandsk varieties, which is – geographically seen – more reasonable. Fræna and Herøy are one group in the traditional plot, but in the other plots Fræna is much closer to the Trøndsk varieties.

In both the traditional and perceptual plot, the variety of Bø is found among the Austlandsk varieties, but in the computational plot Bø is located on top of the Trøndsk varieties. We expected Larvik close to Halden in the computational plot, but the variety is found much higher and more distant from the variety of Halden. The higher positions of Bø and Larvik may be explained by the fact that these varieties are, just as Bodø and Herøy, pronounced by male speakers. This gives us the impression that the first dimension, represented by the vertical axis in the plot, represents gender to some extent. Despite the gender unbalance in the data which may bias the multidimensional scaling projection, we are still convinced that gender significantly affects the first dimension.

6.2 Filtering away the influence of gender

In Section 6.1 we found that the first dimension of the multidimensional scaling solution probably represents gender to some extent. Therefore, it may be better to ignore the first dimension, and examine the second and higher dimensions. In order to find out whether higher dimensions may be interesting, we scaled our computational distances to the largest possible number of dimensions allowed by the R program: 12. Next we calculated distances between the 15 varieties per dimension, resulting in 12 distance matrices. Next we correlated each of the matrices with the perceptual distance matrix. We squared the correlation coefficients and multiplied them by 100. In this way for each dimension we got a percentage which represents the amount of variance which that dimension explains of the perceptual distances. The variances are shown in Figure 12. Variances which are the result of squared negative correlations, are still visualized as negative values. The figure suggests that especially the first, second and third dimension are important.

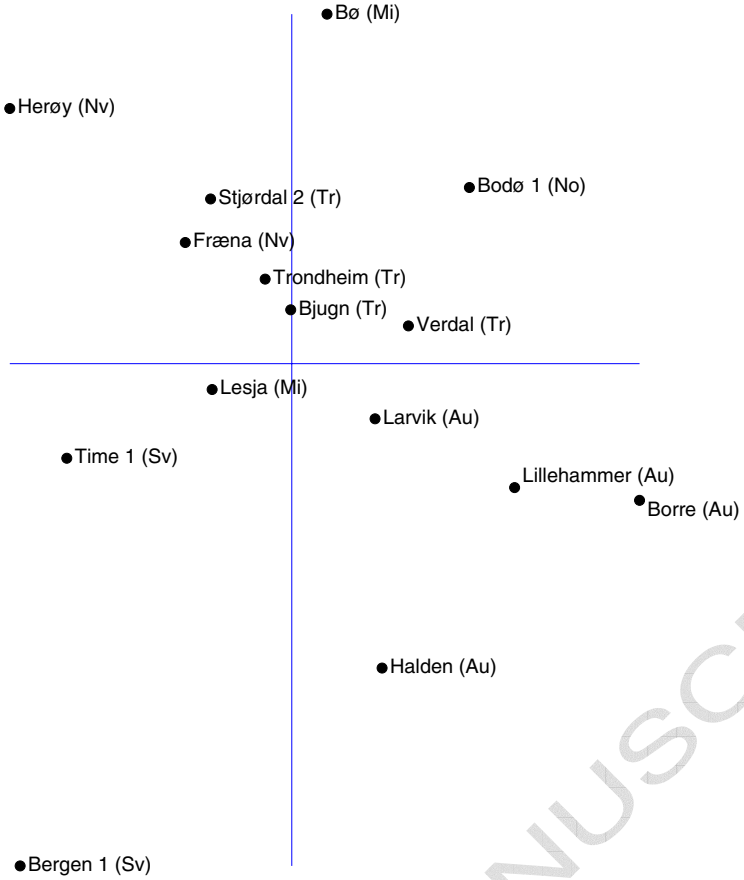


Fig. 11. Multidimensional scaling plot obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The vertical axis represents the first dimension, and the horizontal axis the second dimension. The two dimensions explain 76.1% of the variance. The varieties of Bø, Bodø, Herøy and Larvik were represented by male speakers.

We found that the first dimension distances correlate (nearly) significantly more strongly with the perceptual distances than the fourth and higher dimension distances do (highest P-value was equal to 0.14). The same applies for the second dimension distances (highest P-value was equal to 0.12) and the third dimension distances (highest P-value was equal to 0.07). Therefore we focus on the first, second and third dimension.

In the results of the perceptual measurements we found no influence of gender differences (see Figures 4 and 5). Nevertheless we found a relatively high variance for the first dimension. Therefore we cannot conclude that the first dimension just represents voice quality. However, we want to exclude this dimension since it is the only way to exclude the influence of gender differences. Therefore we scaled the computational dimensions to three dimensions, and drew a plot on the basis of the second and third dimension. The plot is shown in Figure 13. The three dimensions explain 79.6% of the variance of the original computational distances, and the second and third dimension together

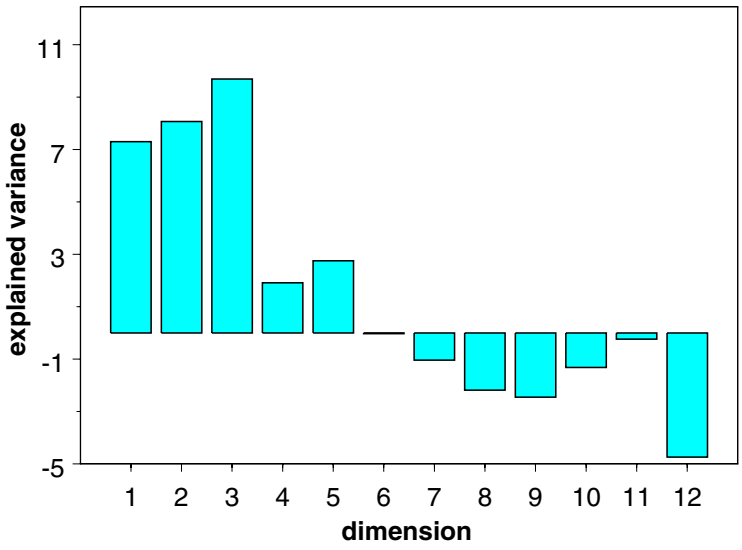


Fig. 12. The computational distances are scaled to 12 dimensions. For each dimension a bar shows how much perceptual variance that dimension explains. The variances are given in percentages. Variances which are the result of squared negative correlations, are still visualized as negative values.

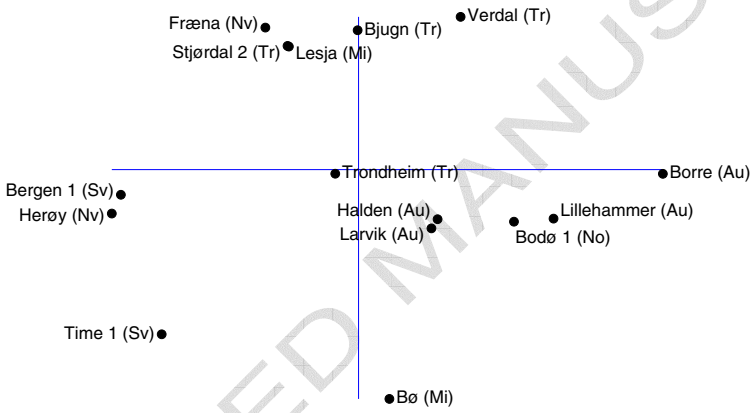


Fig. 13. Multidimensional scaling plot obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The horizontal axis represents the second dimension, and the vertical axis the third dimension. The two dimensions explain 31.3% of the variance.

explain 31.3%.

Different from Figure 11 and more similar to Figures 3 and 5 is that the different groups are distinguished more sharply in Figure 13. The global pattern is the same again: The Sørvestlandsk varieties and Herøy on the left, the Trøndsk varieties on top and the Austlandsk varieties in the lower right corner. Larvik is close to Halden, just as in the traditional and perceptual plot. Bø is south of the Austlandsk varieties which is better than north of the Trøndsk varieties as in Figure 11. Trondheim is close to the Austlandsk varieties in our new plot which agrees with the traditional plot, but disagrees with the perceptual

plot. This may indicate that linguistic features are weighed differently in the traditional and perceptual plot. Lesja becomes more distant to the Austlandsk varieties and is still close to the Trøndsk varieties. Again this agrees with the traditional plot and disagrees with the perceptual plot. A bit unexpected may be the position of Bodø close to the Austlandsk varieties. This may be explained by the fact that Bodø shares the conjugation of the past tense of the third person singular with the southern varieties (see 2), and for four lexically varying words it has the same lexeme as the Austlandsk and Sørvestlandsk varieties have.

6.3 The representation of the linguistic features

In this section we investigate to what extent the three multidimensional scaling dimensions represent the linguistic features we identified in Section 2. In order to do so, we first calculate distances per dimension. We do not consider toneme variation feature, since this information is not represented by the combined normalized formant track/zero crossing representation (see Sections 4.1.1 and 5.4).

In Figure 5 the y-axis represents the first dimension. The distance between two varieties based on the first dimension only is calculated as the absolute difference between the corresponding y-coordinates. In this way, for each pair of varieties the distance is calculated. Similarly distances for the second dimension are based on the x-coordinates. Distances for the third dimensions are based on the y-coordinates as shown in Figure 13.

In Section 4.4 we explained the way in which acoustic distances between dialect varieties are calculated. When we have acoustic samples of the pronunciations of one particular word for each of the 15 varieties, Levenshtein distances between all pairs of varieties are calculated. The results presented in the previous sections are not based on measurements between pronunciations of one single word, but on aggregate distances of 58 words. However, in this section we use the single word distances of each of the 58 words individually. For each word the distances are correlated with the distances based on the first, second and third multidimensional scaling dimension respectively.

When looking at the results of the first dimension, we find that this dimension cannot be associated with one particular feature but rather with the combination of lexical variation, palatalization and the pronunciation of the /r/. Strongly correlating words representing lexical variation are *kjekla* or *kjeklet* 'disputed' (0.67), *til sist* 'at last' (0.54) and *straks* 'immediately' (0.51). Relatively strongly correlating words (partly) representing the palatalization feature are *innrømme* 'admit' (0.45), *mannen* 'the man' (0.36), *mann* 'man'

(0.36), *kunne* 'could' (0.36) and *nordavinden* 'the north wind' (0.33). Relatively strongly correlating words (partly) representing the variation in the pronunciation of the /r/ are *trakk* 'drew' (0.44), *frakken* 'the cloak' (0.41) and *nordavinden* 'the north wind' (0.33).

Among the words most strongly correlating with the distances based on the second dimension, words representing the /r/ pronunciation feature are found most frequently: *frakk* (0.58) 'cloak', *fram* 'out' (0.37) and *frakken* 'the cloak' (0.31). Looking at the correlations of single word distances with the third dimension, words with lexical variation occurs most frequently: *av* 'of' (0.34), *av* 'off' (0.32) and *til sist* 'at last' (0.25).

We may conclude that especially the linguistic phenomena dominantly represented in the data – except for tonemes – are processed with our combined normalized formant track/zero crossings representation. The first multidimensional scaling dimension represents the combination of all of the phenomena, the second one represents the /r/ pronunciation feature and the third one represents lexical variation.

7 Conclusion

The aim of this paper was to show that one can classify the varieties of Norwegian on the basis of acoustic features only, without phonetic transcriptions. Therefore we seek a fully acoustic and transcription-independent measure for finding distances between dialect varieties. Heeringa and Gooskens (2003) presented a semi-acoustic measure and obtained the best results when using formant tracks where frequencies are represented in the Bark scale. Looking in Table 2 we find $r = 0.53$ for all 15 varieties, and $r = 0.58$ for the 11 varieties, represented by female speakers. In Table 3 we find the results for the fully acoustic measures. The best results are obtained when using a combined representation (formant tracks and zero crossing rates) and normalizing formant frequencies per frame: r=0.48 for all 15 varieties and r=0.58 for the 11 varieties. This means that we have found a fully acoustic measure which performs (nearly) as well as the semi-acoustic one proposed by Heeringa and Gooskens (2003).

When classifying the 15 Norwegian varieties on the basis of results obtained with the fully acoustic method, we found that they largely agree with both the traditional and the perceptual classification. In all three classifications the Trøndsk, the Austlandsk an the Sørvestlandsk varieties show up as groups. Both the similarity with the classification based on traditional dialectology criteria and the significant correlation with the results of the perceptual classification experiment confirm our hypothesis that the varieties of Norwegian

can be classified on the basis of acoustic features only, without the use of any information from phonetic transcriptions.

However, gender affected the acoustical results. When using multidimensional scaling, this influence was found in the first dimension. In a plot based on the second and third dimension, the influence of gender is no longer found. However, besides gender-specific variation, the first dimension also represents dialect specific information which is lost when leaving out this dimension. Therefore further research is necessary to filter out the influence of voice quality. Adank et al. (2004) propose different ways of formant frequency normalization which can be examined in future research.

Another issue is speech rate normalization. Our research shows that the results of the semi-acoustic measure are still higher for the complete set of 15 varieties when using formant tracks and formant frequencies represented in the Bark scale (0.53 vs. 0.48). Therefore it may be useful to seek for a procedure which automatically determines the number of phonetic segments on the basis of the acoustic signal. We found that especially the zero crossing distribution represents the segmental structure to some extent, therefore zero crossings can possibly help in finding the number of segments.

The 'winning' method among the fully-acoustic alternatives uses normalized formant tracks. Speaker-dependent variation is normalized in this way. The drawback is that toneme variation is lost as well. Future work may be to find a solution for this, probably by adding pitch as an extra acoustic feature.

Finally results can be improved by using more data. First, some words occur more than once in the text, for example 'the north wind' occurs four times. We only used the first occurrence, but results may be improved when using all occurrences of a word and averaging over them. Second, we use one speaker per variety. Useful future research may be to base results on multiple recordings per variety.

References

Adank, P., Smits, R., and van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116:3099–3107.

Bonnet, E. and Van de Peer, Y. (2002). zt: a software tool for simple and partial Mantel tests. *Journal of Statistical Software*, 7(10):1–12. Available via: `http://www.jstatsoft.org/`.

Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press, Cambridge, 2nd edition.

Christiansen, H. (1954). Hovedinndelingen av norske dialekter. In Beito,

T. B. and Hoff, I., editors, *Frå norsk målføregranskning*, pages 39–48. Universiteitsforlaget, Oslo.

Fintoft, K. and Mjaavatn, P. E. (1980). Tonelagskurver som målmerke. *Maal og Minne*, pages 66–87.

Frankel, J., Richmond, K., King, S., and Taylor, P. (2000). An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing.

Goebl, H. (1982). *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, volume 157 of *Philosophisch-Historische Klasse Denkschriften*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of W.-D. Rase and H. Pudlatz.

Goebl, H. (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In Viereck, W., editor, *Proceedings of the International Congress of Dialectologists*, volume 1, pages 37–81, Stuttgart. Franz Steiner Verlag.

Gooskens, C. and Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3):189–207.

Heeringa, W. (2005). Dialect variation in and around Frisia: classification and relationships. *Us Wurk: Tydskrift foar Frisistyk*, 54(3-4):125–167.

Heeringa, W. and Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, 37(3):293–315.

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen, Groningen.

Hunt, M. J., Lennig, M., and Mermelstein, P. (1999). Use of dynamic programming in a syllable-based continuous speech recognition system. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 163–187. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Yersey.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–67, Dublin. EACL.

Kristoffersen, G. (2000). *The Phonology of Norwegian*. The Phonology of the World's Languages. Oxford University Press, Oxford.

Kruskal, J. B. (1983). An overview of sequence comparison. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 1–40. Addison-Wesley, Massachusetts.

Kruskal, J. B. (1999). An overview of sequence comparison. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 1–44. CSLI, Stanford,

2nd edition. 1st edition appeared in 1983.

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

Nerbonne, J., Heeringa, W., Van den Hout, E., van der Kooi, P., Otten, S., and van de Vis, W. (1996). Phonetic distance between Dutch dialects. In Durieux, G., Daelemans, W., and Gillis, S., editors, *CLIN VI, Papers from the sixth CLIN meeting*, pages 185–202, Antwerp. University of Antwerp, Center for Dutch Language and Speech (UIA).

Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech*. The Bell Telephone Laboratories Series. Van Nostrand, New York.

Rietveld, A. C. M. and Van Heuven, V. J. (1997). *Algemene fonetiek*. Coutinho, Bussum.

Sankoff, D. and Kruskal, J. (1999). *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, 37:1–24.

Skjekkeland, M. (1997). *Dei norske dialektane: tradisjonelle særdrag i jamføring med skriftmåla*. Høyskoleforlaget, Kristiansand.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. A Series of Books in Biology. W. H. Freeman and Company, San Francisco.

Traunmüller (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88:97–100.