Using Accent Information in ASR Models for Swedish

Giampiero Salvi

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology), Stockholm, Sweden
giampi@speech.kth.se

Abstract

In this study accent information is used in an attempt to improve acoustic models for automatic speech recognition (ASR). First, accent dependent Gaussian models were trained independently. The Bhattacharyya distance was then used in conjunction with agglomerative hierarchical clustering to define optimal strategies for merging those models. The resulting allophonic classes were analyzed and compared with the phonetic literature. Finally, accent "aware" models were built, in which the parametric complexity for each phoneme corresponds to the degree of variability across accent areas and to the amount of training data available for it. The models were compared to models with the same, but evenly spread, overall complexity showing in some cases a slight improvement in recognition accuracy.

1. Introduction

Statistical models for automatic speech recognition (ASR) are often described by a large number of parameters. The need for this complexity arises from the attempt to model variations caused by a long list of phenomena that disturb the classification. These are, for example, variations in channel or speaker, such as *gender*, *age*, *level of education*, *anatomical characteristics*, *emotions* and *accent*, but can also be intrinsic in the process of speech production, e.g. *co-articulation effects*.

One solution to this problem is blindly increasing the complexity of the models (e.g. in the standard HMM framework, the number of Gaussian mixture components). This process is limited by the ability of the training algorithm to efficiently estimate the model parameters, and to avoid over-training. A possibly more successful approach would be to isolate (when known) the source of variation, and build different models according to the state of the source in each particular case. One example of this approach is the creation of context dependent models (di- and tri-phones) [1],[2], that attempt to reduce/use the effects of co-articulation.

This study intends to apply the second approach to variations introduced by speaker accent. One aim is to study if the accent information, provided by phoneticians, corresponds to acoustic variations retained by the features that are usually employed in ASR. This is not obvious because the feature extraction procedure commonly used [3] was designed in the attempt to optimize phoneme discrimination, possibly at the expense of a detailed representation of other characteristics of the speech sounds.

The other aim is to investigate if models built on the base of accent information can outperform models of the same complexity that ignore it. In particular we will test the possibility to use accent information in order to allocate more efficiently the resources in the model design.

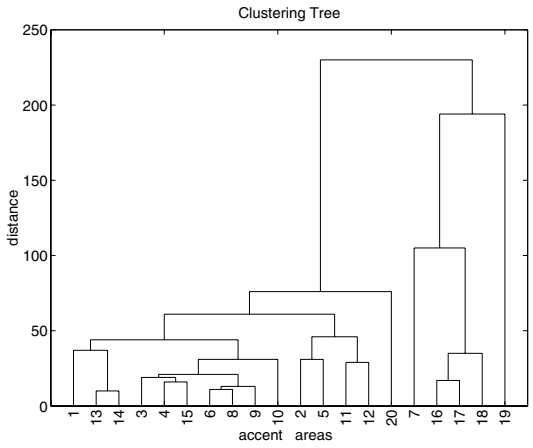

Figure 1: Clustering tree for the vowel [ɔ]. The x-axis shows the allophones for each accent area. As the distance that is allowed between members of the same cluster is increased, the allophones form fewer and larger groups, until they all merge in a single cluster (top).

2. Method

A standard paradigm for ASR is employed in this study: 10ms spaced Mel frequency cepstrum coefficients (MFCCs) are modeled by three states left-to-right phonetic HMMs with Gaussian mixtures as state-to-output probability estimators. At first M allophonic HMMs are assigned to each of the N phonemes, where M is the number of accent areas defined in the database. In the following we will refer to the set of parameters for each model as λ_{ij} with $i \in [1, N]$ as phoneme index and $j \in [1, M]$ as accent index, while b_{ij} will indicate the set of parameters defining the state-to-output distribution (i.e. ignoring the transition probabilities). The model parameters λ_{ij} were estimated, independently, on the data corresponding to each accent j, using embedded Baum-Welch training. The state-to-output distributions of the allophonic models, that are used as a base for the following process, were single Gaussian distributions but the methods here described can be easily extended to the case of mixture of Gaussian components.

2.1. Clustering

If we consider the model parameters for each HMM as spanning a D dimensional space, each point of this space represents an acoustic realization of the particular sound modeled by the HMM. After training, the points reached by each allophonic model b_{ij}, represent the average acoustic realization of the phoneme i in the particular subset of the population j. If we

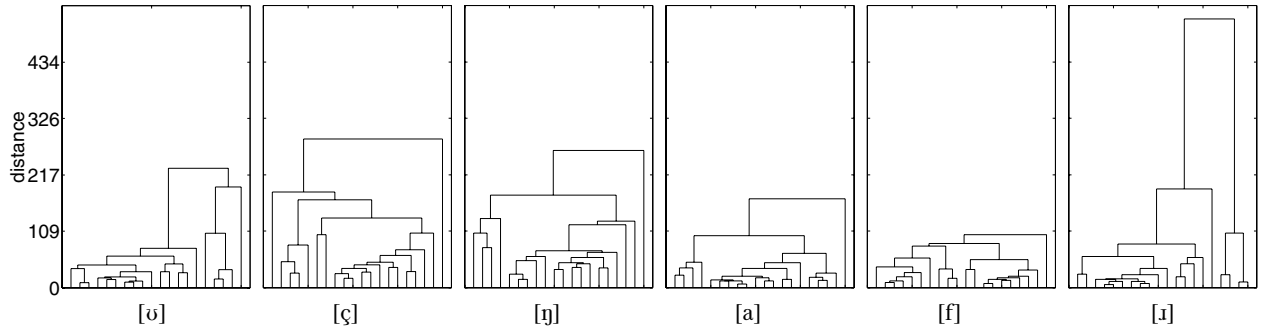

Figure 2: Example of comparison between clustering trees for six phonemes. Phonemes are indicated with the IPA symbols corresponding to their most common pronunciation in Sweden. As expected [ɪ] is the phoneme that varies most in the selection. Particularly expensive, in terms of distance associated with the resulting cluster, is the step that merges the last four allophones on the right (southern Sweden) to the rest.

augment the space with a metric we can measure the pairwise acoustic (dis)similarities in the pronunciation of a phoneme between two of accent areas. The distance measure chosen for this study is the *Bhattacharyya distance* [4] defined as:

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

Where M_i and Σ_i are the parameters (mean vector and covariance matrix) defining a generic multidimensional Gaussian distribution. The mean Bhattacharyya distance over the states of the HMMs, in conjunction with agglomerative hierarchical clustering [5] with complete linkage, is used to define the strategy employed to merge allophonic realizations of each phoneme. The resulting tree structure of Figure 1 defines at each level the allophones b_{ij} that should be merged if the distance d (on the y-axis) is allowed between members of the same cluster c_{ik}. At level 0 each allophone defines its own cluster. At each step (change of level) the two closest clusters in the previous level are merged, and the distance between the new formed and the original clusters is updated with the maximum distance between the original clusters and the members forming the new cluster. At level M, where M is the number of initial allophones, all models are merged into one.

2.2. Definition of accent "aware" models

The use of accent dependent models poses problems in the decoder design. While context information can be inserted as simple rules in the recognition network, the areas of accent homogeneity for each phoneme after clustering are in general overlapping, preventing a simple rule based solution. The method proposed in this study uses the information gathered in the clustering phase, as well as information on the amount of training data for each allophone, to perform a Gaussian selection procedure and define accent "aware" models. With this term we indicate a set of accent independent Gaussian mixture models, in which the complexity (number of Gaussian components) is decided, for each phoneme, by its variability in different areas and the amount of training data available for it. Moreover the mixture components of the resulting models set are obtained selecting, from the pool of accent dependent Gaussian distributions b_{ij} that are members of the same cluster, the ones that correspond to the largest amount of data. The algorithm is depicted in Figure 3: the clustering procedure described in the previous section constitutes the base for this procedure. For each phoneme, and each level l in the corresponding tree, the available parameters are: the definition of the clusters c_{ik} at level l, based on the original accent dependent allophones; the largest distance between elements in the same cluster; and the amount of training data for each cluster. As indicated in Figure 3, the procedure starts assigning one single cluster to each phoneme. This corresponds to being at the root (top) of the clustering trees in Figure 2. At each iteration in the while loop, the phoneme i containing the optimal cluster to be split is selected (FINDBESTCLUSTER). The optimality criterion is defined to maximize the reduction in inter-cluster distance when splitting a particular cluster. The splitting is done only if the resulting clusters correspond to a sufficiently high amount of training data; otherwise the phoneme is excluded from further processing. The loop stops either when the required number of clusters is reached or when no cluster have sufficient data to be split. Finally, for each cluster c_{ik}, the allophone b_{ij} corresponding to the largest population is chosen as representative of c_{ik}, and the corresponding Gaussian distribution is copied to the model for the phoneme i in the new model set. As an example, looking at Figure 2, the procedure will select the phoneme [ɪ] at the first iteration and split it into two allophones (that by inspection correspond to southern and rest of Sweden). Next to be split will be [ç] into one allophone for the Finnish and one for the Swedish pronunciation. Then in a sequence [ŋ], [ʊ], [ɪ] again, and so forth.

GAUSSSELECT (*finalNumClusters,minData*)
1 start with one cluster for each phoneme;
2 **while** *numClusters* $<$ *finalNumClusters* **do**
3 *bestC* $\leftarrow$ FINDBESTCLUSTER(*listOfActivePhonemes*)
4 **if** *data after splitting bestC* $>$ *minData* **then**
5 SPLITCLUSTER(*bestC*)
6 *numClusters* $\leftarrow$ *numClusters* $+ 1$
7 **else**
8 remove the phoneme i from the *listOfActivePhonemes*
9 **endif**
10 **endwhile**
11 **for** $k \leftarrow 1 \ldots$ *numClusters* **do**
12 find allophone $b_{ij} \in$ *cluster(k)* with largest amount of training data
13 copy its Gaussian to the model for phoneme i
14 **endfor**

Figure 3: The Gaussian selection algorithm

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.	1044
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].	1098
III (8,9) East, middle Swedish	Diphthongization into [ɐ/ɛ] in long vowels (possibly with a laryngeal gesture), short [ɐ] and [ɛ] collapses into a single vowel, open variants of [ɛ] and [œ] before [ɹ] ([æ, œ]).	1332
IV (7) as spoken in Gotland	Secondary Gotland diphthongization, long [ʊ] pronounced as [ɔ].	76
V (5,6) as spoken in Bergslagen	[ɵ] pronounced as central vowel, acute accent in many connected words.	307
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.	975
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.	25

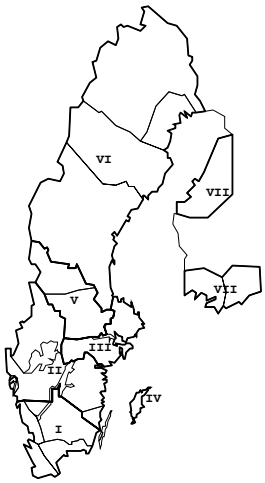

Figure 4: Summary of pronunciation variation (left) in the seven main accent areas in Sweden and part of Finland [6] and their geographic representation (right, thick borders). Arabian numbers in parenthesis and thinner borders in the map indicate a finer subdivision. Phonemes that are subjected to important variations in each area are indicated in the table with the IPA symbols corresponding to their most common pronunciation. The last column contains the number of speakers in the database for each area.

3. Experiments

3.1. Data

The Swedish SpeechDat FDB5000 telephone speech database [7] 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 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.

3.2. Clustering experiments

One of the aims of this study is to verify whether the information about pronunciation variations, retained by the speech processing used in ASR, is consistent with what is found in the phonetic literature. The reference adopted in this study is the work by C.-C. Elert [6] that is summarized in the SpeechDat documentation, as displayed in Figure 4. 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 subdivision type was used in the study as a starting point. The number of speakers is more balanced when considering this subdivision compared with the numbers showed in Figure 4, even though some areas (e.g. Stockholm) are over-represented in the database. Some of the properties described in Figure 4 (left) refer to prosody and will be ignored as 1) pitch information is discarded by the feature extraction procedure, 2) the method here described doesn't consider the timing information contained in the transition probabilities.

In order to simplify the visual analysis of the clustering data, we developed a graphical tool [8]. This application links the clustering tree of Figures 1 and 2 to a map that shows the geographic distribution of the clusters at each level in the tree. The level can be chosen interactively by the user by means of a slider.

3.3. Recognition experiments

Our second aim is to test whether the phonetic models obtained with the procedures described in this paper perform better than standard models with the same complexity. The baseline in these experiments are the monophones trained with the RefRec system [9]. This is a set of scripts that implement a standard training procedure based on the HTK toolkit and the SpeechDat database format. The test task used in the experiments consists of small vocabulary isolated phrase recognition (SVIP). The test set consists of the "A" corpus, containing 1481 utterances spoken by 500 speakers. Each utterance contains one of 30 isolated application words or short phrases.

4. Results

4.1. Clustering results

Inspection of the natural clusters emerging from the data shows interesting properties. Many phonemes follow rules that resemble the ones given in Figure 4. 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 information. As an example the geographical distribution of allophones for four phonemes are depicted in Figure 5. 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 [ʉː] forms a cluster in Gotland (black) where it is pronounced as [oː] according to [10]. The gray area in part of the south indicates another allophonic variation of the phoneme [ʉː]. The fricative [ç] as described in Figure 4 is rather homogeneous in Sweden (white), but is 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 unexpected is, in this case, the fact that Värmland (northern part of region II, see Figure 4) clusters with the north instead of the south of Sweden.

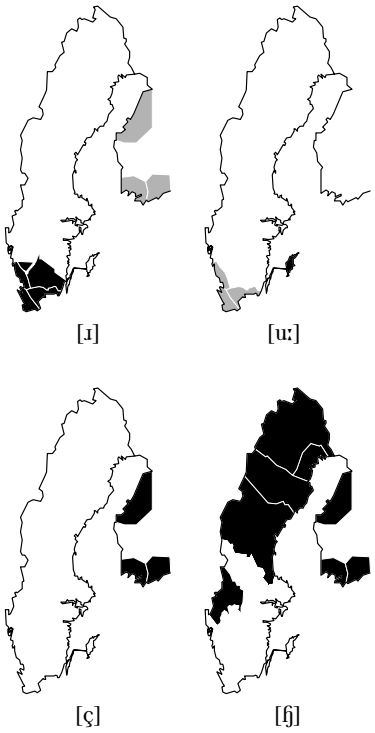

Figure 5: Four examples of pronunciation variation across Sweden and part of Finland. White, black and gray regions represent areas of acoustic homogeneity emerging from the clustering procedure.

4.2. Recognition results

Table 1 displays some preliminary results. The accent "aware" models perform better than RefRec with the same total number of Gaussian Mixture Components (GMCs) in the case of 300 GMCs, but further retraining decreases performance. In the case of 600 GMCs, the performance of our models is worse than RefRec.

5. Conclusion

This study showed how Mel-cepstral features extracted from narrow band (telephone) speech retain information about accent variation in Swedish. The way these variations affect the parameter estimation in models for ASR is consistent with what is known in the phonetic literature. An attempt to apply this information to improving ASR models has shown encouraging results, even though further refinement to our methods seem to be needed. Possible improvements in the Gaussian selection procedure may be: a more efficient use of the information on the amount of data for each cluster; substituting the selection of the most representative Gaussian distribution within one cluster with the computation of a new distribution for the cluster as average over its members.

# GMCs	RefRec	acc_0	acc_1	acc_2	acc_3
300	95	**89**	92	97	90
600	**72**	93	-	-	-

Table 1: Number of recognition errors over 1481 utterances, depending on the number of Gaussian mixture components (GMCs) in the model sets. RefRec models are mono_2_2 and mono_4_2 where the first number is the number of GMCs per state (corresponding to a total of 300 and 600 GMCs respectively) and the second the number of retraining iterations The accent "aware" models are indicated with acc_i where i is the number of retraining iterations after the Gaussian selection procedure.

6. 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 [11].

7. References

[1] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context.dependent modeling for acoustic-phonetic recognition of continuous speech," in *IEEE International Conference of Acoustics Speech and Signal Processing*, vol. 10, April 1985, pp. 1205–1208.

[2] K.-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions of Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, April 1990.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions of Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[4] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *ICSLP96, The Fourth International Conference on Spoken Language Processing*, vol. 4, 1996, pp. 2005–2008.

[5] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241 – 254, 1967.

[6] 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.

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

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

[9] B. Lindberg, F. T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, "A noise robust multilingual reference recogniser based on SpeechDat(II)," in *6th Internetional Conference on Spoken Language Processing*, vol. III, 2000, pp. 370–373.

[10] C.-C. Elert, *Allmän och svensk fonetik*, 7th ed., Norstedts, Ed. Norstedts Förlag, 1995.

[11] G. Sjöberg, "Accent modeling in the Swedish SpeechDat," Master's thesis, Dept. Speech, Music and Hearing, KTH, Stockholm, Sweden, 2003.