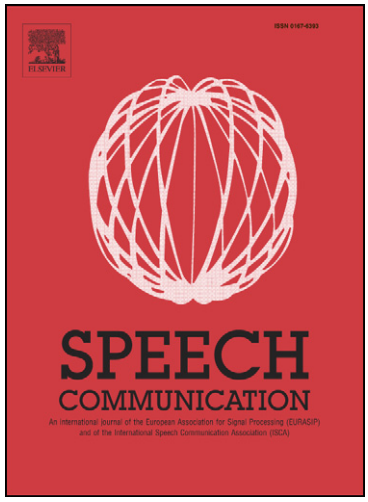
# Feature Subset Selection for Improved Native Accent Identification

Tingyao Wu, Jacques Duchateau, Jean-Pierre Martens, Dirk Van Compernolle




Feature Subset Selection for Improved Native Accent Identification

Tingyao Wu, Jacques Duchateau, Jean-Pierre Martens, Dirk Van Compernolle

PII: S0167-6393(09)00132-0
DOI: 10.1016/j.specom.2009.08.010
Reference: SPECOM 1829

To appear in: *Speech Communication*

Received Date: 7 November 2008
Revised Date: 13 August 2009
Accepted Date: 24 August 2009

Please cite this article as: Wu, T., Duchateau, J., Martens, J-P., Compernolle, D.V., Feature Subset Selection for Improved Native Accent Identification, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.08.010

# Feature Subset Selection for Improved Native Accent Identification[☆]

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

[a]*ESAT - PSI, Katholieke Universiteit Leuven, Belgium*
[b]*ELIS, Ghent University, Belgium*

**Abstract**

In this paper, we develop methods to identify accents of native speakers. Accent identification differs from other speaker classification tasks because accents may differ in a limited number of phonemes only and moreover the differences can be quite subtle. In this paper, it is shown that in such cases it is essential to select a small subset of discriminative features that can be reliably estimated and at the same time discard non-discriminative and noisy features. For identification purposes a speaker is modeled by a supervector containing the mean values for the features for all phonemes. Initial accent models are obtained as class means from the speaker supervectors. Then feature subset selection is performed by applying either ANOVA (Analysis of Variance), LDA (Linear Discriminant Analysis), SVM-RFE (Support Vector Machine - Recursive Feature Elimination), or their hybrids, resulting in a reduced dimensionality of the speaker vector and more importantly a significantly enhanced recognition performance. We also compare the performance of GMM, LDA and SVM as classifiers on a full or a reduced feature subset. The methods are tested on a Flemish read speech database with speakers classified in 5 regions. The difficulty of the task is confirmed by a human listening experiment. We show that a relative improvement of more than

[☆]This research was supported by the Research Fund of the K.U.Leuven, the Fund for Scientific Research Flanders (Projects G.008.01 and G.0260.07)

[*]Corresponding author

*Email addresses:* Tingyao.Wu@esat.kuleuven.be (Tingyao Wu), Jacques.Duchateau@esat.kuleuven.be (Jacques Duchateau), Jean-Pierre.Martens@elis.ugent.be (Jean-Pierre Martens), Dirk.VanCompernolle@esat.kuleuven.be (Dirk Van Compernolle)

20% in accent recognition rate can be achieved with feature subset selection irrespective of the choice of classifier. We finally show that the construction of speaker based supervectors significantly enhances results over a reference GMM system that uses the raw feature vectors directly as input, both in text dependent and independent conditions.

*Key words:* Accent Identification, Language Identification, Feature Selection, Gaussian Mixture Model, Linear Discriminant Analysis, Support Vector Machine

## 1. Introduction

Recognizing native and non-native accents has become an important issue for many commercial applications such as information services over the telephone. Clients of these applications often have strong accents resulting in poor automatic speech recognition (ASR) performance (ten Bosch, 2000). Accent identification (AID) may alleviate this problem as it would enable the system to use both speech models and pronunciation dictionaries specific to the accent. Simply adding pronunciation variants to the dictionary does not solve the problem, since it increases the number of alternatives and tends to generate additional confusion which may worsen performance (Adda-Decker and Lamel, 1999; Cremelie and Martens, 1999). Another potential application for AID is that it may be used in speaker adaptation when a speaker model is adapted from an accent-specific acoustic model rather than a universal model due to the limited speech data for the speaker (Liu et al., 2000).

In this paper, we will study automatic ways to identify accent differences among native speakers, where a native accent is defined as "a way of speaking typical of a particular group of people and especially of the natives or residents of a region" (Merriam-WebsterOnlineDictionary, 2008). Native AID is similar to language identification (LID) and foreign accent identification. Phonotactic approaches widely used in LID are not suitable to AID due to the marginal differences between accent-specific phonetic language models. On the other hand maximum likelihood (ML) based approaches are often used in most current AID systems (Lincoln et al., 1998; Lamel and Gauvain, 1995; Chen et al., 2001; Huang and Hansen, 2006; Torres-Carrasquillo et al., 2004). These ML based approaches have in common that they treat all phonemes and features uniformly. While these methods work well in identifying native accents with non-trivial pronunciation variations, they may fail to discriminate native accents with subtle differences, as we will show in the experiments. We argue that not all phonemes contribute equally to identify an accent, and that certain phonemes only change in certain regions; for instance, in Chinese Mandarin, a single pair of phonemes /$n$/ and /ŋ/ plays an important role to distinguish the southern accent from the northern accent. We also state that not all features of a certain phoneme differ between accents. For example, in the Dutch word 'vijf' (five), the voiced fricative /v/ is devoiced by Dutch speakers and not by Flemish speakers. These phenomena imply that the differences among accents do not widely spread over all phonemes and all phoneme features. Accent irrelevant phonemes or features

may introduce noise and generate confusion in the modeling and the decision. To focus on the salient differences between native accents, selecting accent relevant features on a phoneme specific basis becomes crucial.

The usefulness of feature selection has been shown for many different classification problems especially when the number of training examples is small with respect to the dimensionality of the feature vector (Kohavi and John, 1997; Guyon and Elisseeff, 2002). These are very much the circumstances under which our accent classifier needs to operate. Hence this paper will apply, adjust and tune general pattern classification solutions to the specific problem of native accent classification.

The paper is structured as follows. In section 2 we explain the similarity and difference between LID and AID, then introduce the motivation of using feature selection in AID. Afterwards we discuss our methods for AID based on feature subset selection within a speaker model based framework in section 3. In section 4, we present the experimental results of our system in text dependent and text independent conditions, and demonstrate the advantages of feature selection in the tasks. Finally some accent issues are discussed and conclusions are given.

## 2. Identifying Languages and Accents

Language identification is about recognizing the language a speaker speaks. There are two popular techniques for doing this. One is based on phone tokenization, which attempts to model the differences in phonotactics or higher level linguistic information. Most related approaches are derived from PRLM (phone recognizer followed by language model) or PPRLM (Parallel PRLM) (Zissman, 1996; Zissman and Berkling, 2001), where the language model is generated either by one or a parallel of phone recognizers, or a GMM tokenization system (Torres-Carrasquillo et al., 2002). If specialized language knowledge (i.e. phone labeling) is given, the PPRLM approaches would be the ideal technique. But most of the time this knowledge is absent. As a result, an alternative of phone labels is to use the tokens of Gaussian mixtures generated by a GMM tokenizer. The other technique solely depends on spectral similarity. By combining a discriminatively trained GMM recognizer and the shifted delta cepstral (SDC) coefficients, this technique obtains comparable performance with the phonotactic approaches (Matejka, 2006; Burget et al., 2006; Castaldo et al., 2007; Campbell et al., 2008). However, the best result reported on LID was obtained by a fusion of these two

techniques (Torres-Carrasquillo et al., 2008), probably due to their complementarity to each other.

The family of phone tokenization approaches in LID probably can not be transplanted to the AID task directly since the variations of grammar and morphology are marginal between different accents (Shriberg et al., 2008). Oppositely, the acoustic information based LID techniques have been widely used in accent/dialect classification (Lincoln et al., 1998; Lamel and Gauvain, 1995; Chen et al., 2001; Huang and Hansen, 2006; Torres-Carrasquillo et al., 2004), most of which are applied in a maximum likelihood framework. These studies, in terms of the nativity of speakers, can be further separated into two types, namely foreign AID (Teixeira et al., 1996; Berkling et al., 1998) and native AID (Hansen et al., 2004; Huang and Hansen, 2006; Shen et al., 2008). The foreign AID is a task on either verifying the nativity/non-nativity of a speaker or identifying the native language (L1) from non-native speech (L2), whereas the native AID identifies the accent of a native speaker from his speech. These two tasks, as well as their solutions, are quite similar, except that some features useful in foreign AID may not work in native AID (For example, Shriberg (Shriberg et al., 2008) discovered pauses were particularly efficient to detect low-proficiency non-native speakers, which is not the case in the native AID).

In this paper, we focus on identifying native accents. The difficulty of native accent identification may vary considerably from case to case depending on the granularity of the accent classes. For instance, Chen (Chen et al., 2001) identified four Chinese Mandarin accents using brute force GMMs and obtained about 86% accuracy; when we apply the same straightforward GMM approach for distinguishing Dutch (as spoken in the Netherlands) from Flemish (as spoken in Flanders, Belgium - sometimes also referred to as 'Southern Dutch') we achieve a 83.8% accuracy (Wu, 2009). However when applying the same methodology to identify five subregions within the Flemish region only a meager 27% accuracy is obtained, which is not much better than the 20% chance level (see (Wu et al., 2005) and section 4.2). We conjecture that the big gap between the above results is mainly due to the different magnitude of pronunciation variations between the accents. Apart from intrinsic differences, speaking style (spontaneous vs. read, formal vs. informal) may play a role as well.

The magnitude of acoustic difference between accents also affects the human to distinguish accents. Concerning Dutch, Knops (Knops, 1984) investigated if listeners from Flanders and the Netherlands could identify whether

semi-spontaneous speech fragments were spoken by a person from Belgium or from the Netherlands, and, if possible, could specify the regional identity of the speaker. Country identification was nearly perfect (96%). Identification was mainly done on pronunciation ($\sim$ 90%), morphology ($\sim$ 42%) and on intonation for the Dutch listeners (44%). Identifying the region (open choice, 10 regions) was much more difficult, only 16% of all speakers were classified correctly by the Belgian listeners while 18% were classified correctly by Dutch listeners. Results are significantly better ($\sim$ 45%) if Flemish listeners only need to identify Flemish accents and Dutch listeners only Dutch accents. These results are quite in line with the machine results we mentioned before for Dutch/Flemish and confirm the huge performance gaps that are possible for different accent identification tasks even within the same language group.

Much of the linguistic research on accents and dialects deals with phenomena occurring in a specific phonemic context or word differences between two small regions. For instance, Purnell (Purnell et al., 1999) showed that listeners could identify a dialect with only one word as information (closed choice, African American Vernacular English, Chicano English and Standard American English). Another study (Thomas, 2000) showed that a single segment in a certain context might be enough to distinguish between non-Hispanic whites from central Ohio and Mexican Americans from southern Texas.

While specific phenomena can be great indicators, their occurrence might be too rare to be applied in automatic systems and a focus on more systematic distinctions will be more rewarding. A large study by Labov on regional differences of American English (Labov, 1996) showed that there were two major types of sound changes that affected the success rate in speech recognition: mergers and chain shifts. Mergers occur if in a certain region the differences between two sounds disappear. A first type of mergers are the unconditioned mergers which affect the phonemes wherever they occur. On the other hand, conditioned mergers occur in a particular phonetic context. An example of the former is the distinction between the short /o/ (eg. cot) and long open /oh/ (eg. caught). In about half of the United States and all of Canada, these sounds are pronounced the same. The second major sound changes are the chain shifts which rotate speech sounds. Two major patterns of chain shifting, which rotate the vowels of English in opposite directions for some dialects, have been identified. Labov made maps in the formant space ($F_1 - F_2$) that clearly showed these chain shifts.

All the above studies point to a few inherent difficulties that make the task of native AID quite challenging and quite different from LID: only a fraction

of the speech contains discriminative (useful) information and the distinctions may be quite subtle. Ghesquiere (Ghesquiére and Van Compernolle, 2002) selected informative normalized formant frequencies and duration to identify Flemish accents in a text dependent mode and an absolute improvement of 20.2% was obtained; a similar experiment was done by Hansen (Hansen et al., 2004), who showed that only using a few discriminative phonemes based on the Fisher criterion could boost the system from 30% to 42% for identifying seven American English dialects also in a text dependent mode. These results imply that not all phonemes are discriminative among accents and the existence of non-discriminative phonemes may heavily degrade the performance. Even for accent-relevant phonemes, their contributions may be different; some are more informative while others are less. This motivates us to systematically investigate the behavior of phonemes in the native AID task and to focus on the most accent-relevant phonemes or features during the recognition.

## 3. Method

### *3.1. Framework*

A straightforward approach to automatic speaker classification - of which accent identification is a particular case - is to train class specific GMMs on the basis of all data for the class. In recognition mode the likelihood of the input feature stream is computed for each of the models and the highest scoring model is selected. Such a scheme has shown to reach quite satisfactory results on identification tasks such as gender identification or speaker recognition.

Although such system can achieve acceptable performance on native accents with big differences, it performs poorly on discriminating subtle accents, as will be shown in the section 4.2. This indicates that the brute force GMMs are not capable of modeling tiny accent differences. As stated before, many of the input frames correspond to sounds that do not contribute to accent discrimination and ideally these frames should score identically on all accent specific GMMs. However, in practice these segments will score slightly different on the respective accent models due to random differences in these models. Therefore it is plausible that significant improvements are possible by limiting the scoring to relevant phonemes only and furthermore to discriminative features for those phonemes.

To study the behaviors of phonemes and to select the prominent phoneme features for different accents, we first build a text dependent system with the assumption that a state-level segmentation of all phonemes is known *a priori*. The segmentation is realized by a forced alignment of the speech with the phonetic transcript. The proposed native AID system is shown in Figure 1. The raw speech features, the building of the speaker model, the techniques for feature selection and the selection of classifiers are described in more detail hereafter.

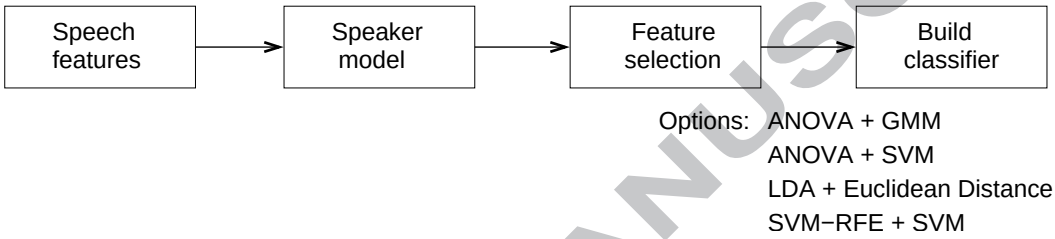



Figure 1: System overview

### *3.2. Speaker Models*

#### *3.2.1. The Raw Speaker Model*

The initial speech feature set for a 30ms-length speech frame consists of 12 mean normalized Mel scaled cepstral coefficients and the mean normalized log energy. After feature extraction, given the state-level segmentation, normally there are two ways to organize the data. One is a *frame-based model*, where the accents are represented by a collection of accent specific phoneme GMMs and the likelihood of a test segment is computed as the sum of individual frame likelihoods. The other is a *speaker-based model*, in which a speaker is represented by a supervector consisting of the averages of the speech features for each of the observed phonemes, or alternatively for the sub-phonetic states of those phonemes. For example, suppose there are $K$ phonemes in a phonetic set, and three states per phoneme in the acoustic Hidden Markov Model (HMM) based speech recognizer, a speaker-based supervector would then be composed of $D = K \times 3 \times (12\ cepstra + 1\ logEn)$ features.

A weakness of the frame-based model is that its performance depends on the number of phoneme occurrences, especially for short segments. On the other hand, the speaker-based model uses the average of the features of

each phoneme, which reduces the sensitivity of phoneme occurrences in our modeling. Therefore unless stated, we adopt the speaker-based model in our system setup. However we also present the results of the frame-based system in both text dependent and independent modes in our reference experiments.

One concern for the speaker-based model is that some phonemes may not occur in a given segment, leading to missing elements in the speaker supervector. Our fallback strategy is to replace the missing features by the overall mean value estimated from all occurrences in the training set. Using the global mean minimizes the impact of missing features. Given that for short segment lengths the probability of a feature not appearing in a speaker model is very high, using the class mean (instead of the global mean) for missing features would greatly bias the transformation, especially for infrequent phonemes.

*3.2.2. Feature Selection*

The introduction of the speaker-based model leads to the mathematical expression for the native AID problem. Suppose that we have a set of $n$ $D$-dimensional speaker models $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ ... \\ \mathbf{x}_n \end{bmatrix}_{n \times D}$ with the class label $\mathbf{Y} = \begin{bmatrix} 1 \\ ... \\ N \end{bmatrix}_{n \times 1}$, $n = \sum_{c=1}^{N} n_c$, $n_c$ the number of speaker models in the class $c$, $N$ the number of accent classes, and $\forall i \in \{1, ..., n\}$, $\mathbf{x}_i = [x_{i1}, ..., x_{iD}]$. Then the feature selection is to look for a subset $B \subseteq \{1, 2, ..., D\}$ to minimize the generation error of $\tilde{\mathbf{X}}$ given $\mathbf{Y}$ on a development set with respect to a classifier, where $\tilde{\mathbf{X}}$ represents the columns of $\mathbf{X}$ with the indices in $B$.

*3.3. Classifiers*

A classifier is used to model the variations between the speaker models of native accents. In this paper we explore three classifiers: GMM, linear discriminant analysis (LDA) and support vector machine (SVM), based on different modeling criteria.

*3.3.1. GMM*

The speaker models of an accent are modeled as linear combinations of a few Gaussians. The posterior probability of a test sample $\mathbf{x}$ is calculated as

$$f_c(\tilde{\mathbf{x}}) = \sum_{j=1}^{M} \lambda_{c,j} \mathcal{N}_c(\tilde{\mathbf{x}}; \mu_j, \mathbf{\Sigma}_j), \tag{1}$$

where $\tilde{\mathbf{x}}$ is the columns of the test sample $\mathbf{x}$ with the indices in $B$, and $\lambda_{c,j}$, $\mu_j$ and $\mathbf{\Sigma}_j$ are the weight, the mean and the covariance of the $j$-th Gaussian in accent $c$. The vector $\tilde{\mathbf{x}}$ is classified to the accent $c^*$ by maximizing the likelihood criterion:

$$c^* = \arg\max_c \ log(f_c(\tilde{\mathbf{x}})) \tag{2}$$

The covariance matrices of the Gaussian are restricted to be diagonal to ensure that all parameters can be estimated reliably from the training data by means of an iterative expectation maximization (EM) training algorithm.

*3.3.2. Linear Discriminant Analysis*

With LDA a speaker model is transformed to be optimal for classification, under the assumptions of normal distributions and homoscedascity in all classes. Formally, the aim of LDA is to maximize the ratio of the between-class scatter, $\mathbf{S}_b$, to the within-class scatter, $\mathbf{S}_w$. This approach uses the optimizing criterion, $\mathbf{S}_w^{-1}\mathbf{S}_b$ to transform the data sets irrespective of their class identity. The two scatters are computed as:

$$\mathbf{S}_b = \sum_{c=1}^{N} (\mu_c - \overline{\mu})(\mu_{\mathbf{c}} - \overline{\mu})^T, \tag{3}$$

$$\mathbf{S}_w = \sum_{c=1}^{N} p_c \Sigma_c, \tag{4}$$

with $\mu_c$, the mean of the models of the training speakers $\tilde{\mathbf{X}}$ belonging to class $c$ and $\overline{\mu}$ the mean of the entire training set. The term $p_c$ refers to the prior probability of class $c$ ($p_c = 1/N$ in our case), and $\Sigma_c$ to the covariance matrix of class $c$. In ideal circumstances, the transformation matrix $\mathbf{T}$ is found as the eigenvector matrix of the optimizing criterion.

The transformed data is now assumed to be distributed sufficiently normal with equal variance per class such that a simple Euclidean distance can be used to classify data points. Hence a speaker is classified by searching the accent, $c^*$, which minimizes:

$$c^* = \arg\min_c \|\mathbf{T}^T(\tilde{\mathbf{x}} - \mu_c)\| \tag{5}$$

*3.3.3. Support Vector Machine*

SVM (Burges, 1998) is one of the most popular supervised classifiers on a wide range of data sets due to its superior performance. It looks for a *maximum-margin hyperplane* to separate data from two categories. For a two-class problem, suppose input speaker models $\mathbf{x}$ (or $\tilde{\mathbf{x}}$) are mapped into a higher dimensional space by a mapping function $\varphi(\centerdot)$: $\mathbf{z} = \varphi(\mathbf{x})$, where $\mathbf{z}$ denotes vectors in the high dimensional space. The optimal hyperplane $f(\centerdot)$ is of the form:

$$f(\mathbf{z}) = \mathbf{w}^T\mathbf{z} + b, \tag{6}$$

where the weight vector $\mathbf{w}$ is a linear combination of training patterns, and $b$ is a bias value. SVM can be formulated in a primal and equivalent dual form. An interesting property of the SVM solution is that a sparsity pattern in the dual variables is observed. The samples which correspond to non-zero dual variables are called support vectors. To a non-linear classification problem, the data can be projected onto a high dimensional space where the maximum-margin hyperplane is found. The mapping function is implicitly defined by a kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j)$. In this paper we focus on the linear and the radial basis function (RBF) kernels as they are the most commonly used kernel functions.

An important issue of the SVM approach is that the complexity of the classifier is described by the number of support vectors and not by the dimensionality of these vectors. Thus, it is insensitive to the *curse of dimensionality*. This property is suitable for our high dimensional native AID task.

The two-class SVM classifier can be extended to the multi-class case by combining several binary SVMs with the one-versus-one (OVO) strategy or the one-versus-all (OVA) strategy. Both strategies decompose the multiple classification problem into a set of binary classification problems. The OVA strategy is to build one SVM for each class, using as negative examples all the patterns of the other classes ; while the OVO strategy trains one SVM for each pair of classes and a voting decision is made to assign an unknown pattern to the class with the maximum number of wins. In our experiments we found no significant difference between both approaches; so only the results of the OVA strategy are presented in this paper.

*3.4. Feature selection and dimensionality reduction*

As we stated, the features $x_j$ $(1 \leq j \leq D)$ may contain different amounts of accent relevant information. A feature selection is necessary to obtain robust models.

Three feature selection and/or dimensionality reduction techniques are studied. In the first method, we use an analysis of variance (ANOVA), to rank features by conducting a series of statistical hypothesis tests, then we select a desired number of features whose p-values are smallest in the feature set (see section 3.4.1). LDA is implemented as the second technique. By multiplying the vector $\mathbf{x}$ with the transformation matrix $\mathbf{T}$, the dimensionality is reduced, as explained in section 3.4.2. As a last method, we use SVM-RFE to evaluate subsets of remaining features and we eliminate the least influential feature recursively with respect to an SVM classifier. This procedure will be explained in more detail in section 3.4.3.

*3.4.1. ANOVA*

One-way ANOVA is used to determine the significance of each feature in the speaker models. It assumes that the observations are normally distributed and that the variances are equal in all $N$ groups.

The null hypothesis, i.e. that the means of a single feature are equal, is rejected if the *p−value*:

$$p = P(F_{N-1,n-N} > F{-}ratio) \tag{7}$$

is smaller than a chosen level of significance, $\alpha$. The *F−ratio*

$$F{-}ratio = \frac{\sum_{c=1}^{N} n_c(\overline{x}_c - \overline{x})^2}{\sum_{c=1}^{N}(n_c - 1)\sigma_c^2} \frac{n-N}{N-1} \tag{8}$$

can be thought of as a measure of how different the means are relative to the variability within each class, where $\overline{x}_c$ is the average of a single feature $x$ within class $c$, $\sigma_c^2$ the variance and $\overline{x}$ the global mean. Fisher showed that, under the given assumptions, this ratio follows an $F$-distribution with $N-1$ and $n-N$ degrees of freedom. The larger this value, the greater the likelihood that the differences between the accent specific means are due to something else than chance alone. So, the accent specific means of the features with a small p-value are statistically unequal due to accent variation instead of chance. The features with small p-values can be used as a feature set on which a classifier is built to classify the accents.

ANOVA is a fast and efficient way to evaluate the importance of each feature. The significances of all features can be ranked by sorting their corresponding p-values in an ascending order. However, there are two concerns about this technique. One is that it requires the assumptions of normality

and equal variance for each feature. Fortunately, 88.4% of the features in our speaker models satisfy this assumption as verified by a normal quantile test. The other issue is that ANOVA evaluates the features independently. Consequently some highly correlated features may obtain rather small p-values and thus are selected into the feature subset simultaneously. Nevertheless, if one of the correlated features exists in the selected feature set, the presence of other correlated features would not help to considerably improve the discriminability.

ANOVA can also be used as a pre-selection procedure for other high computational feature selection or dimensionality reduction techniques, such as LDA and SVM-RFE, as long as the empirically pre-defined size of the feature subset selected by ANOVA is large enough to avoid losing important features.

*3.4.2. LDA as dimensionality reduction*

The between-class scatter matrix $\mathbf{S}_b$ is the sum of $N$ matrices of rank one or less, and because only $N-1$ of these are independent, $\mathbf{S}_b$ is of rank $N-1$ or less. So for our $N$-class problem, multiple discriminant analysis primarily provides a way of reducing the dimensionality from a $D$-dimensional space to an $(N-1)$-dimensional space.

The dimensionality of a speaker model $D$ can be much larger than the number of speaker models $n$. This results in difficulties in estimating the covariance matrices, which are required in the computation of the LDA transformation matrix. The solution can be either to reduce the dimensionality $D$ or to increase the number of speaker models $n$, or both. We first use ANOVA to make a rough selection of the most promising features to decrease $D$. Then instead of building one speaker model per speaker, we split the data of all training speakers in shorter segments, yielding multiple models per speaker. While these multiple models are highly correlated, they will differ due to differences in infrequent events and they will differ more for shorter segment lengths. Then multiple speaker models obtained from short segments are used to estimate a reliable LDA transformation matrix.

*3.4.3. SVM-RFE*

Recursive feature elimination based on SVM (SVM-RFE) (Guyon, 2002; Rakotomamonjy, 2003), derived from the classical SVM, is a feature ranking algorithm to evaluate the contribution of each feature to the classification error in the sense of a maximum margin criterion. In Equation 6, the squared

weight $w_i^2$ of the $i$-th feature in the weight vector $\mathbf{w}$ is a measurement calibrating the contribution of this feature to the margin of the hyperplane. The larger the margin is, the more discriminative this feature will be. The feature with the smallest contribution to $||\mathbf{w}||^2$ is removed. The feature elimination thus follows an iterative procedure that can be described as follows:

(1) Train the SVM classifier with the current feature set

(2) Compute the contribution of each feature

(3) Eliminate the feature with smallest contribution to the norm of $\mathbf{w}$ from the current feature set

(4) Start over again from the step (1) unless a desired number of features is reached

We followed the method by Duan (Duan, 2005) proposed to extend the SVM-RFE from a binary classification to a multi-class case by re-organizing the contribution criterion of features. Suppose some linear binary SVMs are obtained by the OVO or the OVA strategy, the ranking score of each feature is computed as:

$$r_i = \frac{\underset{j}{mean}(w_{ji}^2)}{\underset{j}{variance}(w_{ji}^2)}, \tag{9}$$

where $w_{ji}$ is the weight value associated with the $i$-th feature from the $j$-th binary SVM.

For computational reasons, we also use ANOVA to coarsely pre-select important features, as we did for LDA. We may also simultaneously remove several features at each iteration to speed up the selection, without significant loss of accuracy.

## 4. Experiments and Discussion

### *4.1. Experimental Setup*

Experiments are done on the read speech part of the CoGeN corpus[1]. The CoGeN corpus contains 174 Flemish speakers of which 101 are male speakers.

[1]The read speech part of CoGeN is a part of component 'o' of the CGN corpus (Corpus Gesproken Nederlands): *http://tst.inl.nl/cgndocs/doc_English/start.htm* .

In the read speech part speakers were asked to read five paragraphs of standard Dutch, yielding about 200 seconds of speech per speaker. Paragraphs are different from speaker to speaker.

The experiments are divided into a text dependent mode and a text independent mode. In the text dependent mode, phoneme or state segmentations are obtained from a Viterbi alignment using an in house large vocabulary HMM system (Demuynck et al., 2008) with phonetic dictionary and orthographic sentence transcriptions as inputs. Although the text dependent mode is not realistic in many circumstances, it helps us better understand the role of features and phonemes in native AID. In text independent mode no knowledge of what was said is assumed *a priori*, but it is inferred from a phone recognizer.

The phonetic dictionary is adopted from Fonilex (Mertens and Vercammen, 1998), a list of more than 200,000 Dutch word forms with their Flemish pronunciation. Fonilex uses the YAPA (Yet Another Phonetic Alphabet) phonemic encoding scheme. A description of the 38 symbols and their sonorant/ nonsonorant nature can be found in Appendix A. After having discarded four phonemes (ɡ, ʃ, ʒ and ø) with a low probability of occurrence, the dimensionality of the speaker vector is $D = 1326 = 34 \times 3 \times 13$.

We use the five Flemish provinces as accent clusters as they correspond fairly well to the dialect regions mentioned in (Van Hout et al., 1999) and we use the place of birth to assign a speaker to a province. Using this classification, we can not expect a 100% identification accuracy as the class membership can not be justified for all speakers on the basis of acoustic-phonetic properties. The following abbreviations are used in figures and tables to refer to the provinces: Ant for Antwerp (42 speakers), Bra for Brabant (26 speakers), Lim for Limburg (34 speakers), E-Fl for East-Flanders (36 speakers) and W-Fl for West-Flanders (36 speakers). In order to increase the statistical significance of the obtained results on the relatively small CoGeN corpus a leave-one-out evaluation scheme is adopted. In this approach 174 separate experiments are performed in which each time 173 speakers are used for training and the excluded one for testing. The parameters used in the SVM classifier are obtained by a 10-fold cross validation over the training set. The 95% confidence interval for the leave-one-out setup can be computed to be $\pm 3.2\%$ with a very slight dependency on the expected outcome.

In all experiments, unless claimed otherwise, the default segment length for estimating the speaker models, both in training and test, is 200 seconds. This is well within the region of convergence of the speaker models.

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

Table 1: Accent identification rates for brute force GMMs

*4.2. A difficult database*

As Flanders is a small region, one might expect only small accent differences. Traditionally, however, Flemish has been a very dialectal language with extreme differences over small distances. Due to increased communication, dialects are disappearing quickly and most younger people speak (or can speak) the standard language with only a minor accent. For the CoGeN database, speakers were asked to speak the standard language in a read style, hence accent differences are expected to be relatively small and challenging to detect.

Indeed in informal listening tests with 8 untrained native listeners and speaker segments of 30 seconds, an average accent identification accuracy of 45% was obtained with a maximum of 62% for the best listener and a minimum of 36% for the worse one. In an automatic classification, the brute force GMM without using any phonetic segmentation, shown to be effective in identifying four Chinese accents (Chen et al., 2001) and distinguishing Flemish and Dutch (with an accuracy of 83.8%), gives very poor performance for discriminating the five Flemish accents, as shown in Table 1. The best accuracy, 27.0% is only slightly higher than the 20% chance level. The failure of the traditional GMM in acoustic modeling also indicates the difficulty of the task on this database.

*4.3. Speaker based accent identification*

From this section on until section 4.5, we assume that the read text is known and that the phoneme segmentation is obtained by forced alignment. We first show the experimental results of accent identification without feature selection. Then we demonstrate the benefits brought by the feature ranking of ANOVA. The feature ranks given by ANOVA are analyzed afterwards, and compared to those of the other two feature selection/dimensionality reduction techniques we adopt. Finally we present a frame based GMM framework for the AID incorporated with the feature selection.

*4.3.1. Accent identification without feature selection*

A speaker based supervector is constructed by averaging the phoneme-specific speech features for each speaker, as explained in section 3.2.1. GMM and SVM with two different kernels, a linear kernel and a RBF kernel are taken as the classifiers to discriminate the accents. We tested the performance of GMM using different number of Gaussian mixtures and the best performance was achieved when only a single Gaussian density (SGD) distribution is used to model an accent class. The number of speaker vectors per class is probably not large enough in our database to accommodate more complex models. The accuracies of SGD, linear SVM and RBF-SVM are 50.6%, 63.2% and 62.6% respectively, as shown in Table 2. The superiority of the SVM classifiers is due to their insensitivity to the *curse of dimensionality*.

| Classifiers | SGD | Linear SVM | RBF-SVM |
|---|---|---|---|
| AID accuracy (%) | 50.6 | 63.2 | 62.6 |

Table 2: Accent identification rates for speaker-based models without feature selection

*4.3.2. Accent identification with ANOVA ranking*

Given the speaker models, features $x_i$ are evaluated by ANOVA and ranked according to their p-values. The features with smallest p-values are selected to be part of the feature set. The best accuracy is 71.8% achieved by the RBF-SVM when 100 feautures are selected, compared to 62.6% with the full feature set. For the GMM classifier, a single Gaussian density distribution still performs better than GMM systems with multiple Gaussians. With the ANOVA feature selection, the performance of SGD is boosted from 50.6% to 70.1% when the size of selected feature subset is 80. Figure 2 demonstrates the performance obtained by the RBF-SVM, the linear SVM and the SGD respectively for different numbers of selected features. Note that the end points of the curves represent the accent identification without feature selection. We can also see that the best accuracies for all classifiers are in the range of 40-120 features, indicating that 90% of the features in the original speaker vectors do not help in discriminating the accents. Overall, the RBF-SVM classifiers outperform the SGD classifiers with the same number of features. Moreover, the performance of SVMs degrades much slower

with increasing number of features than that of the SGD classifier. Nevertheless, feature selection remains an essential ingredient for success for all classifiers. Finally, there is only a marginal difference between the linear and RBF kernel, with a small preference for the latter.

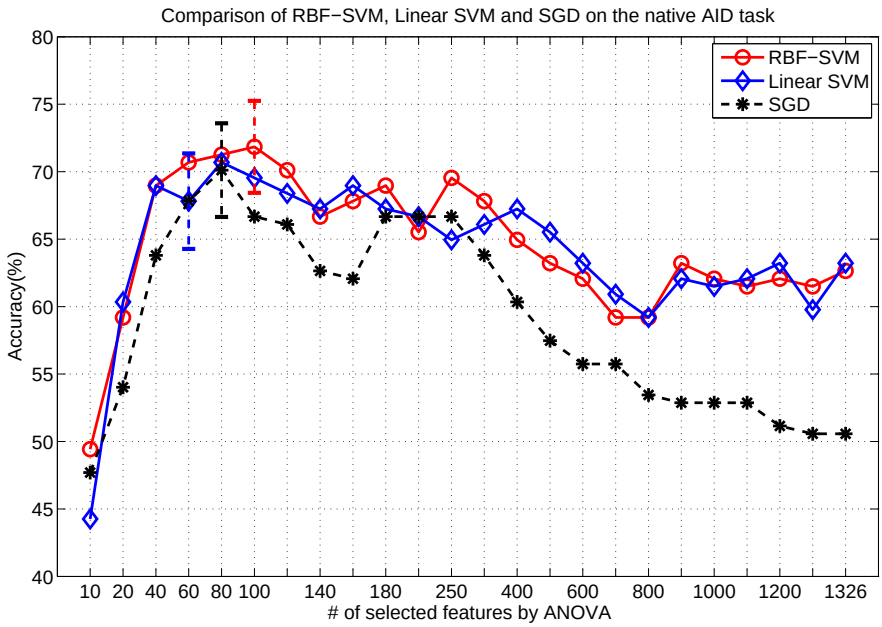



Figure 2: Comparison of RBF-SVM, linear SVM and SGD classifiers with the ANOVA feature selection in the text dependent mode

Table 3 gives the confusion matrix when the number of selected features is 100 with the RBF-SVM classifier and ANOVA. We observe that the selected features are able to discriminate relatively well between speakers from on the one hand Ant and Bra, and on the other hand E-Fl and W-Fl. Greater confusion within these regions is expected because these are geographically close and were politically tied in the middle ages, even when Flanders as a whole was divided between France and Germany. The rather big confusion between Lim and the cluster of E-FL and W-Fl is still unclear, given that Flemish listeners are quite good at discriminating Lim speakers from E-Fl or W-Fl. One possible explanation may be that human listeners make use of

| % | true class Ant | Bra | Lim | E-Fl | W-Fl |
|---|---|---|---|---|---|
| Ant | **92.8** | 26.9 | 2.9 | 8.3 | 5.6 |
| Bra | 7.2 | **61.5** | 2.9 | 8.3 | 11.1 |
| Lim | 0 | 0 | **73.6** | 2.8 | 2.8 |
| E-Fl | 0 | 11.6 | 5.9 | **63.9** | 19.4 |
| W-Fl | 0 | 0 | 14.7 | 16.7 | **61.1** |

Table 3: Normalized confusion matrix on a five forced-choice classification of Flemish accents (global recognition rate is 71.8%) when the number of features is optimal. The RBF-SVM is used as the classifier and ANOVA is used to select features. The columns represent the true class, the rows the recognized class.

intonation, a feature which is not incorporated in our raw feature extraction.

We also investigate the performance of AID under different test lengths. A fixed length speech segment is randomly selected from a test speaker and a test speaker model is generated on the basis of this segment (missing phonemes are represented by the average over all classes). The data segmentation is repeated 10 times (except for the 200-second test length case) and the average accuracies are plotted in Figure 3 with 5, 10, 30, 60 and 200 seconds test lengths. As can be seen, the performance of the AID improves as the test length grows. Moreover, the ranges of the best accuracies in different test times are quite different; short test lengths seem to require more features, or are less sensitive to feature selection. This is probably because in short lengths some very informative features are not available, or may not occur often enough, thus other less reliable but maybe correlated features are included and helpful.

*4.3.3. Discriminative phonemes, states and features*

Figure 4 shows which features contribute most to the identification task. To avoid overloading, the contributions of the three states of each phoneme are taken together because if one state is selected, the other two are mostly selected too. Next to the cepstra and energy, we also show the most useful formants by the ANOVA selection, although the AID result of the formants is not presented since it is absolute 15% worse than that of the cepstra (Ghesquiére and Van Compernolle, 2002). The darkness corresponds to the number of times a feature is selected. As can be seen, most of accent differences

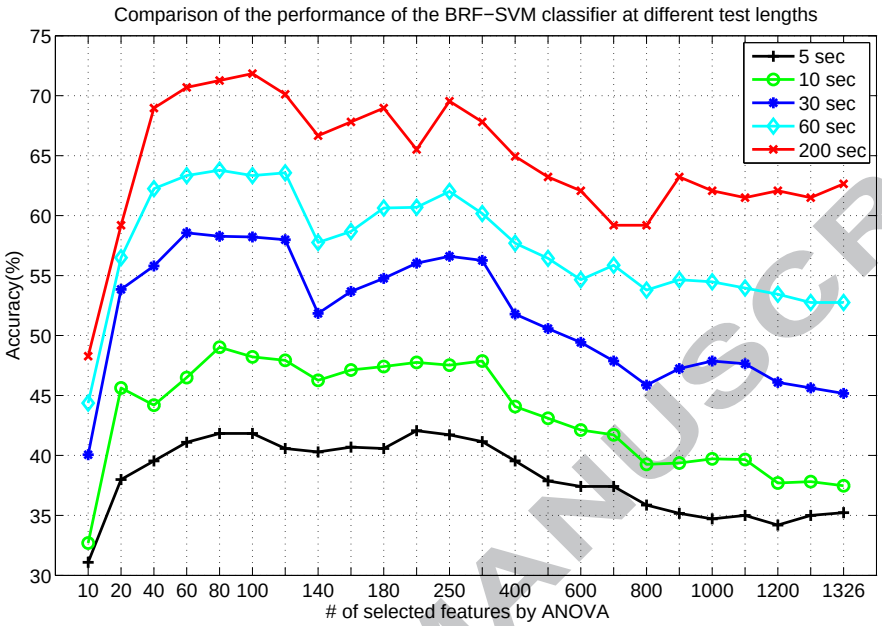



Figure 3: Comparison of the performance of the RBF-SVM at different test lengths.

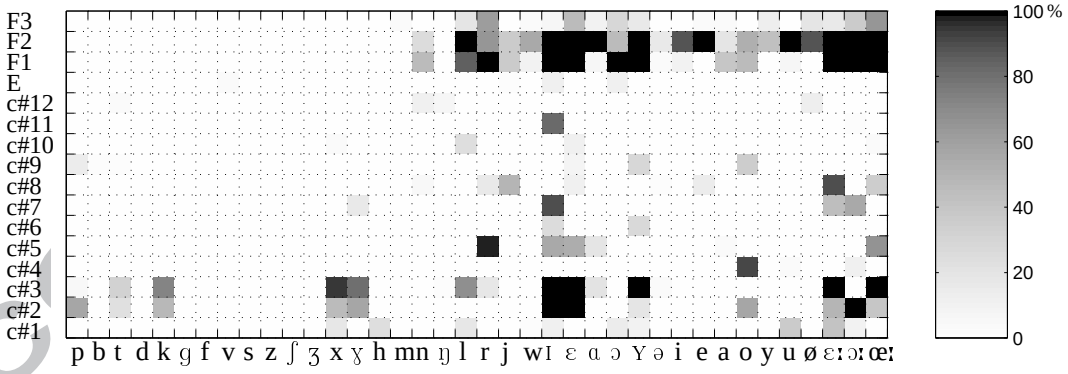



Figure 4: Number of times a feature is used at the optimal number of features. Speech features shown here include 12-order Mel cepstra, log energy and the first three formants (only applicable for sonorants)

come from vowels, although some consonants, like /p/, /t/ and /x/, etc, also contribute a little. A boxplot analysis[2] of each distinctive feature shows that most vowels contribute to the distinction between the cluster Ant - Bra, and the cluster E-Fl, W-Fl and Lim. The vowels /ɛ/ and /ɑ/, are responsible for a correct identification for speakers of Ant. The voiceless plosives, /p/, /t/ and /k/, and the fricative /x/, mainly contribute to differentiate between the cluster E-Fl and W-Fl and the cluster consisting of Ant, Bra and Lim. The correct identification of the speakers of Lim is mainly due to the phonemes /r/ and /ɣ/. The latter phoneme also contributes to the identification of speakers of W-Fl.

Besides of the different contribution of phonemes and features to accent classification, the states of a discriminative phoneme also behave differently. Figure 5 shows the phoneme states that are used at least once when the optimal number of features is selected by ANOVA. The darkness indicates the number of features selected from a state. Having 3 states per phoneme allows for more detailed discriminative modeling as certain differences may only pertain to beginning or end of a phoneme. For example, speakers from Ant usually tend to pronounce diphthongs as monophthongs by neglecting the second vowel in the diphthongs, resulting in the importance of the last two states in the diphthongs. Modeling such detailed differences was only possible by using three states per phoneme instead of using one state.

In Figure 6, we show how many phonemes and cepstral coefficients appear as a function of the size of the selected feature subset. As can be seen, at the optimal number of selected features (100 for RBF-SVM and 80 for SGD), about 22 phonemes are involved, indicating the remaining 12 phonemes do not contribute to the accent discrimination at all. Although all 13 cepstral coefficients (including 12 cepstra and 1 log-energy) show up in the first 100 features, their importance is quite unbalanced: the second and the third cepstral coefficients represent 53% of selected features. The importance of $c_2$ and $c_3$ is also revealed by the dark cells in Figure 4.

[2]A boxplot (Tukey, 1977) is a convenient way to graphically depict groups of numerical data. It helps to find out the degree of dispersion for a discriminative feature in different accents

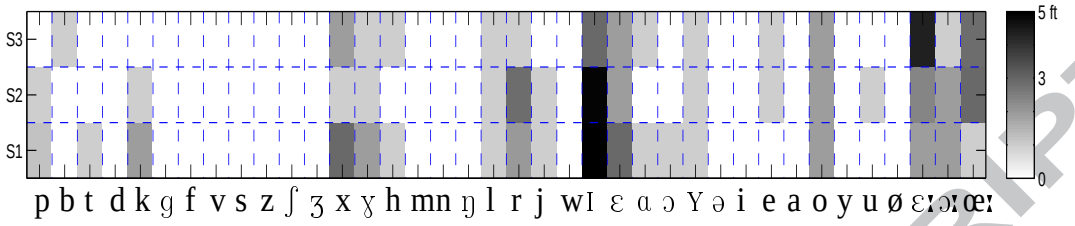



Figure 5: Contribution of phoneme states at the optimal number of features. $S_i$ is the $i$-th state of a phoneme.

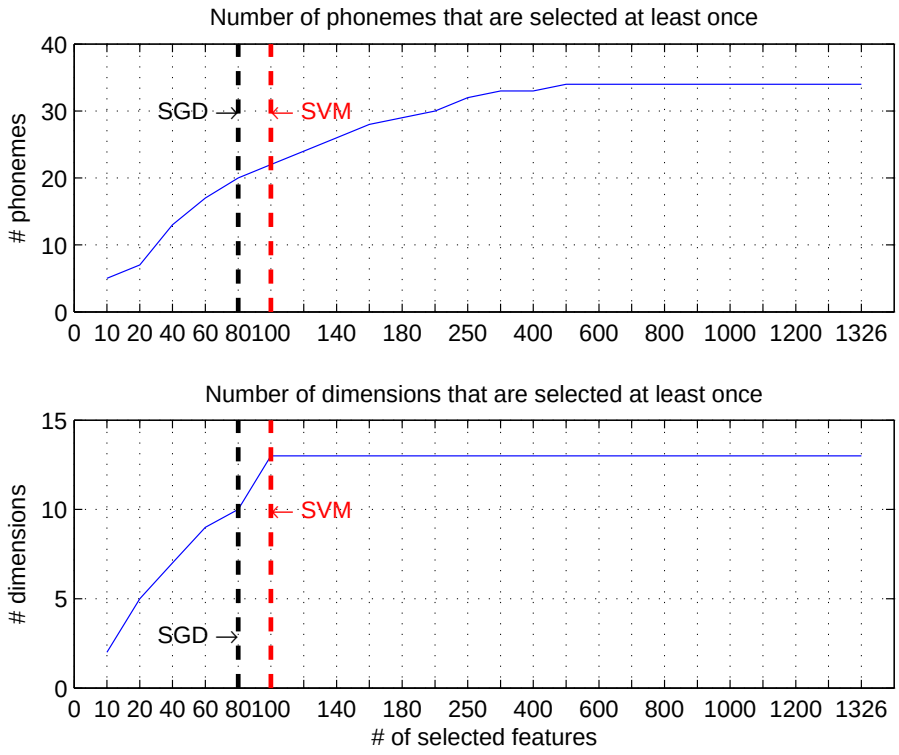



Figure 6: The number of phonemes (top) and dimensions (bottom) that are selected at least once as a function of the size of the reduced feature set

### *4.4. Feature selection: ANOVA, LDA and SVM-RFE*

In this experiment we will compare three feature selection/dimensionality reduction techniques, ANOVA, LDA and SVM-RFE. The main reason to look

for alternative techniques is that ANOVA handles the features independently, without taking the correlation between the features into account. On the contrary, the other two methods start with a large candidate feature set. LDA transforms this high-dimensional feature space to a 4-dimensional space in order to achieve the maximum inter-class separation; SVM-RFE eliminates the feature whose absence influences the decision hyper-plane the least from a present feature set.

However, because of the expensive calculation burden, it is impractical for LDA and SVM-RFE to start with the whole feature set. Thus, ANOVA is used first as a pre-selection to reduce the full set of features to a reasonable size of 500. This reduced feature set is presumed to cause no loss of important information and an acceptable computational cost.

#### *4.4.1. LDA as dimensionality reduction*

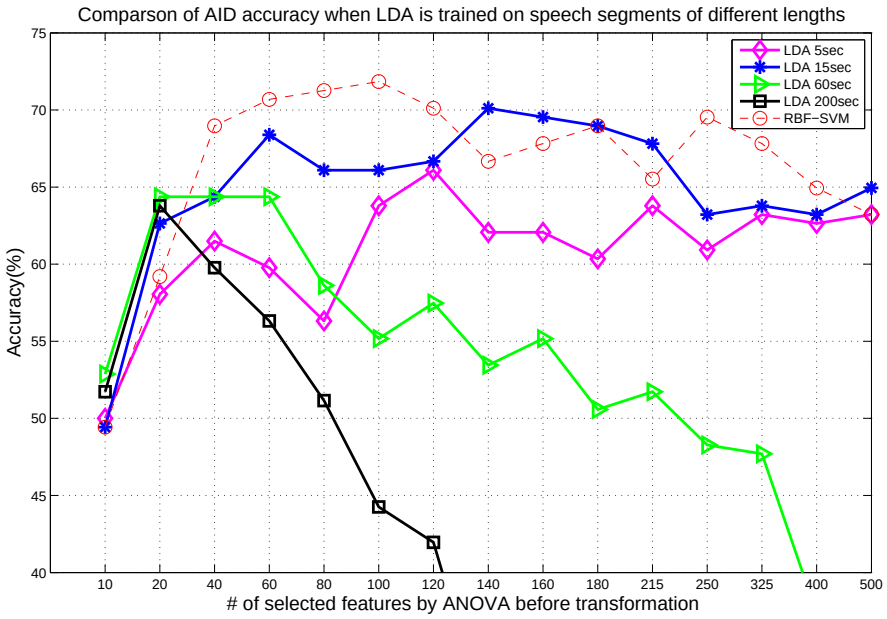



Figure 7: Comparison of identification accuracy when LDA is trained on speech segments of different lengths. ANOVA is adopted to select the most important features (500 or less) as the initial set. The length of the test segment is always 200 seconds.

LDA requires many training speaker models, preferably more than the size of the dimensions; therefore we split the data of a training speaker in shorter segments to construct multiple speaker models using the method described in section 3.4.2. A problem is that there is a trade-off between the number of training speaker models and the accuracy of a speaker model. Thus the segment lengths are arbitrarily set to be 5, 15, 60 and 200 seconds. We still keep the test length, 200 seconds unchanged to generate the test speaker model. The identification rates are shown in Figure 7. The horizontal axis indicates the number of features ranked by ANOVA. As a reference, we also give the accuracy of RBF-SVM identically in Figure 2. Typically the accuracy for the 200-second model is 63.8% optimal at 20 features while for the 15-second model 70.1% at 140 features.

The optimal segment length in our experiments is 15 seconds. For shorter segment lengths the performance is uniformly worse, which is caused by poor estimates of the speaker models. For longer segment lengths the results are slightly better when working with a small number of selected features. This improvement may be attributed to a better estimation of the speaker models. As the number of retained features increases, the size of the transformation matrix (thus the number of parameters that needs to be estimated) increases as well. This transformation matrix will be better estimated as the number of (pseudo) speakers increases, i.e. as the segment length gets shorter. This optimum of 15 seconds is hence very much a database dependent compromise between optimizing speaker model estimates (better with longer segments) and optimizing the estimate of the transformation matrix (better with more speakers).

*4.4.2. SVM-RFE*

SVM-RFE is performed to evaluate the contribution of each feature to the margin of the classifier. As long as the remaining number of features is larger than 250, the five features that contribute the least are eliminated simultaneously. From then, only one feature is eliminated at each step. The accuracies along with the remaining number of features are plotted in Figure 8, together with the RBF-SVM classifier using the ANOVA selection and the LDA based dimensionality reduction using 15-second segment length. Note that except for the initial 500 features, the feature set for SVM-RFE could be different from ANOVA and LDA even for the same test speaker.

At a moderate number of features, there appears to be no significant difference between ANOVA, SVM-RFE and LDA; however, when only a small

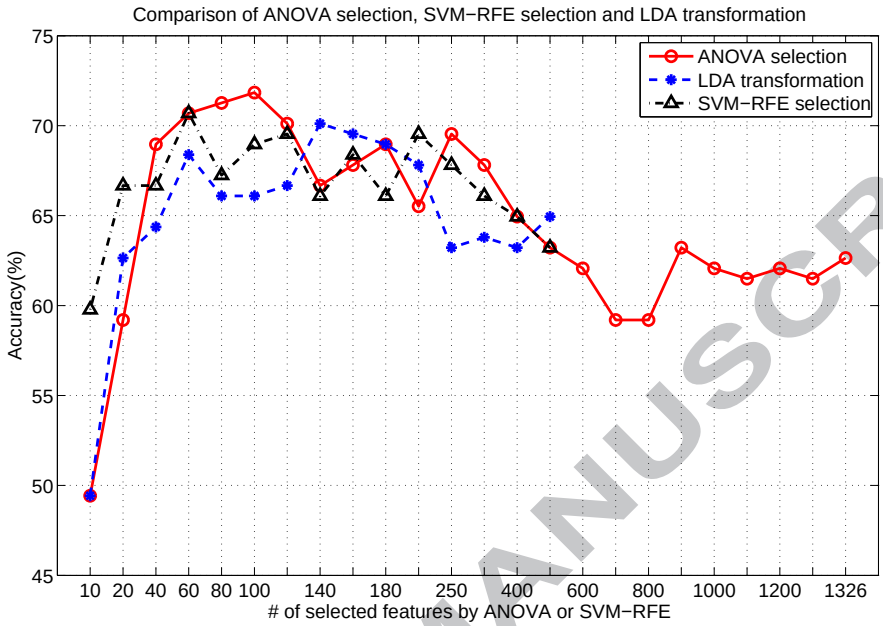



Figure 8: Comparison of the ANOVA selection, SVM-RFE selection and the LDA dimensionality reduction. ANOVA is adopted to select the most important features (500 or less) as the initial set for the SVM-RFE selection and the LDA. The length of the test segment is always 200 seconds.

number of features are being selected, SVM-RFE is superior to the other two. This phenomenon can be explained by the fact that the p-values emerging from ANOVA do not take the correlation between potential feature candidates into account, while the SVM-RFE does. For example, the second coefficient of Mel cepstrum of phoneme $I$, denoted as $/I/_2$, whose p-value is extremely small, seems to be highly discriminative for the accents Ant, Bra against the accents Lim, E-Fl and W-Fl. In our speaker modeling, the three correlated $/I/_2$'s from three HMM states are all highly ranked by ANOVA, implying they have a high chance to be selected even when the desired number of features is limited. But selecting more than one of $/I/_2$'s features does not largely increase the discrimination, as long as the set already holds one $/I/_2$. As a backward elimination method, SVM-RFE would throw away two of these correlated features in an earlier elimination and just keep the most

important one.

*4.5. Frame based accent identification*

In order to show the general applicability of the feature subset selection concept for difficult classification tasks such as accent identification, the method is incorporated in a GMM accent classifier in which the cepstral feature stream is used directly as input, i.e. without the preliminary step of speaker modeling. The diagram of the system is illustrated in Figure 9. We refer to this approach as the *frame based* system and to our speaker modeling approach as the *speaker based* system.

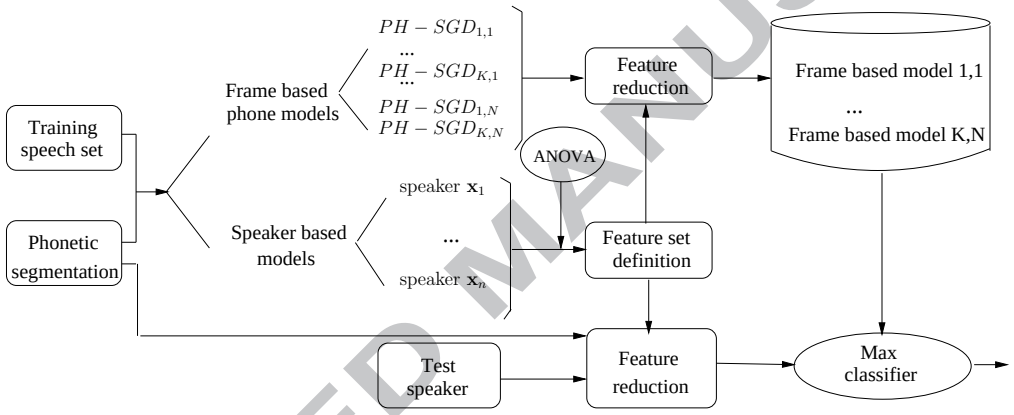



Figure 9: The diagram of frame based accent recognition

We progressively eliminate the less relevant cepstral coefficients for some or all phonemes and eventually the full contribution of certain phonemes in the likelihood computation. Which features and phoneme states to eliminate is derived from the speaker based system. The training procedure is text dependent in the same way for both systems. For each phoneme in each accent a single Gaussian model is built. Frame based log-likelihoods are added and a simple max-classifier is used. Eliminating dimensions in this frame based system may be interpreted as follows: eliminated features in specified phonemes are modeled by a global mean and variance instead of a class (accent) based one. As the contribution to the likelihood is the same in all accent models, they do not contribute to the discrimination and do not need to be computed.

In Figure 10 we compare the results for 5-class forced accent identification using a frame based classifier and a speaker based system with SGD as the

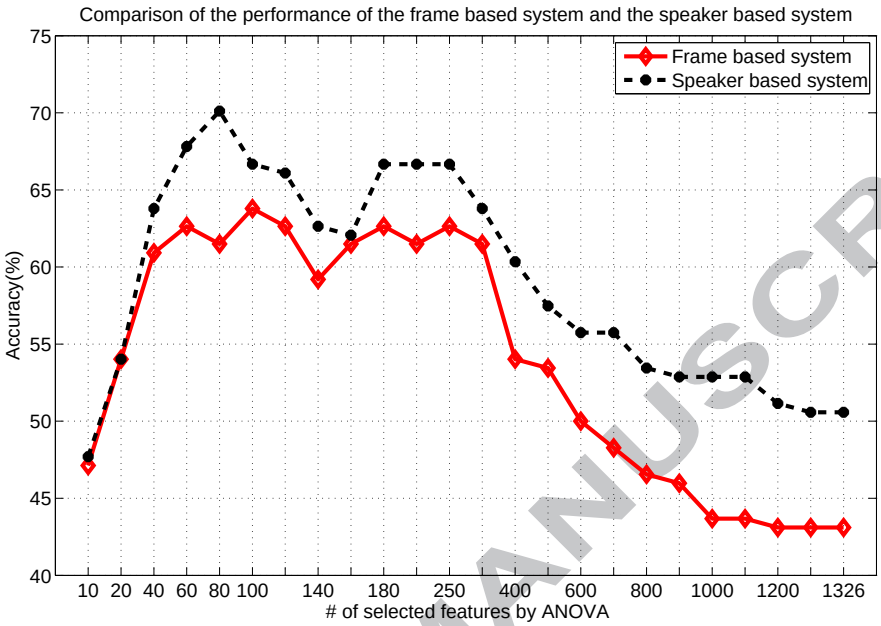



Figure 10: Identification accuracy of the speaker based versus the frame based SGD system as a function of the size of reduced feature set dimension.

classifier for different numbers of selected features. The number of model parameters in the two systems is the same. The effect of feature subset selection is highly similar for the frame based and speaker based systems, i.e. the optimal performance is achieved when 60-200 features are retained from the original 1326. On average the speaker based system is about 10% relative better than the frame based system.

Incorporating phonetic knowledge into the system significantly enhanced results over using one large brute force GMM per accent with the same number of parameters. Without phonetic information and without any feature selection a brute force GMM based classifier only obtains a performance of less than 30% (see Table 1). The best frame based system with feature selection and phonetic supervision reaches an accuracy of less than 65% while the best speaker based system achieves an accuracy above 70%. As in both the frame based system and the speaker based system, only a single Gaussian distribution is used to model accent dependent phonemes, or speaker vectors, the models are very similar. The means are almost identical in both

systems, while the variances are quite different. However this is not the crucial difference between the frame and speaker based systems. The better performance of the speaker based system comes from the embedded temporal normalization. In the speaker based systems phones are weighted equally (assuming there were enough observations), while in the frame based systems phones will contribute proportional to the number of observed frames for each phoneme. We verified this by following simple experiment. When weighting the score of each phoneme in the speaker based system according to their observed number of frames we obtained results that were fully in line with the frame based system.

### *4.6. Text independent experiments*

In order to go from a single brute force GMM per accent without phonetic knowledge to a frame based system or to our classifier using a speaker model, we rely on a phonetic annotation of the incoming speech obtained in a text dependent manner. This may only be possible under very specific circumstances. Therefore we should investigate how good the system performs in a text independent way for a fairer comparison with the brute force GMM classification system. The best phonetic segmentation will then be obtained from a large vocabulary speech recognition system. If speech recognition is the goal, such a system is obviously available. In other situations the overhead involved in running such a system may not be acceptable. Therefore as an alternative to supervision we obtain the phonetic segmentation from a phoneme recognizer for both the training set and the test set, including apart from the acoustic model only a phoneme trigram. The phoneme recognition rate on our database was 69.9%. As illustrated in Figure 11, the degradation from a reference text dependent SGD system to a speaker based text independent SGD system is about 10%, but the speaker based system still substantially outperforms the frame based system. The feature selection done by ANOVA consistently boosts the accuracies of accent identification with about a 50% relative improvement.

## 5. Final Discussion and Conclusions

In this paper, we have shown that feature subset selection significantly improves accent identification. The best performance is obtained when only 10% of the parameters in the raw models are retained for classification. Under these conditions of optimality, about one third of the phonemes are not taken

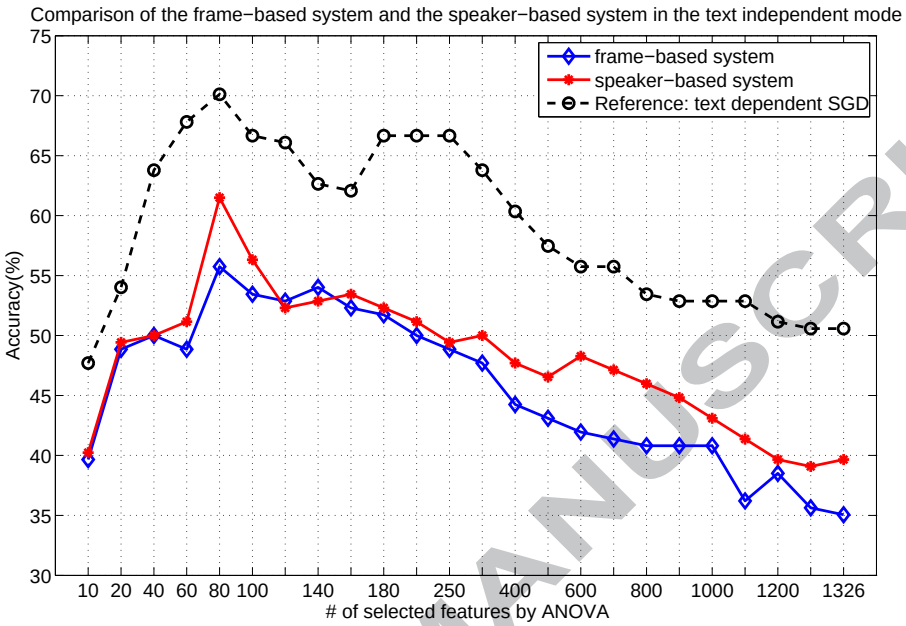



Figure 11: Identification accuracy of the speaker based versus a frame based SGD system as a function of the reduced feature set dimension in the text independent case.

into account in the classification process. Low order cepstral coefficients also seem to be much more discriminative than high order ones.

We found that ANOVA is a quick and efficient method to evaluate the significance of each feature individually. Therefore it is ideally suited as an initial feature filter independent of the back-end classifier. However its inability to take dependencies between features into account makes it suboptimal for feature selection when a very small number of features is desired. In this case embedded feature selection methods, such as SVM-RFE, lead to higher performance, be it at a high computational cost. When the number of selected features increases, there is no significant difference between the three feature selection techniques. Among the different classifiers tested (GMM, LDA and SVM) the latter is to be preferred due to its inherent lower sensitivity to dimensionality and implicitly making it less sensitive to feature selection.

While feature selection is a critical component of our AID classifier, nu-

merous other system design options play an important role as well. For example, in previous work (Ghesquiére and Van Compernolle, 2002), formants of sonorant phonemes were used as features. Although formants have the advantage of being robust against disturbances, they can not be reliably estimated and they are bad descriptors for non-sonorant phonemes. Their performance is therefore 15%-20% worse than that of cepstra. In our experimental designs, we also considered the first and second order derivatives of the cepstra. But as they only caused a marginal improvement for significantly increased computational complexity, we did not present their performance in our experiment section. In Table 4 we summarize the observed contribution of each of these choices for our problem. We believe that similar dependencies and trends will be observed on other tasks and databases in other languages. However the exact numbers may differ substantially, mainly depending on three aspects: the difficulty of the task, the amount of training data and the back-end classifier.

| | |
|---|---|
| usage of feature selection | 20-50% |
| cepstra on all phonemes vs. formants on sonorants | 15-20% |
| speaker vs. frame based | 10% |
| 3 states per phoneme vs. 1 state per phoneme | 5-10% |
| inclusion of delta-cepstra | $< 2\%$ |
| text dependent vs. text independent | 10-20% |

Table 4: AID performance degradation by different experiment's setups

Finally we want to address the question to what degree the feature selection is due to limited training data or to the presence of truly irrelevant features. From a Bayesian perspective there is no need to remove features under the assumption of infinite amounts of training data. In practical circumstances finite amounts of data lead to estimation errors and irrelevant features will have different values for different classes hence inducing noise into the classifier. In the case of accent identification we spread our speakers over the different accent classes. Moreover we have argued that only part of the available speech will contribute to accent recognition. Hence, if speech recognition in general is plagued by sparse data problems, accent recognition will fare significantly worse. Therefore our belief is that the effect of larger databases will be as follows: more features might be discovered as being rel-

evant and the net result of feature selection may be less dramatic than for the database that we used as the estimation variance of irrelevant features gets smaller with more training data. Nevertheless we believe that feature selection will be beneficial, especially for highly confusing accents as their recognition is more hampered by small amounts of classification noise than in cases when the intrinsic differences are larger.

## References

Adda-Decker, M., Lamel, L., 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2-4), 83-98.

Burget, L., Matejka, P., Cernocky, J., 2006. Discriminative training techniques for acoustic language identification. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Vol. I., pp. 209-212.

Berkling, K., Zissman, M., Vonwiller, J., Cheirigh, C., 1998. Improving accent identification through knowledge of English syllable structure. In: Proc. International Conference on Spoken Language Processing, Vol. II., pp. 89-92.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, Vol. II., 121-167.

Campbell, W.M., 2008. A covariance kernel for SVM language recognition, In: Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 4141-4144.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C., 2007. Acoustic language identification using fast discriminative training. In: Proc. Interspeech, pp. 346-349.

Chen, T., Huang C., Chang, E., Jiang, W., 2001. Automatic accent identification using Gaussian mixture models. In: Proc. IEEE workshop on Automatic Speech Recognition and Understanding, pp. 343-346.

Cremelie, N., Martens, J.P., 1999. In search of better pronunciation models for speech recognition. Speech Communication 29 (2-4), 115-136.

Demuynck, K., Roelens, J., Van Compernolle, D., Wambacq, P., 2008. SPRAAK: an open source "SPeech Recognition and Automatic Annotation Kit". In: Proc. Interspeech, pp. 495-495.

Duan, K-B., Rajapakse, J.C., Azuaje, F., 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on Nanobioscience. Vol. 4, no. 3., 228-234.

Ghesquiére, P., Van Compernolle, D., 2002. Flemish accent identification based on formant and duration features. In: Proc. International Conference on Acoustics, Speech and Signal Processing. Vol. I., pp. 749-752.

Guyon, I., Weston, J., Barnhill, S., Vapnik, N., 2002. Gene selection for cancer classification using support vector machines. Machine Learning. Vol. 46, no. 1-3., 389-422.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research, Vol. 3, 1157-1182

Hansen, John H.L., Yapanel, U., Huang R., Ikeno A., 2004. Dialect analysis and modeling for automatic classification. In: Proc. International Conference on Spoken Language Processing, pp. 1569-1572.

Huang, R. Q., Hansen, John H.L., 2006. Gaussian mixture selection and data selection for unsupervised Spanish dialect classification. In: Proc. Interspeech, pp. 445-448.

Knops, U., 1984. Cognitieve en evaluatieve reacties met betrekking tot regionale standaard-variëteiten. Een vergelijking tussen Vlamingen en Nederlanders. Taal en Tongval (36), 25-49 and 117-142.

Kohavi R., John G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence, Vol. 97, 273-324.

Labov, W., 1996. The organization of dialect diversity in North America, In: Proc. International Conference on Spoken Language Processing, http://www.ling.upenn.edu/phono_atlas/ICSLP4.html.

Lamel, L., Gauvain, J., 1995. A phone-based approach to non-linguistic speech feature identification. Computer Speech and Language Vol. 9 , no. 1, 87-103.

Lincoln, M., Cox, S., Ringland, S., 1998. A comparison of two unsupervised approaches to accent identification. In: Proc. International Conference on Spoken Language Processing. Vol. II., pp. 109-112.

Liu, M., Xu, B., Huang, T., Deng, Y., Li, C., 2000. Mandarin accent adaptation based context-independent/context-dependent pronunciation modeling. In: Proc. International Conference on Acoustics, Speech and Signal Processing. Vol. II., pp. 1025-1028.

Matejka, P., et al, 2006. BRNO University of technology system for NIST 2005 language recognition evaluation. In: Proc. IEEE Odyssey Speaker and Language Workshop, pp. 57-64.

Merriam-WebsterOnlineDictionary, 2008. Available from: http://www.merriam-webster.com/dictionary/accent.

Mertens, P., Vercammen, F., 1998. The Fonilex Manual.

Purnell, T., Idsardi, W., Baugh, J., 1999. Perceptual and phonetic experiments on American English dialect identification. Journal of Language and Social Psychology. Vol. 18, no. 1, 10-30.

Rakotomamonjy, A., 2003. Variable selection using SVM-based criteria. Journal of Machine Learning Research, Vol. III., 1357-1370.

Shen, W., Chen, N., and Reynolds, D., 2008. Dialect recognition using adapted phonetic models. In: Proc. Interspeech, pp. 763-766

Shriberg, E., Ferrer, L., Kajarekar, S., Scheffer, N., Stolcke, A., Akbacak, M., 2008. Detecting nonnative speech using speaker recognition approaches. In: Proc. Odyssey Speaker and Language Recognition Workshop.

Teixeira, C., Trancoso, I., Serralheiro, A., 1996. Accent identification. In: Proc. International Conference on Spoken Language Processing, pp. 1784-1787.

ten Bosch, L., 2000. ASR, dialects, and acoustic/phonological distances. In: Proc. International Conference on Spoken Language Processing. Vol. III., pp. 1009-1012.

Thomas, E., 2000. Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. Journal of Phonetics 28, Vol. I, 1-25.

Torres-Carrasquillo, P.A., Reynolds, D.A. and Deller, J.R. Jr., 2002. Language identification using Gaussian mixture model tokenization. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Vol. I., pp. 757-760.

Torres-Carrasquillo, P.A., Gleason, T.P., Reynolds D.A., 2004. Dialect identification using Gaussian mixture models. In: Proc. Odyssey, pp. 297-300.

Torres-Carrasquillo, et al, 2008. The MITLL NIST LRE 2007 language recognition system. In: Proc. Interspeech, pp. 719-723.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Van Hout, R., De Schutter, G., De Crom, E., Huinck, W., Kloots, H., Van de Velde, H., 1999. De uitspraak van het Standaard-Nederlands. Variatie en varianten in Vlaanderen en Nederland. In: E. Huls & B. Weltens (red.). Artikelen van de Derde Sociolinguïstische Conferentie. Delft: Uitgeverij Eburon, 183-196, http://www.cnts.ua.ac.be/Publications/1999/VDDHKV99/.

Wu, T., Van Compernolle, D., Duchateau, J., Yang, Q., Martens, J-P., 2005. Improving the discrimination between native accents when recorded over different channels. In: Proc. Interspeech, pp. 2821-2824.

Wu, T., 2009. Feature selection in speech and speaker recognition, Ph.D. dissertation, Katholieke Universiteit Leuven.

Zissman, M. A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing, Vol. 4, 31-44.

Zissman, M. A., Berkling, K. M., 2001. Automatic language identification. Speech Communication vol. 35 (1-2), 115-124.

APPENDIX A

| phoneme | Dutch example | Close corresponding English example | sonorant |
|---|---|---|---|
| p | /p/aars | /p/urple | |
| b | /b/al | /b/all | |
| t | /t/afel | /t/able | |
| d | /d/ansen | /d/ance | |
| k | /k/iel | /k/eel | |
| ɡ | za/k/doek | /g/un | |
| f | /f/ilm | /f/ilm | |
| v | /v/eel | /v/ery | |
| s | /s/inds | /s/ome | |
| z | /z/eel | /z/eal | |
| ʃ | /sj/aal | /sh/ort | |
| ʒ | gara/g/e | mea/s/ure | |
| x | a/ch/ter | lo/ch/ | |
| ɣ | /g/eval | | |
| h | /h/uren | /h/orse | |
| m | /m/es | /m/ess | X |
| n | /n/acht | /n/ight | X |
| ŋ | ri/ng/ | ri/ng/ | X |
| l | /l/eef | /l/ive | X |
| r | /r/ond | a/r/ound | X |
| j | /j/ij | /y/ou | X |
| w | /w/andel | /w/alk | X |
| ɪ | p/i/t | b/i/t | X |
| ɛ | p/e/n | p/e/n | X |
| ɑ | n/a/t | | X |
| ɔ | l/o/g | b/ou/ght | X |
| ʏ | n/u/t | | X |
| ə | d/e/ | /a/bout | X |
| i | d/ie/r | /ea/se | X |
| e | v/ee/l | | X |
| a | m/aa/n | | X |
| o | v/oo/r | | X |
| y | m/uu/r | | X |
| u | d/oe/n | p/u/t | X |
| ɸ | d/eu/r | | X |
| ɛː | /ij/s | h/a/te | X |
| ɔː | k/ou/s | b/oa/t | X |
| œː | h/ui/s | | X |

Table 5: The Dutch phonemes. A Dutch example and, if possible, a close corresponding English example is given.