Advances in Regional Accent Clustering in Swedish

Giampiero Salvi

KTH, (Royal Institute of Technology)
Department of Speech, Music and Hearing
Stockholm, Sweden

giampi@kth.se

10.21437/Interspeech.2005-719

Abstract

The regional pronunciation variation in Swedish is analysed on a large database. Statistics over each phoneme and for each region of Sweden are computed using the EM algorithm in a hidden Markov model framework to overcome the difficulties of transcribing the whole set of data at the phonetic level. The model representations obtained this way are compared using a distance measure in the space spanned by the model parameters, and hierarchical clustering. The regional variants of each phoneme may group with those of any other phoneme, on the basis of their acoustic properties. The log likelihood of the data given the model is shown to display interesting properties regarding the choice of number of clusters, given a particular level of details. Discriminative analysis is used to find the parameters that most contribute to the separation between groups, adding an interpretative value to the discussion. Finally a number of examples are given on some of the phenomena that are revealed by examining the clustering tree.

1. Introduction

Advanced statistical methods have long been used in speech recognition for acoustic modelling. The resulting models are often of overwhelming complexity and hard to interpret. This is common for data-mining techniques where the focus is on classification performance with large amounts of data.

More traditional statistical methods, on the other hand, provide a more transparent interpretation of the phenomena under study, but are generally less appropriate for large/huge data sets. This is a particularly severe limitation in the study of language and speech, where the great amount of variables involved makes restricted collections of data poorly representative of the problem.

In [1] it was proposed to use techniques from the data mining field, such as the ones used in speech recognition, as a tool for analysing pronunciation variants in students learning a second language. Similarly, we used related techniques to analyse regional accent variation on a large sample of the population in Sweden [2, 3]. Here the term *accent* is used following Crystal's acception [4] of regional pronunciation variation as opposed to the word *dialect* which describes variations in vocabulary and grammatical rules. The same distinction is made by Elert [5] who talks about the opposition between the regional variants of the standard language ("regionala rikssprĺksvarianter") and genuine dialects ("genuina dialekter").

The EM algorithm [6] was used to collect statistics over phonemic units, using only partially labelled material, and a canonical pronunciation dictionary, thus overcoming the problem of creating hand made detailed transcriptions of such a large corpus. Clustering techniques were used to visualise and interpret the results.

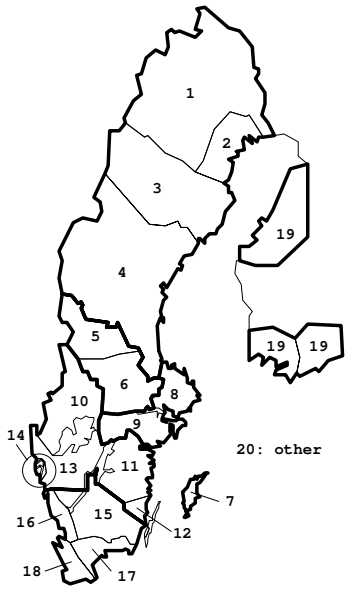

Figure 1: A map of Sweden and part of Finland with broad and finer dialect region definition. Number 20 is assigned to people with a language other than Swedish as mother tongue

This study is an evolution of the previous investigations [2, 3] in two main aspects. Firstly we overcome the limitation of studying each phoneme independently of the others, allowing for cross-phoneme similarities to emerge. Secondly, a number of methods are explored in order to quantify the fit of the clustering models to the data and to interpret their nature.

2. Method

2.1. Parameter estimation

The EM algorithm [6] is a well known method in the speech community that allows estimating a set of unknown parameters Θ given some measured data X when we lack information on some hidden variable, whose effect needs to be integrated out.

If we model continuous speech as a Markov chain, the hidden variable is the assignment of each data point (frame) to each state of the model. Representing the observations as Mel-frequency cepstral coefficients and selecting Gaussian distributions as "state-to-output" probability models, gives us a probabilistic framework to model the pronunciation of each phone-

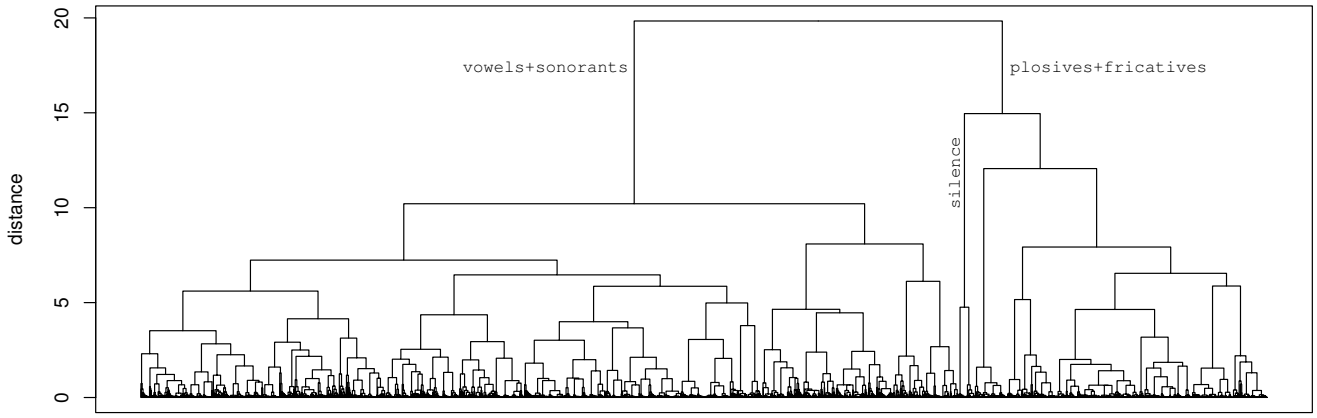

Figure 2: Dendrogram of the full clustering tree. The y-axis shows the dissimilarity level, on the x-axis are the states in the form phoneme-segment-region. Given the number of states it is not possible to display each label, Figure 4 shows three blow-ups for phoneme [ɪ]. Broad classes are also shown in the picture.

mic unit. A common practice, also adopted in this study, is to separately model the initial, middle and final part of each phoneme. I the following we will refer to those as *segments* of each phoneme.

In this framework we can introduce a number of sources of contextual information that are known to affect speech production. A common practice in speech recognition is, e.g., to model co-articulation effects by splitting each phoneme into context dependent units. The regional accent information investigated in this study can be interpreted as contextual information in the same manner, resulting in phonemic units that represent the characteristics of pronunciation patterns in specific geographical areas.

2.2. Clustering

Given our framework, a dissimilarity measure defined between distributions (states) together with agglomerative hierarchical clustering can be used to evaluate pronunciation differences and similarities. As in [2, 3] the metric used in this study is the Bhattacharyya distance. Differently from [2, 3], we do not make here the restriction of analysing each phoneme separately. Furthermore any segment (initial, middle and final) is free to cluster with any other. This gives us a large number of units that can freely form groups of similar pronunciation models. The exact number is $R \times P \times S$ where R is the number of homogeneous regions defined in the database, P the number of phonemes in the canonical pronunciation model, and S the number of segments for each phonemes (3 in our case).

Hierarchical clustering was chosen over partitional or density-based clustering because of its inherent property of displaying relational features at different levels of details. As will be clarified in the following, this is the main focus of this study, as opposed to finding the model that best fits the data in an absolute sense. The practical limitations with hierarchical procedures, e.g. the memory requirements for large datasets, are intrinsically avoided in this framework as the "objects" to be clustered are not single data points, but rather models representing (large) subsets of data points.

2.3. Cluster validation

A number of methods exist that evaluate the fit of the clustering model to the data. Milligan [7, 8] evaluates thirty cluster validation indexes with artificially generated data. Most of these methods try to evaluate the spread of points within and across clusters. Some relay on the pairwise dissimilarity measure alone, others on the relation between data points in each cluster and the corresponding cluster centroids in the space spanned by the data; finally some methods such as the log likelihood and the Bayes information criterion (BIC) [9] use a parametrised probabilistic framework.

In our case, clustering is applied to model parameters (means and covariances of the Gaussian distributions) rather than to single data points. Thus indexes of the first kind, such as the Ball [10] or Calinski [11] are not easily interpretable. On the other hand methods based on dissimilarities e.g. the C-index [12] are directly applicable to our case. Methods based on the likelihood of the data given the model such as the BIC, can also be applied as the log likelihood can be easily computed with some simplifying assumptions.

The log-likelihood of the data given a Gaussian model whose parameters are estimated according to the maximum likelihood criterion is:

$$L = \frac{1}{2} nD[\ln 2\pi + 1] - n \sum_{j=1}^{D} \ln \sigma_j$$

where D is the dimensionality of the data and n is the number of data points. Note that we do not need to refer to the actual data points to compute this quantity.

When the EM algorithm is used, the number of data points per Gaussian model (state) is unknown, but can be approximated by the so called *state occupation count* [13].

Making common assumptions on the effect of clustering on the state assignment, and on the possibility to approximate the total log likelihood with an average of the log likelihoods for each state, weighted by the probability of state occupancy, we can write the total log likelihood as a function of the means, covariances and occupation counts of the original states, for each cluster configuration along the clustering tree, without referring to the original data.

The BIC [9] is an approximation of the Bayes factor that tries to find a trade-off between the complexity of a model (number of free parameters m_M), the model fit to the data (log likelihood l_M) and the number of data points available for model parameter estimation (n). Its definition is

$$BIC \equiv 2l_M(x, \theta) - m_M \log(n)$$

and it is used for example in model-based clustering to select the best model according to the above criteria.

As we will see later, given the amount of data we are dealing with, the best possible model, includes a much larger number of parameters than those used in this study. As a consequence the above methods give monotonically increasing indexes, as our models are always under-dimensioned given the amount of data.

An alternative way to interpret these indexes, that will be used in this study, is to select among several models at different levels of details.

2.4. Cluster interpretation

When we select a merging (or splitting) point for each level of details of interest, we might want to know which acoustic parameters are more important to discriminate between the newly formed clusters. The way we do this is to apply discriminant analysis to the members of the two most recent clusters, and order the parameters according to the scaling factors in the resulting linear model (normalised by the variance of each parameter). We can then compute how much of the discrimination between the two groups is accounted for by one, two or more parameters, by running a prediction on the same cluster members with a linear model of the corresponding order. This gives us insights on the contribution of each coefficient to the global distance.

3. Data

The Swedish SpeechDat FDB5000 telephone speech database [14] was used for the experiments. It contains utterances spoken by 5000 speakers recorded over the fixed telephone network. All utterances were labelled at the lexical level and a lexicon is provided containing pronunciations in terms of 46 phonetic symbols. The database also contains information about each speaker including *gender*, *age*, *accent*, and more technical information about recordings, for example the type of telephone set used by the caller. In this study, only accent information was considered.

Figure 1 shows a subdivision of Sweden into a number of areas as defined by Elert [15]. Seven broad and twenty finer areas are defined according to dialectal homogeneity. The same areas were considered in this study to represent regions of uniform pronunciation (regional accents). They will be referred to as $\{r_1$–$r_{20}\}$.

The official SpeechDat training set was used for a total of 4500 speakers, 270 hours of speech. Ten milliseconds spaced frames were extracted for a total of circa 97 millions frames, containing 13 Mel cepstrum coefficients $\{c_0$–$c_{12}\}$, and the first and second order differences $\{d_0$–d_{12}, a_0–$a_{12}\}$. As already mentioned three states $\{s_1$–$s_3\}$ were used for each phoneme resulting in 2940 states (20 accent regions times 46 phonemes plus 1 silence and 2 noise models times 3 segments). Of these the two noise models and the retroflex allophone [ɭ] were removed (the last for lack of data points in certain regions). Thus a total of 2760 states (distributions) were used in the clustering procedure.

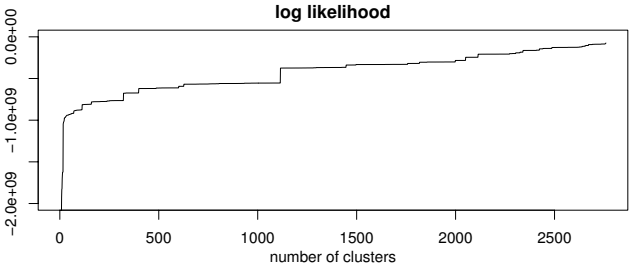

Figure 3: Log-likelihood versus number of clusters. Clear jumps can be seen when certain "merges" take place.

4. Results

Figure 2 shows the full clustering tree for the 2760 distributions used in this study. The figure is shown to give a general idea of the clustering process and is not intended as a complete representation of the results.

Figure 3 shows the corresponding log likelihood computed as described in Section 2.3. The values for number of clusters close to 1 have been truncated in the figure to better display the evolution for large number of clusters. The function is monotone and presents jumps at particular numbers of clusters. The corresponding BIC values have a very similar shape, meaning that, given the enormous amount of training data, the number of parameters in the model has a negligible effect compared to the log likelihood.

Starting from the top of the tree, the first split divides essentially vowels and sonorants from non-sonorants. The discriminant analysis predicts this classification with 0.5% error. The most important cepstrum coefficients are, in decreasing order, c_0 (energy), d_0 (derivative of the energy), c_1 (related to the spectral slope) and it's derivative d_1. Truncating the discriminant function to 1,2,3 and 4 variables, the prediction error is respectively 20%, 10%, 9% and 8%, meaning that the first four of the 39 coefficients account for 92% of the discrimination.

The second split divides the initial and final part (s_1 and s_3) of the silence model (for every accent region) from the rest of the states in group two from the previous split. Note that the middle state of silence clusters with the middle states of the voiceless plosives at a lower distance level. Successive splits divide, for example, sonorants from vowels and fricatives from plosives.

An interesting phenomenon is that the first and last segments (s_1,s_3) of the vowels tend to be separated from the middle part (s_2). This is possibly due to the latter being more stable in time (lower derivatives).

A general trend is that, in most cases, the splitting process separates states (distributions) belonging to different segments (initial, middle and final) and phonemes, while the effect of the accent region comes last in the process.

The exceptions to this are particularly interesting because they show the cases in which the pronunciation of one phoneme in the language is changed into the pronunciation of another phoneme that is also part of the phonetic inventory of that language. For example, [ʂ] (retroflex allophone of [s]) in the south of Sweden (r_{15}–r_{18}) clusters always with the main pronunciation of [s]. Similarly [ɧ] in Norrland and Finland (r_1–r_4, r_{19}) is more similar to [ç] and [ʂ] in the rest of the country. The vowel [ɵ] in Gotland (r_7) groups with the group including all variants of [ʊ] and [ɛː] in south-west Skåne (r_{18}) groups with the [æː] cluster, to give some examples with vowels.

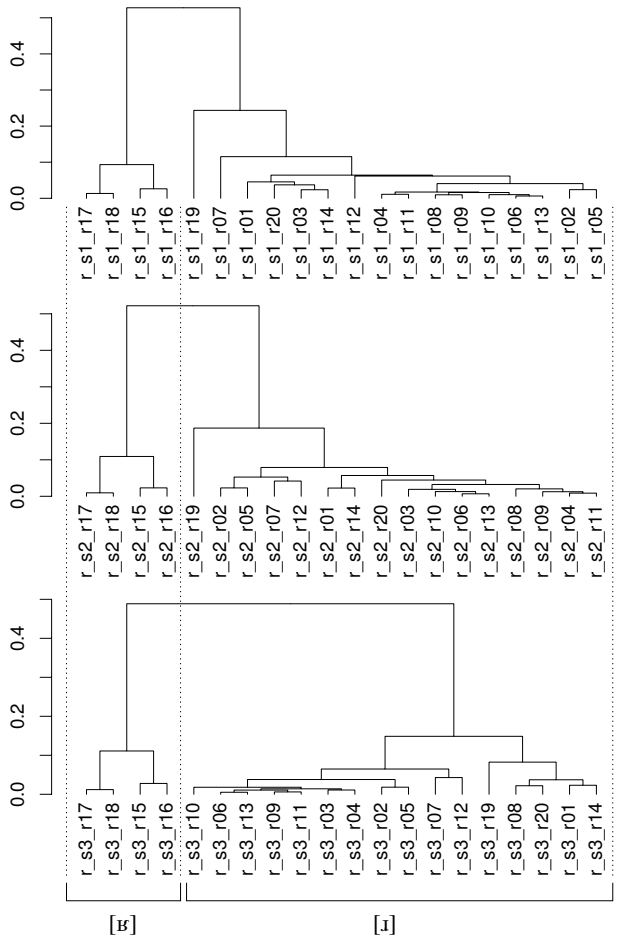

Figure 4: Subtrees for the phoneme [ɪ], states s_1, s_2, s_3 and regions r_1–r_{20}

In other cases, although the pronunciation varies greatly across regions, there is no corresponding sound in the standard pronunciation that can be confused with the regional variants. One example is [ɪ] that in the south (r_{15}–r_{18}) is retracted to [ʁ]. In this case both [ɪ] and [ʁ] group together, but the dendrogram (Figure 4) clearly shows a subgrouping that corresponds to this difference. Note that the figure shows three subtrees extracted from the complete tree in Figure 2.

5. Conclusions

This study proposes the use of methods from the data mining field to analyse pronunciation variation in a language using huge amounts of data. The intension is to grasp phenomena that may be missed with more restricted data sets.

The combination of the EM algorithm and clustering methods permits to find similarities and differences in the segments of each phoneme as pronounced in different regions in Sweden.

Discriminative analysis was used to interpret the results as it gives an indication of the importance of each cepstrum coefficient in discriminating between two groups of states.

The clustering tree was interpreted both at the minimum and maximum level of details (from top and bottom). In the second case two kinds of examples were shown: in the first case a regional variant of a phoneme is found to be closer to the average pronunciation of another phoneme. In the second case, this "merge" does not happen, but plotting the corresponding subtree(s) clearly displays the regional pronunciation variation.

6. Acknowledgements

This research was carried out at the Centre for Speech Technology supported by Vinnova (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. Thanks to Bruno Giordano for interesting discussion on statistical methods.

7. References

[1] N. Minematsu and S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," in *InSTIL'2000*, 2000, pp. 130–137.

[2] G. Salvi, "Accent clustering in Swedish using the Bhattacharyya dista nce," in *15th ICPhS Internamtional Congress of Phonetic Sciences*, August 2003.

[3] ——, "Using accent information in ASR models for Swedish," in *Eurospeech, 8th European conference on speech communication and technology*, sep 2003, pp. 2677–2680.

[4] D. Crystal, *The Cambridge encyclopedia of language*, 2nd ed. Cambridge university press, 1997, ch. 8, p. 24.

[5] C.-C. Elert, *Allmän och svensk fonetik*. Norstedts Förlag, 1996, ch. 6, pp. 34–35.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[7] G. Milligan and M. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, pp. 159–179, 1985.

[8] G. Milligan, "A Monte Carlo study of thirty internal criterion measures for cluster analysis," *Psychometrika*, vol. 46, pp. 187–199, 1981.

[9] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *Computer Journal*, vol. 41, no. 8, 1998.

[10] G. H. Ball and D. J. Hall, "Isodata. a novel method of data analysis and pattern classification." Menlo Park: Stanford Research Institute, Tech. Rep., 1965.

[11] T. Caliński and L. C. A. Corsten, "Clustering means in ANOVA by simultaneous testing," *Biometrics*, vol. 41, pp. 39–48, 1985.

[12] L. J. Hubert and J. R. Levin, "A general statistical framework for assessing actegorical clustering in free recall," *Psychological Bulletin*, vol. 83, pp. 1072–1080, 1976.

[13] S. Young and P. Woodland, "The use of state tying in continuous speech recognition." in *Eurospeech*, 1993, pp. 2203–2206.

[14] K. Elenius, "Experience from collecting two swedish telephone speech databases," *International Journal of Speech Technology*, vol. 3, pp. 119–127, 2000.

[15] C.-C. Elert, "Indelning och gränser inom området för den nu talade svenskan - en aktuell dialektografi," in *Kulturgränser - myt eller verklighet*, E. L.E., Ed. Diabas, 1994, pp. 215–228.