Accent Clustering in Swedish Using the Bhattacharyya Distance

Giampiero Salvi

Dept. Speech, Music and Hearing,
KTH, (Royal Institute of Technology)
Drottning Kristinasv. 31, 10044 Stockholm, Sweden
giampi@speech.kth.se

Abstract

In an attempt to improve automatic speech recognition (ASR) models for Swedish, accent variations were considered. These have proved to be important variables in the statistical distribution of the acoustic features usually employed in ASR. The analysis of feature variability have revealed phenomena that are consistent with what is known from phonetic investigations, suggesting that a consistent part of the information about accents could be derived form those features. A graphical interface has been developed to simplify the visualization of the geographical distributions of these phenomena.

1 Introduction

In automatic speech recognition (ASR), acoustic features extracted from the digitized speech signal are used to classify different phonetic or phonemic categories. This process, based on statistical methods, is made more difficult by a long list of phenomena that introduce acoustic variations in the speech signal. Among many others are *gender*, *age*, *level of education*, *anatomical characteristics*, *emotions* and *accent* of the speaker. The classic solution is to blindly increase model complexity and let some optimization procedure decide how to model each phenomenon. This approach is often successful, but has some drawbacks: the models obtained this way, given their complexity, are difficult to interpret and provide little information about the phenomena they were optimized for. Furthermore, the rising complexity sets limits to the efficiency of the optimization procedures. These can be technical, as for example the increasing amounts of data needed to adjust the model parameters, and the resulting computational load, or theoretical: models based on imprecise assumptions can be improved only to a certain extent.

Explicitly incorporating information into these models, on the other hand, requires the information to be formulated in a complete and consistent way. Dialectal and pronunciation variations in Swedish have been extensively studied by, among others, prof. Claes Christian Elert. In one of his books [4], that will be the main reference for this study, Elert defines areas of homogeneity as well as general descriptive rules to distinguish between them.

The aim of this study is to verify if data driven statistical models for speech recognition retain some of the accent information in a consistent way. This corresponds to verifying how accent related pronunciation variations influence the distribution of acoustic features. Since the analysis is done for each phoneme independently, it can also provide indications on how complex models are needed to describe the pool of allophones emerging from accent variations.

2 Method

Models for speech recognition are a suitable tool for collecting statistics on speech material that is not annotated at the phonetic level. The standard paradigm consists of a signal processing step aimed at extracting suitable *features*, and a *statistical modeling* step that characterizes the distributions of these features for each phoneme, and the dynamics of the speech production process. While the signal processing step is somehow standardized, the statistical models can have varying complexity depending on the application. In this study we consider simple dedicated models representing the allophones for each geographical area. An acoustic similarity criterion can then show which of these areas can be merged (according to each phoneme) indicating that, from the speech recognition viewpoint these areas are homogeneous.

2.1 Features

Speech feature extraction has two main purposes in speech recognition: the first is to reduce the amount of data per time unit the classification algorithm has to work on; the second is to reduce, as much as a sim-

ISBN 1-876346-48-5

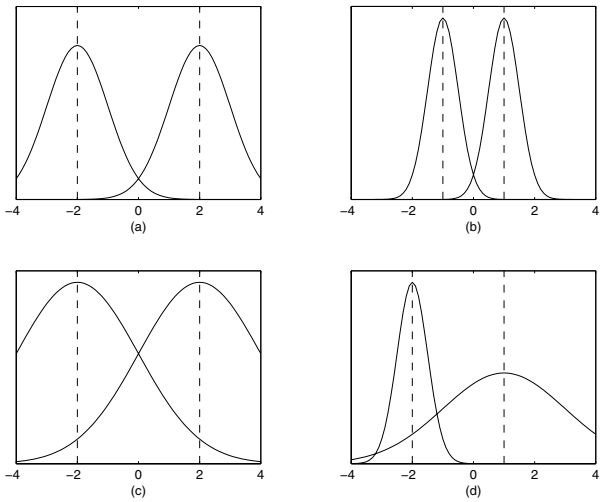

Figure 1: One-dimensional example of pairs of Gaussian distributions. (a) and (c) show pairs with the same mean (Euclidean) distance, but different Bhattacharyya distances; (a) and (b) on the other hand have different mean distances, but similar Bhattacharyya distances. In general (d) the variances of the two distributions can be different.

ple filtering method can, the amount of *spurious* information and noise that can degrade accuracy, while enhancing the phonetic related characteristics of the signal. Features used in this study, and in most speech recognition systems, are Mel scaled cepstral coefficients plus short time energy, first introduced in [1]. One of the properties of these features can be interesting in this context: the cepstrum is based on modeling the spectrum envelope, and thus discarding pitch information. This means that intonation and prosody will not be considered in this study.

2.2 Models

Each Mel-cepstral coefficient is described statistically by its mean and standard deviation. This is a sufficient statistic if we assume the values to be normally distributed and independent. In general this is not true and multivariate models are chosen instead. This study describes only normal distributed models, but all the methods are directly applicable to the multivariate case. Given the dynamic properties of speech, each phone is divided into an initial, intermediate and final part, resulting in three independent statistics, as usually done in ASR. Since the speech material was not phonetically annotated, the estimation process is based on the expectation maximization (EM) algorithm and relies on pronunciation rules defined in a lexicon. When computing independent statistics on different subsets of data, the resulting means and variances can be interpreted as representing each subset they were extracted from. In principle any subdivision can be considered to the limit of individual speakers

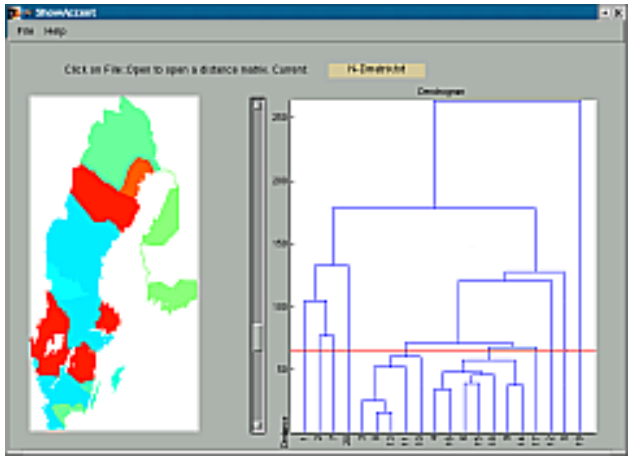

Figure 2: The ShowAccent interface. The map on the left side displays the geographical location of the clusters for a distance level specified by the slider. On the right side the *dendrogram* is a compact representation of the clustering procedure

(so called "lects"). A limitation to this is set by the reduced amount of data that can affect the efficiency of the estimation algorithms and the significance of the result.

2.3 Clustering

If we describe the means and variances as points in a metric space, clustering methods provide means for letting natural groups emerge from the data. That is, if we start from a fine subdivision of the data, we can group the instances that, after statistical analysis are similar to each other. In the current study, for example, the fine subdivision corresponds to accent areas and each mean/variance vector describes an allophone for the corresponding area. The clustering procedure then finds which allophones, in the conditions so far described, are more similar to each other and thus can be merged.

The agglomerative hierarchical clustering method [6] used in this study starts form N different elements and iteratively merges the ones that are closest according to a similarity criterion, in our case the distance between two normally distributed stochastic vectors. This can be defined by the *Bhattacharyya distance* [7]:

$$D_{batt} = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

Where M_i is the mean vector for class i and Σ_i its covariance matrix. Besides the mathematical formulation, it is interesting to consider some of its properties. Figure 1 shows a one-dimensional example: considering (a) and (c) we can see that, while the euclidean

ISBN 1-876346-48-5

I (15,16,17,18) South Swedish	South Swedish diphthongization (raising of the tongue, late beset rounding of the long vowels), retracted pronunciation of /ɹ/, no supra-dentals, retracted pronunciation of the fricative /ɧ/. A tense, creaky voice quality can be found in large parts of Småland.
II (10,11,12,13,14) Gothenburg, west, and middle Swedish	Open long and short /ɛ/ and (sometimes) /œ/ vowels (no extra opening before /ɹ/), retracted pronunciation of the fricative /ɧ/, open /ɔ/ and /l/.
III (8,9) East, middle Swedish	Diphthongization into e/ɛ in long vowels (possibly with a laryngeal gesture), short /e/ and /ɛ/ collapses into a single vowel, open variants of /ɛ/ and /œ/ before /ɹ/ (/æ, œ/).
IV (7) as spoken in Gotland	Secondary Gotland diphthongization, long /ʊ/ pronounced as /ɔ/.
V (5,6) as spoken in Bergslagen	/ɵ/ pronounced as central vowel, acute accent in many connected words.
VI (1,2,3,4) as spoken in Norrland	No diphthongization of long vowels, some parts have a short /ɵ/ pronounced with a retracted pronunciation, thick /l/, sometimes the main emphasis of connected words is moved to the right.
VII (19) as spoken in Finland	Special pronunciation of /ɵ/ and long /a/, special /ɧ/ and /ç/, /ɹ/ is pronounced before dentals, no grave accent.

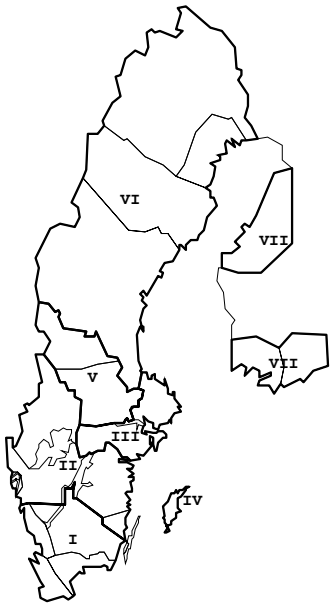

Figure 3: Summary of pronunciation variation (left) in the seven main accent areas in Sweden and part of Finland [3] and their geographic representation (right, thick borders). Arabian numbers in parenthesis and thinner borders in the figure indicate a finer subdivision that is used in this study. Phonemes that are subjected to important variations in each area are indicated in the table with the IPA symbols corresponding to their most common pronunciation.

distance is the same in this two cases, D_{batt} is larger in (a) than in (c). This is because the distance between the means is scaled by the variances and expresses the degree of overlapping of the two distributions. The same idea is confirmed looking at (a) and (b): here D_{batt} is approximately the same, while the distance between the means is different. Finally part (d) in the figure shows how in general the variances of the two variables may be different.

3 Experiments

3.1 Data

The Swedish SpeechDat FDB5000 telephone speech database [2] was used for the experiments. It contains utterances spoken by 5000 speakers recorded over the fixed telephone network. All utterances were labeled at the lexical level and a lexicon is provided containing pronunciations in terms of sequences 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.

The accent areas defined in [3], and adopted in the SpeechDat documentation, are summarized (left) and displayed (right) in Figure 3. The figure shows two degrees of subdivision of the population: the roman numbers (thick borders) refer to broader areas, while Arabian numbers (thin borders) to a finer subdivision. The last will be used in the rest of the study. Some of the properties described in Figure 3 (left) refer to prosody and will be ignored as explained in Section 2.1.

3.2 The ShowAccent Tool

To simplify the visual analysis, the graphical interface depicted in Figure 2, was developed by the author. It is based on the Matlab scripting language and some of the tools in [5]. The interface shows a map on the left side and a *dendrogram* on the right side. The last is a compact representation of the clustering procedure. The numbers on the x-axis correspond to the fine accent subdivision described in the previous section. The tree-like structure depicts the way those regions form clusters depending on the distance measure (y-axis). This representation is complete, but not straightforward: to understand how the clusters are distributed geographically, one needs to look up the region number on a map. To simplify this process, the map on the left is linked to the dendrogram by the slider in the central part. The user can display the clusters on the map at a certain distance level by moving the slider accordingly.

3.3 Results

Inspection of the natural clusters emerging from the data shows interesting properties. Many phonemes follow rules that resemble the ones given in Figure 3. This in spite of the severe reduction of information caused by the telephone line and by the feature extraction procedure, that was not designed to preserve accent

ISBN 1-876346-48-5

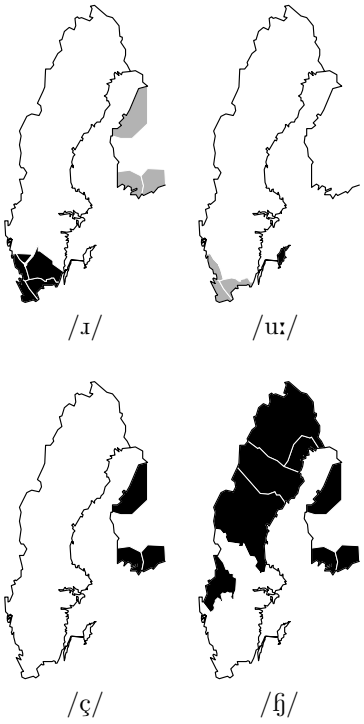

Figure 4: Four examples of pronunciation variation across Sweden and part of Finland. White, black and gray regions represent clusters where the acoustic features are homogeneous

information. As an example the geographical distribution of allophones for four phonemes are depicted in Figure 4. The phoneme /ɹ/ forms three clusters clearly corresponding to the "standard" variant in great part of Sweden (white), to the retracted pronunciation /ʁ/ in the south (black) and to the particular pronunciation in Finnish regions (gray). The vowel /uː/ forms a cluster in Gotland (black) where it is pronounced as /oː/ according to [4]. The gray area in part of the south indicates another allophonic variations of the phoneme /uː/. The fricative /ç/ as described in Figure 3 is rather homogeneous in Sweden (white), but becomes an affricate in Finnish-Swedish (black). Finally an allophone of the fricative /ɧ/ (frontal pronunciation) emerges in the northern part of Sweden and in Finland (black) while the southern and central Sweden form a different cluster, most probably associated with the more retracted pronunciation for /ɧ/. More difficult to explain is, in this case, the fact that Värmland (part of region II, see Figure 3) clusters with the north instead of the south of Sweden.

4 Conclusion

This study investigated the possibility that Melcepstral features extracted from narrow band (telephone) speech, could retain information about accent variation in Swedish. This was done using automatic speech recognition methods to derive allophonic statistical models for accent regions and clustering methods to find natural groupings of the regions. The resulting clusters proved to be largely consistent with what is known in the phonetic literature suggesting that:

- accent information can partly be extracted from the signal processing originally developed for speech recognition,
- explicitly modeling accent variation could improve the discriminative power of speech recognition models.

Acknowledgments

This research was founded by the Synface European project IST-2001-33327 and carried out at the Centre for Speech Technology supported by Vinnova (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

Part of the results here presented have been obtained within Gustaf Sjöberg's Master Thesis work [8].

References

[1] S. B. Davies and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions of Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366, August 1980.

[2] Kjell Elenius. Experience from collecting two swedish telephone speech databases. *International Journal of Speech Technology*, 3:119–127, 2000.

[3] Claes-Christian Elert. Indelning och gränser inom området för den nu talade svenskan - en aktuell dialektografi. In Edlund L.E., editor, *Kulturgränser - myt eller verklighet*, pages 215–228. Diabas, 1994.

[4] Claes-Christian Elert. *Allmän och svensk fonetik*. Norstedts Förlag, 7th edition, 1995.

[5] Jyh-Shing Roger Jang. Data clustering and pattern recognition toolbox. http://neural.cs.nthu.edu.tw/jang/matlab/toolbox/DCPR/.

[6] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241 – 254, 1967.

[7] Brian Mak and Etienne Barnard. Phone clustering using the bhattacharyya distance. In *ICSLP96, The Fourth International Conference on Spoken Language Processing*, volume 4, pages 2005–2008, 1996.

[8] Gustaf Sjöberg. Accent modeling in the Swedish SpeechDat. Master's thesis, Dept. Speech, Music and Hearing, KTH (Royal Institute of Technology), 2003.