Automatic Dialect Identification: A Study of British English

Emmanuel Ferragne and François Pellegrino

Laboratoire Dynamique Du Langage - UMR 5596 CNRS / Université de Lyon, France
{Emmanuel.Ferragne,Francois.Pellegrino}@univ-lyon2.fr

Abstract. This contribution deals with the automatic identification of the dialects of the British Isles. Several methods based on the linguistic study of dialect-specific vowel systems are proposed and compared using the Accents of the British Isles (ABI) corpus. The first method examines differences in diphthongization for the FACE lexical set. Discrimination scores in a two-dialect discrimination task range from chance to ca. 98 % of correct decision depending on the pair of dialects under test. Thanks to the ACCDIST method (developed in [1,2]), the second and third experiments take dialectal differences in the structure of vowel systems into consideration; evaluation is performed on a 13-dialect closed set identification task. Correct identification reaches up to 90 % with two subsets of the ABI corpus (/hVd/ set and read passages). All these experiments rely on a front-end automatic phonetic alignment and are therefore text-dependent. Results and possible improvements are discussed in the light of British dialectology.

Keywords: Automatic dialect identification, accents of English, British Isles, phonetics of vowel systems.

1 Introduction

The specific patterns of pronunciation that are related to speakers' regional origin or social background greatly contribute to the distinctiveness of their voices, and therefore to the variability of speech. Dialect – or rather accent[1] – identification has therefore become an important concern in speech technology. For instance, it has been shown that automatic speech recognition systems can perform tremendously better when the training and the test sets are matched for dialect ([3]). Dialect identification – whether the task be carried out by a computer or a human expert – also has forensic applications ([4,5]) although, as is the case with any other component of somebody's voice, the plasticity issue

[1] The word *accent* quite often refers to foreign-accented speech, and although it is appropriate to designate the pronunciation of dialects, the term *dialect* will be used instead since the present contribution deals exclusively and unambiguously with pronunciation features.

C. Müller (Ed.): Speaker Classification II, LNAI 4441, pp. 243–257, 2007.

(e.g. somebody may alter their accent for sociolinguistic reasons, or in order to deceive) raises daunting challenges for the speech community. Our aim here is to assess to what extent knowledge of the phonetics of dialects can provide an alternative to crude acoustic modelling. A substantial part of this contribution is therefore devoted to some aspects of phonetic vowel variation across dialects. The remainder covers experiments in the automatic classification of the dialects of the British Isles ([6,1,2,7]) with a twofold objective: evaluating classification scores *per se*, and demonstrating how automatic methods can assist researchers in phonetics and dialectology.

2 An Overview of the Dialects of the British Isles

Most of the dialects of the British Isles have been extensively described in the literature; therefore an exhaustive account falls well beyond the scope of this contribution. The reader is advised to consult the following references for thorough information on the phonetic aspect: [8,9,10,11]. However, some features are highlighted in this section because they constitute the necessary background basis for the rest of the discussion. In traditional (areal) dialectology, pronunciation isoglosses, i.e. boundaries demarcating dialects, have commonly been used. The boundaries that delimit differences in vowel systems are of particular interest to us since they are at the heart of the method developed in Experiment 2. By way of example, *gas* does not rhyme with *grass* in the south of England, but it does in the (linguistic) north. Similarly, the vowels of *nut* and *put* are phonologically identical in the north, but a phonemic split caused them to be differentiated in the south. *Good* and *mood* rhyme in Scotland, but not in the rest of the British Isles, while *nurse* and *square* have been reported to have the same vowel in certain speakers from Liverpool and Hull, for example. However, just as surface realization can be affected by sociological factors, vowel systems too may vary within a given location, and speakers sometimes try to "posh up" their accent by adopting the vowel system of a more prestigious variety than their own. This can lead to a phenomenon known as hypercorrection whereby, for instance, a speaker from the north of England (having no distinction between the vowels in *nut* and *put*) tries to imitate a southerner, failing to identify which words should pattern with the southern phonemes of *nut* or *put*, and ends up pronouncing *sugar* with the vowel of *nut* (example taken from [9, page 353]). This may sound trivial, but it has serious consequences on the method we describe in Experiment 2. The question of lexical incidence (roughly speaking: deciding to which phonemic category a vowel token belongs) is indeed crucial here because it suggests extreme caution – and, clearly, expert knowledge – when choosing the key words for creating shibboleth sentences. Suppose a phonetician designs test sentences to elicit the – or the absence of – contrast between *gas* and *grass* or *father* in order to determine whether a speaker is from the north or the south of England. Without prior knowledge of dialectology, he or she may well wrongly infer from the spelling that *mass* patterns with *grass*, or that *gather* rhymes with *father* in southern dialects. Opposing *gather* or *mass*

with *gas*, and therefore failing to identify the correct underlying phonological representation of these words, would lead the phonetician to miss the potential contrast under study. We will return to this question further below. Beside systemic differences, dialect variation is also manifested by different phonetic realizations of the same phoneme; this characteristic also plays an important role in Experiment 2, and it is clearly illustrated in Experiment 1, which focuses on diphthongization.

3 Corpus Description

The material comes from the Accents of the British Isles (ABI) corpus ([12]). The database consists of recordings from 14 geographical areas throughout the British Isles. For each variety of English, 20 speakers on average (equally divided into men and women) participated. In the following experiments, two types of data were used: a list of 19 /hVd/ words spoken 5 times by each speaker, and a read passage, containing approximately 290 word tokens, specifically designed to elicit dialect variation. The recordings took place in quiet rooms (e.g. in public libraries) at the beginning of 2003; the participants spoke through a head-mounted microphone that was connected to a PC via an external sound card. The sound files are mono 16 bit 22050 Hz PCM Windows files. Worthy of mention is the total lack of individual information on the participants (age, occupation, etc.), which precludes the inclusion of highly relevant sociolinguistic factors in the study ([5,13,14]). The dialects and the towns where the corresponding recordings took place are listed in Table 1.

Table 1. Dialects of the ABI Database

LABEL	DIALECT	PLACE
brm	Birmingham	Birmingham
crn	Cornwall	Truro
ean	East Anglia	Lowestoft
eyk	East Yorkshire	Hull
gla	Glasgow	Glasgow
ilo	Inner London	London (Tower Hamlet)
lan	Lancashire	Burnley
lvp	Liverpool	Liverpool
ncl	Newcastle	Newcastle
nwa	North Wales	Denbigh
roi	Republic of Ireland	Dublin
shl	Scottish Highlands	Elgin
sse	Standard Southern English	London
uls	Ulster	Belfast

4 Experiment 1: Diphthongization

4.1 Goal

Diphthongization refers to the stability over time of the formant pattern in a vowel. The concept lies at the phonetic level in that it disregards whether a vowel be phonologically termed a diphthong or not. For example the vowels of FLEECE and GOOSE [2] in Standard British English are often described as monophthongs in manuals for foreign learners, but they are clearly diphthongized. Our aim is to come up with an economical and sufficient set of parameters to describe formant stability and then validate the model with a classifier. For the sake of parsimony, and in order to get rid of part of the individual variation, absolute vowel initial and final formant values are discarded (although they are known to be dialect specific) and only dynamic features are considered. In the first experiment we concentrate on the so-called FACE vowel, which occurs in the corpus in the words *sailor*, *faces*, *today*, *takes*, *same*, *generations*, *way*, *stable*, *unshakable*, *faith*, *later*, *favour*, *great*, *fame*, *Drake*, *sail*, and *make*. We posit for practical reasons that all these words belong to the FACE set. Note however that this may be too much of an assumption, and a more cautious approach is taken in Experiment 3 where we no longer consider lexical sets, but individual words instead. Using formant trajectories (i.e. the formant slopes) as a criterion, the FACE vowel has, roughly speaking, three main realizations in the dialects of the British Isles:

1. a long closing diphthong beginning with an open-mid vowel and gliding towards a close front position, e.g. in the south of England (e.g. Figure 1a);
2. a centring diphthong starting from a mid-close (or even closer) quality and gliding towards schwa in Newcastle (e.g. Figure 1b);
3. a rather short front close-mid monophthong, e.g. in Scotland and some dialects of the north of England (e.g. Figure 1c).

It is hypothesized that the slopes of F1 and F2 will adequately model these three types of vowels.

4.2 Method and Results

A transcription at the phonetic level was generated with forced alignment using the Hidden Markov Model Toolkit (HTK) ([15]). The models had been trained on the WSJCAM corpus[3]. Formant values were estimated with the Praat program ([16]) using the Burg algorithm set with default values. Some formant extraction errors occurred (as confirmed by visual inspection of formant tracks); however, in order to keep the procedure as automatic as possible, no attempt was made to manually get rid of outliers. Then the slopes of F1 and F2 were computed with robust linear regression in Matlab. Knowledge of phonetic variation was taken

[2] These small capitalized key words stand for lexical sets: [9] popularized this practice in the early 80s, and it is still widely used in British English dialectology nowadays.

[3] We are grateful to Mark Huckvale for kindly providing the HMM models.

into account in order to conceptualize the classification problem. Given that one single linguistic variable (i.e. the FACE vowel) does not allow separability between all dialects, the original task with $C = 14$ classes was approached as $C(C-1)/2 = 91$ separate two-class problems. Another reason for building several two-class models, which would be worth exploring, is to gather an optimal - and therefore presumably different - set of parameters for each pair of dialects. In the absence of any *a priori* reason to the contrary, linear separability was assumed and the classification was performed with a single layer neural net implemented with the Netlab toolbox ([17]). The network has two inputs: the slopes of F1 and F2. For each pair of dialects, all the tokens of all speakers except the speaker under test are passed through the network. This cross-validation procedure is adopted because of the very small size of the dataset. The network is trained with 10 iterations of the iterated re-weighted least squares algorithm. Finally the ouput neuron with a logistic activation function makes a binary decision: the test speaker's tokens either belong to the first or the second dialect of the current pair. A correct classification score is therefore computed for each pair of dialects. In order to save space the 91 scores are not reproduced here; instead, the top and bottom ten pairs are shown in Table 2.

The fourth column shows the geographical distance (in km) between towns. Note how, on average, pairs with high classification scores are farther apart than those with low scores. Actually, a rather low but significant correlation exists between discrimination scores and geographical distances for the 91 pairs ($r = .53$, Spearman rank correlation).

4.3 Discussion

This experiment is the most linguistic-oriented one since the correspondence between formant slope values (the input to the model) and the traditional phonetic vowel quadrilateral facilitates phonetic interpretation. In other words, Experiment 1 not only shows that the method works, but also that the results are directly interpretable in phonetic terms. However, one of the flaws lies in that automatic formant estimation is only partially reliable. Besides, the automatic aspect is quite restricted, and the method described here is therefore very unlikely ever to be implemented in real-life applications. It may however prove a useful tool for testing dialectological hypotheses such as the discriminatory power of a given pronunciation trait.

5 Experiment 2: Vowels in hVd Context

5.1 The ACCDIST Method

In Experiment 2, 19 vowels embedded in /h_d/ consonantal contexts were examined. /hVd/ words have often been used in phonetic studies because the acoustic characteristics of vowels are only slightly affected by these consonants, and keeping the same consonantal context rules out coarticulatory differences.

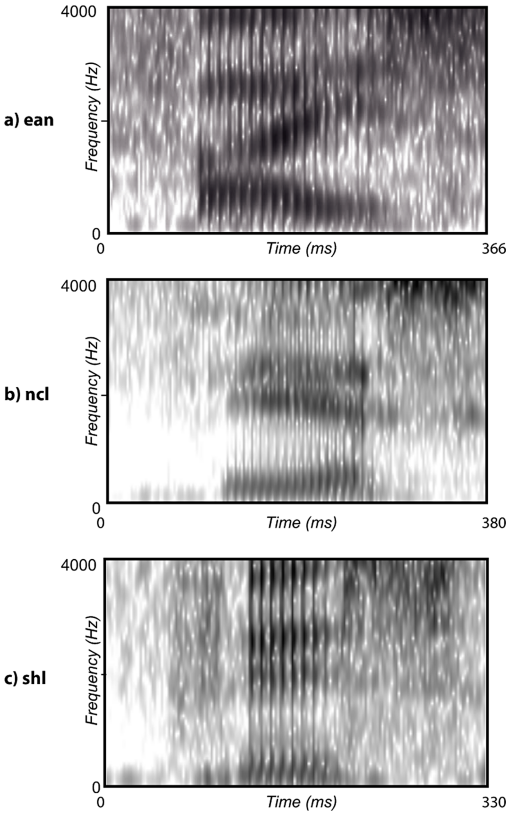

Fig. 1. Spectrograms exemplifying the three realizations of the FACE vowel

In a multi-dialect perspective, the 19 /hVd/ words presumably instantiate all possible phonological contrasts in the dialect that has the biggest inventory[4]. Artificial though the stimuli may seem, they nevertheless give the opportunity to calibrate the system under ideal conditions for subsequent use on data closer to real-life speech (see Experiment 3), and provide a convenient way of studying variation in phonological systems. Prior to the analysis proper, a native English expert phonetician examined the corpus and advised us against including the *ilo* subset on the grounds that the extreme heterogeneity of the speakers could in no way form a single entity (further details are given in section 7.1). More than for any other dialect in the corpus, individual information on speakers would have

[4] This again is an oversimplification: to be more accurate, the 19 stimuli exemplify the phonological vowel contrasts of Standard British English, which implies that the other vowel inventories are assessed with reference to that of Standard English, and not to an ideal panlectal representation. Thus, we have no means of knowing whether increasing the number of /hVd/ words would elicit other contrasts in the remaining dialects.

Table 2. Paired-dialect discrimination based on diphthongization. The ten highest and lowest scores are displayed. All scores, unless specified (ns), are significant at the $p = .05$ level (binomial tests).

DIALECT1	DIALECT2	CORRECT DISCRIMINATION (%)		GEOGRAPHICAL DISTANCE (km)
brm	shl	97.8		581
ean	shl	96.8		658
ean	gla	96.3		541
shl	sse	96.1		712
brm	gla	95.9		406
crn	shl	95.5		829
ilo	shl	95.4		712
brm	ncl	95.2		277
lvp	shl	95.0		471
ean	ncl	94.7		354
		...		
lvp	roi	57.3		219
lvp	sse	56.3	ns	285
nwa	roi	52.6	ns	189
crn	lvp	51.4	ns	380
crn	sse	51.3	ns	374
ilo	sse	51.3	ns	0
gla	ncl	51.0	ns	193
lvp	nwa	49.7	ns	38
eyk	lan	49.3	ns	125
crn	ilo	42.5		374

been essential. ABI comes complete with a word-level segmentation; assuming – although this is not totally accurate – that voiced frames corresponded to vowels, automatic pitch detection with the Snack Sound Toolkit ([18]) was employed to estimate vowel boundaries. 12 MFCC and one energy feature were computed at 25 %, 50 %, and 75 % of the duration of the vowel, and the duration itself was included to form a vector of 40 features. The computation was done with the melfcc routine from the rastamat toolbox ([19]); the options were those that the author recommends to duplicate HTK's MFCC, except that the window length and the analysis step were set to 20 ms and 10 ms, respectively. After removing the speakers from *ilo* and two participants who did not complete the whole set of test words, we were left with 261 speakers. The rationale for the classification method was first introduced, as far as we know, by [6], and it was later adopted by [1,2], who devised the ACCDIST method (Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table), which is central to this section[5]. Speaker normalization is a critical issue in phonetics: differences in

[5] [1,2] also used the ABI corpus; he however worked on a different part of the database, namely, a set of shibboleth sentences.

individual acoustic spaces, either due to physiological constraints or habit, have to be factored out. [6] and later [1,2] got round the problem by representing vowels with reference to a speaker's vowel space structure, and not to average stored values. One way to do this is to compute distances between each pair of vowels. As mentioned above, a vector of size 40 was computed for each vowel. For a given speaker, the values for the five repetitions of each /hVd/ type were averaged. Then, distances were calculated between the 19 vowel types, yielding, for each speaker, a 19×19 symmetric distance matrix. Quite a few distance measures for continuous variables are available in the literature (see for example [20], for a discussion of the properties of some of them), and the choice of the appropriate one depends on the particular kind of data. Central to the problem is the issue of variable weighting: in our $n \times p$ matrices (where n are the 19 vowels of a speaker and p the 40 spectral and duration features), the ranges and scales of the p variables differ substantially. It is common practice to standardize each variable to zero mean and unit variance (i.e. computing a so-called z-score); yet, we assumed that, given that the computation of MFCC is based on an auditory filter bank, the differential weightings induced by differences in scales and ranges reflected perceptually relevant attributes of the spectrum, and should therefore be preserved. A good choice in such cases is to use a family of distance metrics whose general form is the Minkowski distance:

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}} \tag{1}$$

where r must be superior or equal to 1. As the chosen r value increases, the differential weighting of the p variables also increases: large differences are given relatively more weight than small ones. Bearing in mind what has just been said about the perceptual relevance of our feature space, we want to avoid distorting it by using high exponents and will therefore stick to low values such as $r = 1$ and $r = 2$, which correspond to the Manhattan and Euclidean distances, respectively. So, once a 19×19 distance matrix has been computed for each speaker, the classification method described in [1,2] is carried out: 13 dialect matrices are obtained by getting the mean of the individual matrices for each dialect. The validation procedure goes as follows: the dialect matrix of the speaker under test is re-computed without her/his individual matrix and then each individual matrix is compared to the 13 dialect matrices. Matrix similarity is estimated with a matrix correlation coefficient: the two matrices, i.e. the test speaker and the dialect matrix (or rather: either the upper or lower triangular part, since they are symmetric) are unfolded onto a row vector, then the Pearson product-moment correlation is computed. The speaker under test is classified as belonging to the dialect whose correlation with her/his matrix is highest. Percent correct identification scores are 86.6 %, 89.0 %, and 89.7 %, for men, women, and both sexes respectively, using the Euclidean distance. Slight improvements are obtained in both sexes condition with the Manhattan distance: 90.0 %. Correlation measures are insensible to scale magnitude, which solves the question of

Table 3. Confusion Matrix: /hVd/ words; all subjects; Manhattan distance. Overall correct score: 90.0 %.

TEST DIALECT	MODELS BRM	CRN	EAN	EYK	GLA	LAN	LVP	NCL	NWA	ROI	SHL	SSE	ULS
brm	18	-	1	-	-	1	-	-	-	-	-	-	-
crn	-	16	-	-	-	-	-	-	-	1	-	3	-
ean	1	-	15	-	-	-	-	-	-	-	-	3	-
eyk	2	-	-	22	-	-	-	-	-	-	-	1	-
gla	-	-	-	-	18	-	-	-	-	-	-	-	2
lan	-	-	-	-	-	21	-	-	-	-	-	-	-
lvp	-	-	-	-	-	-	19	-	-	-	-	-	-
ncl	-	-	-	-	-	-	-	18	1	-	-	-	-
nwa	1	-	-	-	-	-	1	-	18	-	-	-	-
roi	-	-	-	-	-	-	-	-	1	19	-	-	-
shl	1	-	-	-	1	-	-	-	-	-	19	-	1
sse	1	1	2	-	-	-	-	-	-	-	-	12	-
uls	-	-	-	-	-	-	-	-	-	-	-	-	20

speaker normalization. Note incidentally that the method is unaffected by sex differences.

5.2 Gaussian Modelling

An alternative classification using Gaussian modelling was carried out with the Netlab ([17]) toolbox. The model takes z-scored individual distance matrices as input and estimates one Gaussian model $N(\mu, \sigma)$ per dialect. As before, the test speaker is excluded from the training set; in other words, for each speaker, a new model is trained on all the data minus this speaker's matrix. The estimated dialect identity is then given according to the Maximum Likelihood decision. This statistical decision yields a non significant improvement over the previous method: for the both sexes condition, the model achieves 90.4 % correct classification.

5.3 Discussion

Both methods seem to perform equally well, which might indicate that a ceiling has been reached for this particular corpus. This question will be addressed more in depth in Section 7.1. A close examination of Table 3 suggests that linguistic explanations can often justify some of the misclassifications. For example, the historical link between *ean* and *sse* may account for the 3 *ean* speakers being classified as *sse*, and the 2 speakers of *sse* being classified as *ean*. The fact that 2 speakers of *gla*, and 1 from *shl* were identified as *uls* could be accounted for by saying that the 3 dialects belong to a common super region, namely, the Celtic countries. The high scores were of course facilitated by the absence of

co-articulatory variation; yet, it is worth pointing out that even /hVd/ words – whose weaknesses are constantly condemned – contain essential information about dialect. And, what is more, they probably constitute the quickest and most convenient way to form an opinion about the linguistic quality of a corpus, or the feasibility of a classification task.

6 Experiment 3: Dialect Classification with Read Passages

The ACCDIST procedure is then applied to the read passage part of the corpus. The segmentation was obtained through forced-alignment as in Experiment 1. The number of words uttered by all speakers amounted to 61. When words were polysyllabic, only the stressed syllable was kept for the classification. The same spectral and duration parameters as in Experiment 2 were computed. One option would have been to classify the vowels according to the lexical set they belonged to. However, this would have artificially reduced the diversity of coarticulatory phenomena, possibly leading to poor performances, and it would have necessitated the intervention of a human expert in order to infer lexical set membership of the stressed vowel in a given word. This would in turn have led to a manifold increase in the tedium and the time to carry out the classification, not to mention the questionable theoretical validity of such inferences. A sounder approach that by-passes such linguistic hypotheses was therefore adopted: instead of vowel types, distances were computed between vowel tokens. Note here that 264 speakers are included. The 61×61 individual distance matrices were then classified with the same correlation-based procedure that was used for the /hVd/ words. 89.6 %, 87.6 %, and 90.5 % correct classification are obtained for men, women, and both sexes respectively with the Euclidean distance. The Manhattan distance yields 87.4 % and 89.4 % for men and women; there is no improvement for the third condition.

7 General Discussion

7.1 Guidelines to Assess Classification Scores

One of the questions underlying these experiments is how good a 90 % correct classification score is with respect to the data that has been analysed. A fundamental conceptual discrepancy between language identification and dialect identification should help us come up with a tentative answer. Except for a few borderline cases – including code-switching –, language sets are in principle mutually exclusive; in other terms, a speaker either speaks language A or language B, and certainly not a mixture of the two. Matters get more complicated for dialect corpora: dialect membership for a speaker does not mean that the speaker produces all the phonetic features of that particular dialect, nor does it mean that s/he does not use features from other dialects. And as the distance (however it is measured) of a speaker from its dialect prototype increases,

so does the risk of this speaker being associated (by a naive listener, an expert phonetician, or the machine) with another dialect. In other words, it is undoubtedly more adequate to view dialect classes as fuzzy sets, and language classes as hard sets, although quite circularly, depending on linguistic denominations, we may come across borderline cases: if we use the linguistic criterion of mutual intelligibility, some entities traditionaly termed "languages" can overlap (see the case of Danish, Swedish, and Norwegian) while others called "dialects" may be rather distinct (possibly the case for distant dialects of Arabic). Translating this into figures, it could be said that language identification scores must be judged against the maximal achievable score (i.e. 100 % in almost all cases) whereas, there is no simple way to estimate this figure for dialects. There probably exists a floor (above chance level) below which the scores of an automatic dialect identification system can be considered bad; this floor could be given by classification carried out by naive listeners. And there certainly is another threshold around which scores can be deemed excellent. We tried to estimate the value of the latter threshold with an informal experiment: a native speaker expert phonetician was asked to listen to one third of all the passages spoken by men in the ABI corpus. The experiment was actually divided into 14 (one per dialect) separate verification tasks. In each task, the expert had to listen to a stimulus and say whether it had been uttered by a speaker of the dialect of the current task or not. We will not go into too much detail since this is beyond the scope of the present research, suffice it to say that the expert scored 89.6 %. Of course, proficient though the expert may have been, his degree of acquaintance with dialects probably varied from one to the next, but this is the closest we can get to estimating the highest possible classification score. Ceiling effects in classification accuracy are also suggested by a statement in the documentation of the corpus acknowledging that some speakers, particularly in *crn* and *nwa*, have an accent that might not be regarded as typical.

7.2 Descriptive Scope

Part of the descriptive task of the phonetician is to come up with linguistically interpretable visual representations from multidimensional raw numerical data. Graphical displays, particularly vowel plots, have been frequently used to illustrate phonetic phenomena. This section exemplifies how the methods employed in Experiment 2 for classification can be used as a descriptive tool. The dendrograms in Figures 2 and 3 display the output of hierarchical clustering computed with the single linkage algorithm implemented in Matlab for a selected set of vowels in two female speakers from *eyk* and *shl* respectively. The first tree clearly shows the relative proximity of *hood* and *Hudd*, exemplifying the well-known absence of phonemic split in the north of England we discussed in Section 2. The second tree illustrates the phonemic merger in Scotland involving the vowels of *hood* and *who'd*.

Figure 4 shows the scatter of women from six selected dialects based on individual 19×19 distance matrices computed with the /hVd/ words. Each individual matrix was z-scored and dimensionality was reduced with principal

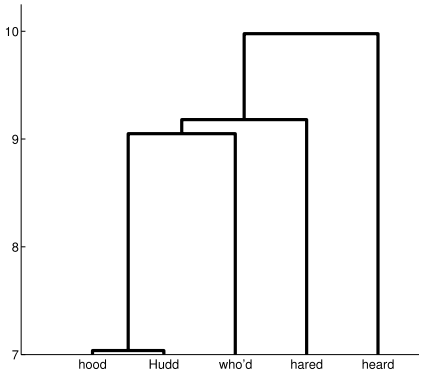

Fig. 2. Dendrogram illustrating the absence of *hood* vs. *Hudd* phonemic split in *eyk*

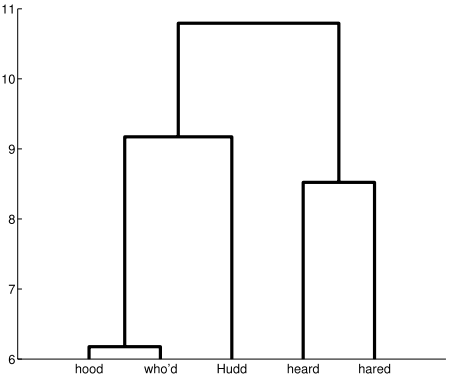

Fig. 3. Dendrogram illustrating the phonemic merger involving *hood* and *who'd* in *shl*

component analysis. The plane is defined by the first two principal components, which account for approximately 35 % of the variance of the original data. High though the distorsion may be, meaningful patterns can still be identified on the graph: an imaginery oblique line separates the dialects of England (*ean*,*lan*, and *ncl*) from those of the Celtic countries (*gla*, *roi*, and *shl*). Then, within the English group, an almost geographical picture emerges: *ean* in the south east, *lan* in the north west, and *ncl* in the north east. In the Celtic group, Scotland and Ireland are neatly split, with *roi* being distinct from *gla* and *shl*[6]. Finally, within the Scottish subset, the situation looks more fuzzy (but this may simply be a consequence of dimensionality reduction), although there is a tendency for *gla* speakers to cluster near the bottom of the graph, and *shl* speakers above the latter. Whatever the goodness of the final display, the efficiency of inter-segment distance matrices to capture dialect specifities is confirmed by the bidimensional map whose interpretation in linguistic and geographical terms makes perfect sense.

7.3 Suggested Improvements

We now turn to the question of how to improve the classification scores. Consider the $n \times p$ matrix where n refers to the 261 speakers and p to the $19(19-1)/2 = 171$ distances (i.e. the unfolded 19×19 individual symmetric matrix) between pairs of vowels. It is very unlikely that all distances possess equal discriminatory power: some may be extremely relevant, e.g. those between two vowels that can be either merged or not depending on the specific vowel system, others may have only slight discriminatory power, for example those implying minute phonetic differences, and others may be irrelevant altogether. In addition, measurements

[6] Note however that one speaker from *roi* ended up with the *ncl* cluster.

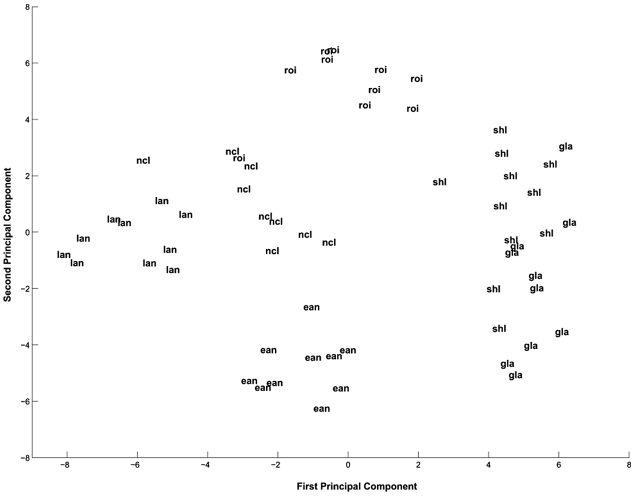

Fig. 4. Female speakers from 6 dialects along 1st and 2nd principal components derived from the distance matrices on /hVd/ words

on continuous scales contain noise, hence we could consider dichotomizing some of the quantitative variables. For example, phomenic mergers, or the absence of phonemic splits, could be regarded as binary events: on a continuous scale, the distance between *hood* and *Hudd* in northern English dialects is never equal to zero, although it should be in systemic (phonological) parlance. Besides, it may vary between speakers despite their producing exactly the same target vowel in these two words. The varying distances between *hood* and *Hudd* in a set of speakers having no *hood* vs. *Hudd* contrast is linguistically irrelevant, and it adds noise to the system. So there must be a threshold in the distance measured on a continuous scale below which the two vowels can be regarded as identical; and above this threshold, the two vowels can be considered different. Feature selection (recall that the features are the $p = 171$ distances between vowel pairs) would be desirable for at least three reasons. Firstly, it would rid the system of noisy variables, possibly improving classifications scores and reducing computational cost. Secondly, some modelling techniques require a subtle balance between the number of examples to classify and the size of the feature space (the n and p dimensions in the matrix respectively); given the small size of n in our data, reducing p is imperative. Thirdly, and most interestingly, special cases of feature selection such as feature ranking and feature weighting can provide explanatory principles: such methods as K-means partitioning may be used to assess the relative weight of each feature ([21]). This assessment could in turn validate linguistic hypotheses on the discriminatory power of each feature. All these methods work *a posteriori* in that they need the data first; another possible improvement would be to include linguistic knowledge prior to data analysis. [6] applied such a procedure to increase the potential differentiation of dialects: for

example, if the distance between the vowels of *father* and *after* is smaller than that between *cat* and *after*, then strong evidence for a southern English dialect is obtained, whereas this weighs against northern English dialects, and neither favours nor disfavours Scottish dialects. So [6] came up with an *a priori* trivalent weight system which somewhat enhances the discrimination on the basis of phonological knowledge after the raw numerical evidence has been accumulated.

7.4 Perspectives

The classication method presented here is text-dependent: what is being said must be known beforehand, and the words of the training and test sets must match. Besides, it is based on phonetic and phonological knowledge of dialect differences, and we must bear in mind that the stimuli (/hVd/ words and read passages) were precisely designed to elicit dialect variation, and therefore facilitate discrimination. So this approach can be termed shibboleth-based. Now, how good would the performance be with a randomly chosen text? More specifically, how could one deal with mismatches between the vowels of the training datasets and those of the test speaker set? Another challenge is the transposition of the method to spontaneous speech. Future research will focus on text-independency and include other phonetic cues such as consonants and suprasegmentals.

Acknowledgements

We are grateful to Mark Huckvale and Francis Nolan for their help. This study was supported by a Eurodoc grant from the Région Rhône-Alpes.

References

1. Huckvale, M.: ACCDIST: a metric for comparing speakers' accents. In: Proceedings of Interspeech 2005, Jeju, Korea, pp. 29–32 (2004)
2. Huckvale, M.: ACCDIST: an accent similarity metric for accent recognition and diagnosis. In: Müller, C. (ed.) Speaker Classification. Lecture Notes in Computer Science / Artificial Intelligence, vol. 4343, Springer, Heidelberg (2007) (this issue)
3. Yan, Q., Vaseghi, S.: A comparative analysis of UK and US english accents in recognition and synthesis. In: Proceedings of ICASSP, Orlando, Florida, USA, pp. 413–417 (2002)
4. Ellis, S.: The Yorkshire Ripper enquiry: Part I. Forensic linguistics 1, 197–206 (1994)
5. Jessen, M.: Speaker Classification in Forensic Phonetics and Acoustics. In: Müller, C. (ed.) Speaker Classification. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
6. Barry, W.J., Hoequist, C.E., Nolan, F.J.: An approach to the problem of regional accent in automatic speech recognition. Computer Speech and Language 3, 355–366 (1989)
7. Huang, R., Hansen, J.: Advances in word based dialect/accent classification. In: Proceedings of Interspeech 2005, Jeju, Korea, pp. 2241–2244 (2005)

8. Orton, H., Sanderson, S., Widdowson, J.: The linguistic atlas of England. Croom Helm, London (1978)
9. Wells, J.: Accents of English. The British Isles, vol. 2. Cambridge University Press, Cambridge (1982)
10. Foulkes, P., Docherty, G.: Urban Voices. Accent Studies in the British Isles. Arnold, London (1999)
11. Kortmann, B., Schneider, E.W.: A Handbook of Varieties of English. Mouton de Gruyter, Berlin, Germany (2004)
12. D'Arcy, S.M., Russell, M.J., Browning, S.R., Tomlinson, M.J.: The Accents of the British Isles (ABI) corpus. In: Proceedings of MIDL 2004 Workshop, Paris, France, LIMSI-CNRS, pp. 115–119 (2004)
13. Schötz, S., Müller, C.: A Study of Acoustic Correlates of Speaker Age. In: Müller, C. (ed.) Speaker Classification. Lecture Notes in Computer Science / Artificial Intelligence, vol. 4343, Springer, Heidelberg (2007) (this issue)
14. Schötz, S.: Acoustic Analysis of Adult Speaker Age. In: Müller, C. (ed.) Speaker Classification. Lecture Notes in Computer Science / Artificial Intelligence, vol. 4343, Springer, Heidelberg (2007) (this issue)
15. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK version 3.2). Cambridge University Engineering Department, Cambridge (2002)
16. Boersma, P., Weenink, D.: Praat. Doing Phonetics by Computer. version 4.4.22 (2006)
17. Nabney, I.T.: Netlab. Algorithms for Pattern Recognition. Springer, London (2002)
18. Sjölander, K., Beskow, J.: Wavesurfer - an open source speech tool. In: Proceedings of ICSLP 2000, Beijing, China, pp. 464–467 (2000)
19. Ellis, D.P.W.: PLP and RASTA (and MFCC, and inversion) in Matlab, online web resource (2005)
20. Gower, J.C., Legendre, P.: Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 5–48 (1986)
21. Makarenkov, V., Legendre, P.: Optimal variable weighting for ultrametric and additive trees and K-means partitioning: Methods and software. Journal of Classification, 245–271 (2001)