HIERARCHICAL CLUSTERING OF SPEAKERS INTO ACCENTS WITH THE ACCDIST METRIC

Mark Huckvale

Department of Phonetics & Linguistics, University College London, U.K.
m.huckvale@ucl.ac.uk

ABSTRACT

Hierarchical clustering of speakers by their pronunciation patterns could be a useful technique for the discovery of accents and the relationships between accents and sociological variables. However it is first necessary to ensure that the clustering is not influenced by the physical characteristics of the speakers. In this study a number of approaches to agglomerative hierarchical clustering of 275 speakers from 14 regional accent groups of the British Isles are formally evaluated. The ACCDIST metric is shown to have superior performance both in terms of accent purity in the cluster tree and in terms of the interpretability of the higher-levels of the tree. Although operating from robust spectral envelope features, the ACCDIST measure also showed the least sensitivity to speaker gender. The conclusion is that, if performed with care, hierarchical clustering could be a useful technique for discovery of accent groups from the bottom up.

Keywords: Accents, Clustering, Instrumental Methods, Socio-phonetics.

1. INTRODUCTION

Hierarchical clustering [1] is a data analysis procedure that aims to uncover structural relationships between objects and between groups of objects. Input is typically a set of pairwise similarity measurements of the objects, and output is a hierarchical tree or dendrogram (see Fig. 1 for an example). The hierarchical clustering of speakers by their accent is a particularly interesting application, because the sociological variables affecting accent are many and interact in complex ways [2]. Indexical variables related to region or class may themselves show some intrinsic hierarchical structure. Hierarchical clustering has been used previously in a small number of socio-phonetic studies of accent variation. For example, it has been used to cluster speakers by frequency of use of phonetic forms [3], to cluster foreign accent types by vowel quality [4], and to investigate Swedish regional accents by allophonic variation [5].

The success of hierarchical clustering analysis applied to speakers clearly relies on the operation of the chosen similarity metric. This metric must be sensitive only to the speakers' accents. A metric which was also sensitive to the physical characteristics of the speakers might, for example, cluster them according to height, age or sex, rather than on the nature of their accent. It would be nonsensical to use cluster analysis to claim that male and female speakers had different accents, if one could not be sure that the similarity metric used was insensitive to the physical differences between the sexes.

This investigation studies the effectiveness of hierarchical clustering for accent analysis. We look at different hierarchical clustering methods in combination with different speaker similarity metrics to examine whether hierarchical cluster analysis can recover the known accent groupings of a set of male and female speakers chosen to be representative of 14 regional accents of the British Isles. Acoustic representations are formant frequencies, normalised formant frequencies and spectral envelope measures of vowels. Similarity metrics are correlation, Euclidean and the ACCDIST metric [5]. Agglomerative clustering techniques are single, complete, average, group and Ward linkage methods. Evaluation is in terms of sub-tree purity, and in terms of the interpretability of the top levels of the cluster tree.

2. SPEECH DATA

Data is taken from the Accents of the British Isles (ABI) corpus [7]. Twenty sentences from each of approximately 10 male and 10 female speakers from each of 14 regions were used. The region codes are shown in Figure 1.

A phonological transcription was generated for each sentence using Southern British English pronunciations, and phonetic segmentation was performed by forced alignment using the HTK Hidden

Markov Modelling toolkit [8]. All subsequent analysis was made using only the vowel segments in the 20 sentences including diphthongs but excluding schwa. This gave up to 145 vowel measurements per speaker.

Formant locations were estimated by LP analysis and single frequency values for each half-vowel were found from the trimmed mean. Formant frequencies were normalised for each speaker, using Z-scores, according to the recommendation of [9]. The spectral envelope representation of each half-vowel was calculated from the average mel-scaled cepstral coefficients (MFCC). Signal analysis was performed with SFS [10]. More detail can be found in [11].

3. CLUSTERING

An agglomerative clustering method was used to combine speakers into groups from the bottom up. At the start of clustering each speaker is placed in their own sub-tree, then at each step two trees are combined. The choice of which sub-trees are to be combined is based on the similarity between speakers and the linkage method. In the 'single' linkage method, the two trees which have the most similar pair of speakers are combined. In the 'complete' method, the two trees which have the least difference between the most different speakers are combined. In the 'average' method, the two trees which have the smallest average distance between all pairs of speakers in the trees are combined. In the 'group' method, the two trees with the smallest distance between the centroids of the clusters are combined. In the Ward method, the two trees which add least to the overall variance of the clustered speakers from their centroids are combined.

To compute the similarity between speakers we used (i) the correlation between the acoustic measurements of the two speakers, (ii) the Euclidean distance between the acoustic measurements of the two speakers, (iii) the weighted Euclidean distance between the acoustic measurements of the two speakers, using the measured parameter variance across all speakers, and (iv) the ACCDIST distance [5]. The ACCDIST measure first computes a table of segment similarities (SS) *within* a speaker (N) from the euclidean distance between all pairs of segments (s_i):

$$SS_{ij}^{N} = dist(s_i^N, s_j^N)$$

It then correlates two segment distance tables *across* speakers:

$$ACCDIST = corr(SS^1, SS^2)$$

Thus it provides a distance measure which is only based on the relative similarities of the segments and not on any absolute properties. This measure has been shown to have superior performance to other similarity measures in an accent recognition task [11].

To provide a formal evaluation of the success of clustering, a sub-tree 'purity' measure was used. For two speakers of the same accent, we can find the smallest sub-tree that contains them both. We can then compute the proportion of speakers in that sub-tree which are of the same accent as the two speakers. The purity of the whole tree is the average of this proportion taken over all pairs of speakers within each accent. A perfect tree, with each accent group in its own sub-tree, would have a purity of 1. For comparison, we can also compute purity on the basis of gender, and ideally this would be 0.5 since there are equal numbers of men and women in the tree and no selection should be performed on the basis of speaker sex.

4. CLUSTER PURITY RESULTS

Table 2 shows the accent purity results for formant, normalised formant and spectral-envelope parameters as a function of similarity metric and linkage type.

A number of observations can be made from these results. The only satisfactory linkage methods are 'complete' and 'average', presumably because these both take into account the extremes of the cluster, not just the closest or the middle point. The best similarity metric was ACCDIST, which outperformed the other measures in all configurations. The second best similarity measure was correlation. The use of a weighted Euclidean measure made no significant improvement over simple Euclidean on these parameters. Although formant frequency normalisation improved the performance of the Euclidean metric, it had little effect on the correlation and ACCDIST measures. Although the ACCDIST measure performed well on all feature sets, it performed best on the spectral envelope parameters, probably because this gave the most robust representation of vowel quality.

Table 3 shows the gender purity measures for the better results in Table 2. Here it is easy to see that the configurations with the best gender purity values (i.e. those close to 0.5) are those with also

the best accent purity values. Correlation of spectral envelope parameters showed the most sensitivity to speaker sex, as could be expected. Formant frequency normalisation did not improve upon the already good results obtained with correlation of un-normalised frequencies. In all configurations ACCDIST showed the least sensitivity to speaker sex.

Table 2. Accent purity of hierarchical clustering of ABI speakers by linkage method, similarity metric and acoustic parameter set [2F=two formants, 2FN=two normalised formants, Env=MFCC spectral envelope] N=275.

Linkage	Similarity	Acoustic Parameters		
		2F	2FN	Env
Single	Correlation	0.137	0.198	0.142
	Euclidean	0.133	0.134	0.128
	Weighted	0.124	0.110	0.142
	ACCDIST	0.236	0.234	0.300
Complete	Correlation	0.508	0.450	0.190
	Euclidean	0.251	0.434	0.155
	Weighted	0.244	0.452	0.192
	ACCDIST	**0.570**	0.565	0.647
Average	Correlation	0.457	0.459	0.203
	Euclidean	0.243	0.349	0.188
	Weighted	0.246	0.367	0.211
	ACCDIST	0.556	**0.576**	**0.724**
Group	Correlation	0.150	0.175	0.090
	Euclidean	0.123	0.116	0.092
	Weighted	0.117	0.096	0.095
	ACCDIST	0.182	0.171	0.171
Ward	Euclidean	0.293	0.558	0.129

Table 3. Accent purity and gender purity for the better clustering results [2F=two formants, 2FN=two normalised formants, Env=MFCC spectral envelope] N=275.

Params	Linkage	Similarity	Accent Purity	Gender Purity
2F	Complete	Correlation	0.508	0.525
		ACCDIST	0.570	0.519
	Average	Correlation	0.458	0.519
		ACCDIST	0.556	0.516
2FN	Complete	Correlation	0.450	0.571
		ACCDIST	0.564	0.514
	Average	Correlation	0.459	0.538
		ACCDIST	0.576	0.516
Env	Complete	Correlation	0.189	0.825
		ACCDIST	0.647	0.516
	Average	Correlation	0.202	0.900
		ACCDIST	**0.724**	**0.512**

5. CLUSTER INTERPRETATION

Figure 1 compares the top levels of the hierarchical cluster analysis performed using complete linkage and (a) correlation of two-formant features, with (b) ACCDIST of spectral envelope features. It is apparent that tree (b) is more balanced and contains nodes which are more readily interpretable in terms of known accent groups. For example, node 2 contains the Scottish accents, while node 6 contains predominantly Northern and node 7 predominantly Southern English accents. Tree (b) also shows interesting socio-phonetics effects, for example node 17 includes both Irish and Newcastle accents, while 20 combines Inner London with Liverpool.

6. CONCLUSIONS

This investigation has attempted to validate the use of hierarchical clustering of speakers in socio-phonetic research. It has shown that with careful selection of cluster linkage method, similarity metric and acoustic features, good accent clustering results can be obtained, both in terms of purity and interpretability. Although this study is only based on vowels, the best method can easily be extended to consonantal and prosodic features.

7. REFERENCES

[1] Johnson, S., 1967, Hierarchical Clustering Schemes, *Psychometrika*, 2:241-254.

[2] Foulkes, P., 2006, Socio-Phonetics, in K. Brown (ed.) *Encycl. of Language and Linguistics* (2nd ed.). Elsevier.

[3] Stuart-Smith, J., Timmins, C., Tweedie, F., 2007, 'Talkin' Jockney'? Variation and change in Glaswegian accent, *Journal of Sociolinguistics*, in press.

[4] Vieru-Dimulescu, B., Boula de Mareuil, P., 2006, Perceptual identification and phonetic analysis of 6 foreign accents in French, *Interspeech-2006 Conference*, Pittsburgh, USA, 441-444.

[5] Salvi, G., 2003, Accent Clustering in Swedish Using the Bhattacharyya Distance, *Proc. International Congress of Phonetic Sciences (ICPhS)*, Barcelona, 1149-1152.

[6] Huckvale, M., 2004, ACCDIST: a metric for comparing speakers' accents. *Proc. International Conference on Spoken Language Processing*, Korea, 29-32.

[7] D'Arcy, S.M., Russell, M.J., Browning, S.R. and Tomlinson, M.J., 2005, The Accents of the British Isles (ABI) Corpus, *Proc. Modélisations pour l'Identification des Langues*, MIDL Paris, 115-119.

[8] htk.eng.cam.ac.uk

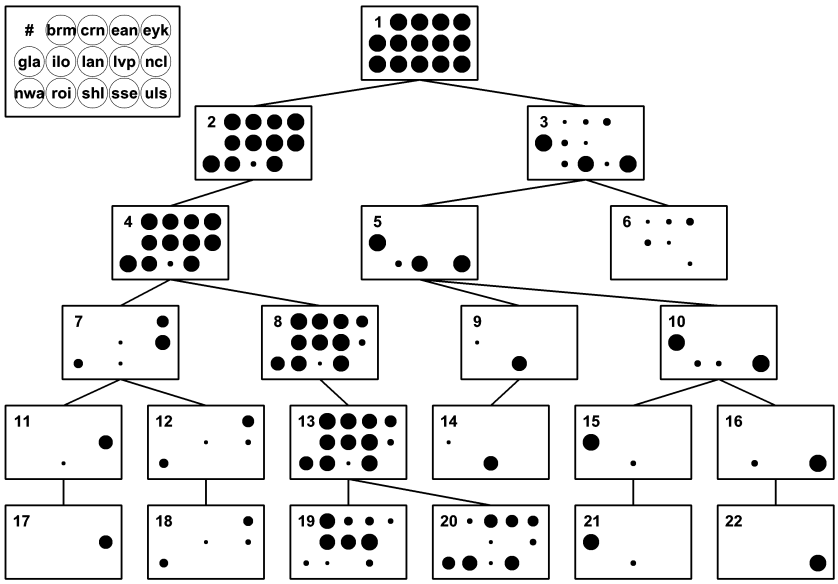
[9] Adank, P., Smits, R., van Hout, R., 2004, A comparison of vowel normalization procedures for language variation research, *J. Acoust. Soc. Am.* 116 (5) 3099-3107.

[10] www.phon.ucl.ac.uk/resource/sfs/

[11] Huckvale, M., 2007, ACCDIST: an accent similarity metric for accent recognition and diagnosis, in *Speaker Classification*, ed. Müller & Schötz, Springer LNAI.

***Figure 1**. Comparison of two approaches to agglomerative hierarchical clustering of 275 speakers analysed by originating accent group. Each figure shows the top levels of a tree of 549 nodes. The area of the disks represents the proportion of the accent group members present in the node. For clarity, nodes which contain fewer than 10 speakers, or whose parent node consists entirely of speakers of one accent have been pruned. Accents codes are: brm=Birmingham, lvp=Liverpool, crn=Cornwall, ncl=Newcastle, ean=East Anglia, nwa=North Wales, eyk=East Yorkshire, roi=Dublin, gla=Glasgow, shl=Scottish Highlands, ilo=Inner London, sse=South East, lan=Lancashire, uls=Ulster*

***(a)** Clustering using a correlation measure on 2 raw formant frequencies and the complete linkage method. Accent purity=0.508. Note how the tree is unbalanced, with the bulk of the speakers in a single node even at the fourth tier of the tree.*



***(b)** Clustering using the ACCDIST measure on 13 spectral envelope features and the complete linkage method. Accent purity=0.647. Note how this tree is more balanced than (a), and how accent groups combine more cleanly. Interesting analyses can be made based on which accent groups cluster together at the highest levels.*