# Improving the Discrimination between Native Accents when Recorded over Different Channels

*Tingyao Wu [a], Dirk Van Compernolle [a], Jacques Duchateau [a], Qian Yang [b], Jean-Pierre Martens [b]*

[a] K.U.Leuven – Dept. ESAT, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{tingyao.wu, dirk.vancompernolle, jacques.duchateau}@esat.kuleuven.ac.be
[b] Ghent University – ELIS - RUG
{qian.yang, jean-pierre.martens}@elis.rug.ac.be

10.21437/Interspeech.2005-714

## Abstract

Acoustic differences between native accents may prove to be too subtle for straightforward brute force techniques such as blindly clustered Gaussian mixture model (GMM) classifiers to yield satisfactory discrimination performance while these methods work well for classifying more pronounced differences such as language, gender or channel. In this paper it is shown that small channel differences are easier to detect by such coarse classifiers than native accent differences. Performance of native accent classification can be improved considerably by incorporating the knowledge of the underlying phoneme sequence and using phoneme specific GMMs. Further improvements are obtained if optimal feature selection is combined with the phoneme dependent GMMs, resulting in usage of less than 10% of the original features. The presented methods result in a reduction of more than 40% in relative error rate in a 5-class classification task.

## 1. Introduction

Speaker variability due to gender, accent, speaking style, age, etc, contributes greatly to the problem of speaker independent recognition. Speaker adaptation may improve performance significantly, however, it requires large amounts of data. Modeling and adapting to properties of groups of speakers makes adaptation quicker and more robust; if the adaptation population is large enough it may be worthwhile to make group specific models as is often done for both genders. Similarly, building accent dependent models [1] can reduce speaker independent error rates, assuming a fast and reliable accent identification method.

In this paper we investigate methods for native accent identification (AID). Some techniques widely used for language identification (LID) seem to be suitable for AID under the condition that they rely on pronunciation information only; others that incorporate linguistic information (eg. [2]) would not be suitable for AID. Along these lines Chen [3] used blindly clustered GMMs to identify four domestic Chinese accents and obtained improved performance with an increasing number of mixtures. In our experiments with the Flemish GoGeN database, blindly clustered mixtures performed very poorly. We see two possible explanations for this discrepancy: (i) the intrinsic differences between the dialects studied in [3] are much larger than the ones present in our CoGeN database; (ii) the experiments in [3] were influenced by a correlation between accent and recording condition, which can easily be overlooked if recordings are made in different regional locations.

Our experience tells us that a different approach is required for AID than for LID. The importance of speaker normalization in our former work [4, 5] suggests that accent differences are considerably more subtle than individual speaker or language differences. One of the key properties of accent differences is that they are definitely prominent for a few sounds in a language but might as well be negligible for the majority of sounds [4]. Hence classifiers relying on the global speech distribution may not be the most suitable approach for native AID.

In this paper we show that in such a global classifier accent properties may be dominated by slight changes in recording channels. Even when using standard normalization schemes such as cepstral mean subtraction (CMS), channel information still dominates accent information in a blindly clustered GMM identification system. The above suggests that the large number of speech frames that contain no accent specific information will introduce a prohibitive amount of noise in any accent classifier relying on the global speech distribution. Therefore we should only consider those frames and phonemes that are likely to contribute to accent discrimination. Moreover only a few features per phoneme may exhibit accent specific differences in a statistically significant way [5]. In order to verify the above intuitive statements three different accent identification schemes, all making use of some form of a GMM-classifier but with varying degrees of phoneme and feature selection are compared on a task of classifying 5 native Flemish accents.

## 2. System description

The primary input feature stream for all methods consists of typical speech features, i.e. 12-th order cepstra and energy after the application of CMS. The three different accent identification schemes are now described in detail.

### 2.1. AID1: GMM based Classifier using the global speech distribution

AID1 is a blindly clustered GMM based classifier building a model of the global speech distribution. This is a simple approach often used for other speaker group classification problems including gender and language identification. The great advantage of this method is its simplicity. Speech data from the different accents are clustered and models are trained using the EM (expectation-maximization) algorithm. The accent of a

This research was supported by the Research Fund for Scientific Research Flanders (FWO-project G.0008.01) and by the Research Fund K.U.Leuven OT/03/32/TBA.

testing speaker $r^*$ with observation set $\boldsymbol{F}$ is recognized as:

$$r^* = \arg \max_{1 \le r \le R} \sum_{\boldsymbol{F}} \log(P(\boldsymbol{F}|G_r)), \quad (1)$$

where $G_r$ is the model for accent $r$, and $R$ is the number of accents.

### 2.2. AID2: GMM based Classifier with Phonetic information

AID2 no longer relies on a single model of the global speech distribution. Individual GMMs are constructed for all phonemes and during classification incoming frames need to be phonetically labeled such that they can be scored by the relevant phoneme model only. Instead of one large GMM for all data, an accent is thus represented by a collection of small GMMs for each individual phoneme. In practice our phoneme models use a single gaussian distribution. The block diagram of a GMM-based classifier with phonetic information is depicted in Figure 1 excluding the blocks in the dashed frame.

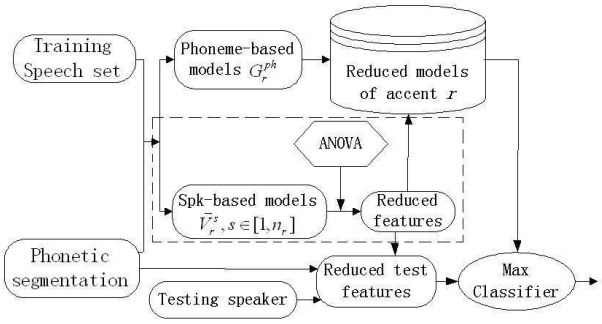


Figure 1: Diagram of GMM-based AID2 and AID3 system. Parts within the dashed frame are extra components for the AID3 system. $G_r^{ph}$ represents GMMs of the phoneme $ph$ for accent $r$, and $\vec{V}_r^s$ is the speaker model for speaker $s$ in accent $r$

### 2.3. AID3: GMM based Classifier with Feature Selection

AID3 is a refined version of AID2 in which for each phoneme only the most distinctive features are retained. For AID3 the blocks within the dash frame in Figure 1 are also included. An accent is still represented by a collection of small(er) GMMs for each individual phoneme. Moreover different selected subsets of features are used for the individual phoneme models.

The feature selection technique is based on a method we developed for finding the most relevant features in AID in [5]. According to that approach we construct a super vector for each speaker, which is called the speaker-based model in Figure 1, containing the mean parameters for each phoneme. From these, accent specific super vectors are computed as the mean across speakers for a given accent. Using a phonetic alphabet of 38 symbols this results in a super vector of dimension 494. Given speaker and accent super vectors, analysis of variance (ANOVA) is then used to determine the significance of each feature assuming that these super vectors are the inputs to the classifier. While the input to the GMMs is not the speaker super vectors but the individual frames, both classifiers are quite related to one another and it seems plausible that feature relevance will carry over from one classifier to another. Thus only the most significantly discriminative features selected by the ANOVA procedure described above are retained in the AID3 classifier. Details on the use of the ANOVA method are given in Appendix A.

## 3. Experiments

### 3.1. Database

All experiments are performed on the CoGeN database, which contains 101 male speakers and 73 female speakers. Speakers are grouped in accent clusters corresponding to the five Flemish provinces, namely Antwerp (A), Brabant (B), Limburg (L), East-Flanders (O) and West-Flanders (W). Dialect studies show that the Flemish provinces correspond fairly well with dialect regions [6].

Speakers were asked to read five paragraphs of standard Dutch, which yielded about 5 minutes of speech per speaker. All recordings were made with the same recording equipment but in three different places, leading to possibly different channel effects. These three places are denoted as CH1, CH2 and CH3 respectively. The distributions of accents and channels are shown in Table 1.

| | CH1 | CH2 | CH3 | sum |
|---|---|---|---|---|
| A | 24 | 16 | 2 | 42 |
| B | 15 | 10 | 1 | 26 |
| L | 12 | 21 | 1 | 34 |
| O | 3 | 8 | 25 | 36 |
| W | 13 | 13 | 10 | 36 |
| sum | 67 | 68 | 39 | 174 |

Table 1: The distribution of accents and recording channels in CoGeN database

All speech was aligned at the phoneme level by the speaker-independent large vocabulary continuous speech recognition system developed by the ESAT speech group at K.U.Leuven [7]. This alignment was obtained either by forced alignment (Viterbi) relying on a written transcript or by a phoneme recognizer using a phoneme trigram language model with approximate phoneme recognition rate of 68%. Unless stated otherwise, phoneme labels obtained from the forced alignment are used.

Because of the limited number of speakers for each accent, a leave-one (speaker)-out scheme is used for evaluation in all experiments.

### 3.2. Phonetically blind GMMs

We first study the performance of the phonetically blind GMMs to identify the accents. Poor performance, shown in Table 2 is obtained whatever the number of mixtures that is used. A possible reason is that as the variation of accents is pretty subtle, the channel effects disturb the discrimination of accents, even after CMS is used.

### 3.3. Channel effects

To verify the channel dominance assumption, a channel recognition experiment is set up: three GMMs are trained on data recorded at each place, and the speech of a testing speaker is identified to one of three channels. Table 3 gives pretty good channel recognition rates, indicating that GMMs of the full speech distribution perform better on channel identification than on accent identification. From the prior channel and accent distributions, we can calculate that a 29.1% AID performance is possible using the prior distribution and perfect channel recognition.

| number of gaussian mixtures | 1 | 4 | 16 | 64 | 256 |
|---|---|---|---|---|---|
| AID accuracy(%) | 27.0 | 20.1 | 23.6 | 25.3 | 23.6 |

Table 2: Accent identification rates for blindly clustered GMM

| number of gaussian mixtures | 1 | 4 | 16 | 64 | 256 |
|---|---|---|---|---|---|
| channel recognition(%) | 74.6 | 69.5 | 72.9 | 76.3 | 78.0 |

Table 3: Channel identification rates for GMM

So, with a channel identification accuracy of 75%, an AID of 25% would be possible by just using prior knowledge. This compares to an intuitive 20% chance level in our 5-choice test. This confirms that the results of AID1 (Table 2) show no accent identification whatsoever and only some channel information.

### 3.4. AID with phonetic information and feature selection

Methods AID2 and AID3 rely on the phonetic labeling of the input speech, as generated by a Viterbi alignment. Figure 2 shows AID accuracy in function of the number of features selected in AID3. Selecting all features ($D = 494$), which corresponds to method AID2, yields an accuracy of 41.4%, compared to 27.0% (Table 2) without phonetic information. With the procedure of feature selection, the AID3 system performance boosts to 65.5%, when only 30 features are selected from the whole feature set. The corresponding confusion matrix is shown in Table 4. Figure 2 also indicates that the exact number of features that is selected is not very critical, as long as it is between 25 and 75. In Figure 3, we plot the number of phonemes and cepstral dimensions appearing in the selected feature set. The phonemes, or cepstral dimensions occurring in the reduced feature set are believed to contribute to the discrimination of accents. Figure 3 illustrates that when the size of the reduced feature set is equal to 30, the number of phonemes that are involved is 15 out of 38, and the number of dimensions is 10 out of 13. Thus, there are some phonemes and cepstral dimensions which are not used at all.

It is interesting to show which phonemes and cepstral dimensions contribute most to the identification task at optimal performance (Figure 4). The *x*-axis indicates the phonemes and the *y*-axis represents the different cepstral dimensions. The darkness corresponds to the number of times a feature is selected in the 174 leave-one-out trials that were conducted in this experiment. We can see that although most of the discrimination capability is contributed by vowels, some non-vowel phonemes, even some plosive phonemes (for example, /k/ and /p/), play an important role in accent identification. In the cepstral domain, the most selected cepstral coefficients are concentrated at low dimensions.

### 3.5. Discussion

Because of tiny differences among accents, blindly clustered GMMs for native accents may not model the accents well. Our experiments show that even after CMS, the channels can be easily retrieved by a set of GMMs, whereas the accents are not retrieved at all. We argue that this is due to the fact that acoustic differences among accents are smaller than the differences

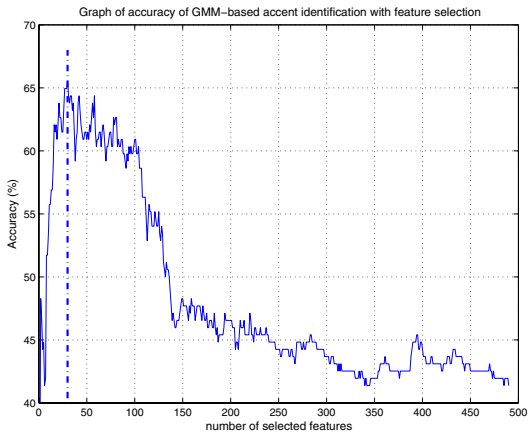


Figure 2: Accuracy of GMM-based accent identification with feature selection

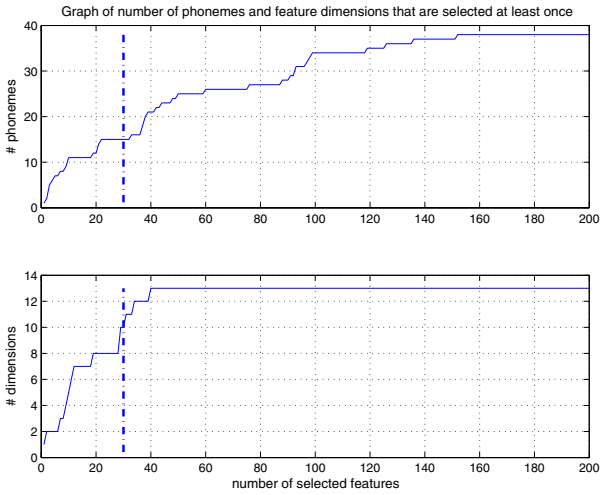


Figure 3: Number of selected phonemes and feature dimensions that are selected at least once

originating from the channel. From AID2 system, we found that phonetic information helps to enlarge the differences among accents. To verify this statement further, we use a phoneme recognizer, instead of Viterbi alignment, to generate frame labels and re-run AID2 and AID3 systems. 37.4% and 52.9% (when 70 features are selected) AID accuracies are achieved by these two systems respectively, compared to 41.4% and 65.5% obtained through Viterbi alignment.

Although incorporating phonetic information and feature selection improves the discrimination among accents, inevitably we still find the shadow of channel effects. For example, in Table 4, the possible reasons for easy confusion of speakers of accents *A* and *B* are not only that they are neighbors geographically, but also that most of their speakers were recorded through the same channel. The same also holds even more for the speakers of accents *O* and *W*.

Our experiments show that one can talk about accent relevant and accent irrelevant phonemes, and that in combination with GMMs, the features of accent irrelevant phonemes mask the contributions of the relevant features. Another observation is that the most accent sensitive features lie in the low order

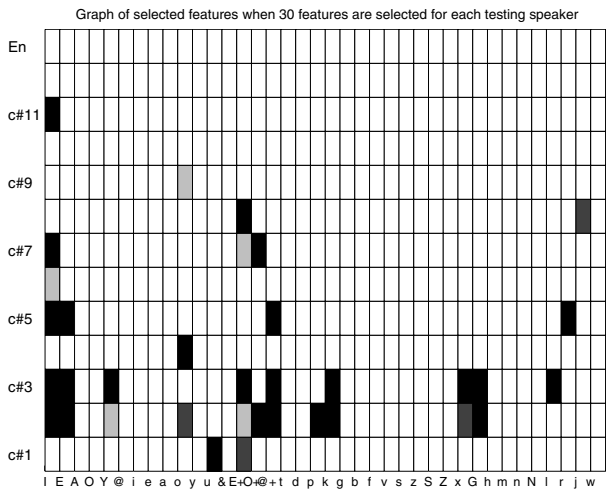


Figure 4: Number of times phonemes and dimensions that are selected (gray scale) when optimal performance (=65.5%) is reached, and the number of selected features is 30

| | recognized accent | | | | |
|---|---|---|---|---|---|
| | A | B | L | O | W |
| A | **32** | 8 | 1 | 0 | 1 |
| B | 11 | **11** | 0 | 2 | 2 |
| L | 3 | 1 | **24** | 1 | 5 |
| O | 1 | 4 | 3 | **20** | 8 |
| W | 2 | 2 | 2 | 3 | **27** |

Table 4: Confusion matrix when optimal performance (=65.5%) is reached. The rows represent the true class, the columns the recognized class.

cepstral coefficients, whereas the high orders are hardly ever selected.

## 4. Conclusions

In this paper we presented three AID systems with varying degrees of phoneme and feature selection, which illustrated the difference across native accents might be an order magnitude smaller than the difference among channels. Such that blindly clustered GMM based classifiers do not model accents well, whereas with the use of phoneme-specific acoustic features and a selection of accent relevent features, the discrimination among accents is revealed.

## 5. Appendix A: some considerations on ANOVA

ANOVA is a known statistical tool which checks whether the means of samples in different groups are equal. To apply ANOVA for the feature selection task, we assume that if the mean of a specific feature in accents significantly differs, this feature will contribute to accent recognition [5]. Consider the $i^{th}$ ($1 \leq i \leq D$) feature in the speaker models. Based on the assumptions that the observations are normally distributed, and the variances are equal in all $R$ accents, the null hypothesis of ANOVA is

$$H_0 : \mu_{1,i} = \mu_{2,i} = \ldots = \mu_{R,i}, \tag{2}$$

where $\mu_{r,i}$ is the mean of $i^{th}$ feature for accent $r$. The alternative hypothesis $H_1$, which then corresponds to the situation that at least one couple $(\mu_{j,i}, \mu_{k,i})$, $j, k \in [1, R]$ satisfies $\mu_{j,i} \neq \mu_{k,i}$, indicates that the $i^{th}$ feature is significantly different among accents. $H_0$ is rejected if the p-value:

$$p_i = P(F_{R-1,N-R} > F - ratio_i) \tag{3}$$

is smaller than a chosen level of significance $\alpha$, where $N$ is the number of speakers in all accents. The $F - ratio_i$ can be thought of as a measure of how different the means are, relative to the variability within each class. The larger this value, the greater the likelihood that the differences between the means are due to something other than chance.

$$F - ratio_i = \frac{1}{R-1} \sum_{r=1}^{R} \{ \frac{\sqrt{n_r}(\mu_{r,i} - \overline{\mu}_i)}{\hat{\sigma}_i} \}^2 \tag{4}$$

where $\overline{\mu}_i$ is interpreted as the global mean, ${\hat{\sigma}_i}^2$ is the estimated pooled class variance, and $n_r$ is the number of speakers in accent $r$.

ANOVA is sensitive to non-normal distributions. If the assumptions, namely a normal distribution for each feature for each accent and equal variances across accents for each feature, are not satisified, it may give inaccurate p-values. Normal quantile plots, which check how the actual data fits the ideal data from normal distribution, and Barlett's test, which tests homoscedasticity accross the accents, were used to check the normality and equal variances for each feature respectively. On a significance level $\alpha = 0.05$, experimental results show that 84.8% of the features satisfy the assumption of normality, and 95.0% of those features also show equal variances on CoGeN database, assuring the validity of ANOVA's prerequisite.

## 6. References


[1] C. Huang, E. Chang, J. L. Zhou, and K. F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition," in *Proc. ICSLP*, vol. 3, Beijing, China, Oct 2000, pp. 818–821.

[2] K. Berkling, C. Cleirigh, J. Vonwiller, and M. Zissman, "Improving accent identification through knowledge of english syllable structure," in *Proc. ICSLP*, vol. 2, Sydney, Australian, Dec 1998, pp. 89–92.

[3] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using gaussian mixture models," in *Proc. ASRU*, Trendo, Italy, 2001, pp. 343–346.

[4] P.-J. Ghesquière and D. Van Compernolle, "Flemish accent identification based on formant and duration features," in *Proc. ICASSP*, Orlando, USA, May 2002, pp. 749–752.

[5] D. Van Compernolle, P. Ghesquière, T. Wu, Q. Yang, and J. Martens, "Feature subset selection for improved accent identification," *Speech Comm.*, 2005, submitted.

[6] R. Van Hout and H. Münstermann, "Linguïstische afstand, dialekt en attitude," *Gramma*, no. 5, pp. 101–123, 1981.

[7] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Comm.*, vol. 24, no. 1, 1998.