Comparison of Variable Selection Methods and Classifiers for Native Accent Identification

Tingyao Wu, Peter Karsmakers, Hugo Van hamme, Dirk Van Compernolle

Dept. ESAT, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{Tingyao.Wu, Peter.Karsmakers, Hugo.Vanhamme, Dirk.VanCompernolle}@esat.kuleuven.be

10.21437/Interspeech.2008-140

Abstract

Acoustic differences are so subtle in a native accent identification (AID) task that a brute force frame-based Gaussian Mixture Model (GMM) fails to discover the tiny distinctions [1]. Apart from the frame-based framework, in this paper we propose a vector-based speaker modeling method, to which common support vector machine (SVM) kernels can be applied. The vector-based speaker model is composed of the concatenation of the average acoustic representations of all phonemes. SVM and GMM classifiers are compared on the speaker models. Moreover, based on the observation that accents only differ in a limited number of phonemes, a variable selection framework is indispensable to select accent relevant features. We investigate a forward selection method, Analysis of Variance (ANOVA) , and a backward selection method, SVM - Recursive Feature Elimination (SVM-RFE). We find that the multiclass SVM-RFE achieves comparable performance with the ANOVA on optimally selected variable sets, while it obtains excellent performance with very few features in low dimensions. Results demonstrate the effectiveness of the proposed speaker models together with the SVM classifier both in low dimensions and in high dimensions as well as the necessity of variable selection.

Index Terms: variable selection, native accent identification, support vector machines, recursive feature elimination, cross validation, Gaussian mixture models

1. Introduction

Native accent identification (AID) consists in identifying a speaker's accent when the speaker is asked to speak his/her standard native language as much as possible. A possible application of AID is the selection of appropriate speech models and pronunciation dictionaries in automatic speech recognition (ASR) systems. Techniques used for LID may also be applicable to native AID if they only consider acoustic information. However, other approaches in language identification (LID), e.g., Parallel Phone Recognizer followed by Language Model (P-PRLM) [2], using phone tokenization of speech combined with a phonotactic analysis, may not be appropriate to the native AID because phonotactic differences between native accents are minimal or do not exist.

Recently, a support vector machine (SVM)-based approach [3] has been applied to LID and speaker recognition by deriving a new kernel, Generalized Linear Discriminant Sequence kernel (GLDS) to handle a frame-by-frame feature sequence in a Gaussian Mixture Model (GMM) and SVM fusion framework. However, unlike LID or speaker recognition, a straightforward phonetically blind GMM has shown to be insufficient to explore the tiny accent differences in the native AID task [1], whereas the maximum accuracy was only 27.0% on a 5 native accent classification problem. In this paper, instead of using the frame-based SVM framework, we propose an alternative approach, in which a speaker is represented by a high dimensional super vector, consisting of the acoustic representations of all phonemes (see Sec. 2). Then an accent model can be trained with a cluster of super vectors from the same accent. The benefits of this representation include two aspects. First, the number of features in each super vector is identical for each speaker, resulting in the feasibility of using the common SVM kernels, such as the linear kernel or the Radial Basis Function (RBF) kernel. The other advantage is to utilize the merits of SVM in handling the high-dimension, small-sample case. The major weakness of this approach is that a phonetic segmentation is required in order to build the speaker model. This can be achieved with high accuracy when using a kind of enrollment procedure. Otherwise one will need to rely on a segmentation provided by a generic speech recognizer.

For native AID, one crucial problem is that not all phonemes change under the influence of an accent, and that certain phonemes only change in certain regions/accents; furthermore, not all features of a certain phoneme contribute to accent discrimination. For example, in the Dutch word 'vos' (fox), the voiced fricative /v/ is devoiced by Dutch speakers and not by Flemish speakers. This is a simple illustration that only a fraction of the features contains accent discriminative information and the distinction may be quite subtle. In [1], we proposed to use ANOVA to evaluate the importance of features, from which features with small p-values are selected into the variable set incrementally. Recently a recursive feature elimination method has been proposed in the context of SVM classification [4]. Unlike ANOVA, it gradually eliminates a variable whose removal changes the objective function used by the SVM the least in a sequential backward elimination manner. The SVM-RFE was extended to the multiclass case in [5] afterwards.

In this paper we will investigate different variable selection and classification algorithms and how to combine them in an optimal way. The remainder of this paper is organized as follows. In Section 2 we describe the construction of the speaker vectors. In Section 3 we develop multi-class SVM and different variable selection techniques. In Section 4 we present experimental results and give an interpretation of the different performances, followed by the conclusions.

2. Vector-based speaker model for AID

The diagram of our AID system is shown in Fig. 1. A series of 12th-order Mel cepstra plus energy is extracted from 30ms-length speech frames, followed by cepstral mean subtraction (CMS). In order to investigate the different behaviors

This work is sponsored by the FWO project G.0260.07 TELEX: Combining acoustic TEmplates and LEXical modeling for improved speech recognition, and by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical an Educational applications.

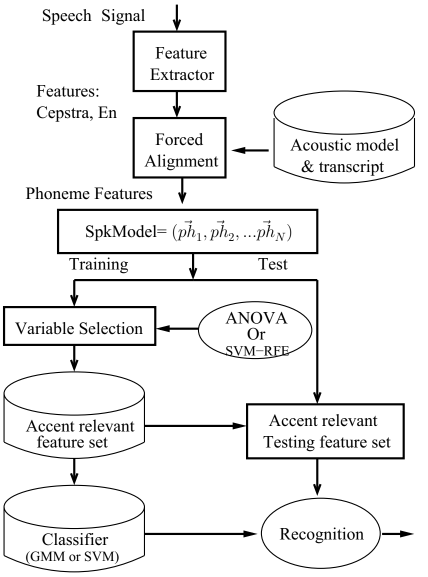

Figure 1: The diagram of the native AID system

of phonemes between accents, we assume that a text transcript is priorly known, with which then all speech is aligned to the state level by the ESAT speaker independent large vocabulary recognizer [6]. To reduce the influence of random fluctuations, context-independent phoneme features are averaged over all occurrences of the phoneme spoken by the same speaker. Given three-state acoustic Hidden Markov Models (HMMs), a speaker model $\mathbf{S}$ is represented by a super-vector composed of all phonemes: $\mathbf{S} = (\vec{ph}_1, \vec{ph}_2, ..., \vec{ph}_M)$, where $\vec{ph}_m, 1 \leq m \leq M$ is the concatenation of all cepstra and energy features in the three states for this phoneme and M is the number of phonemes. Since we can not guarantee the existence of all phonemes, non-occurred phoneme features for a training speaker model are replaced by the means of the corresponding occurred features of other speaker models in the same accent, while in testing, they are substituted by the global means over all accents. This representation results in a speaker model with dimension $38(phonemes) \times 3(states) \times (12(mel) + 1(En)) = 1482$. Four phonemes which do not appear frequently are removed before the variable selection, leading to a 1326-dimensional vector for a final speaker model. Suppose there are K accent classes, and in each class there are $n_k (1 \leq k \leq K)$ speakers, all speaker models $\mathbf{X}$ and their labels $\mathbf{Y}$ can be written as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{S}_{1,1} \\ ... \\ \mathbf{S}_{1,n_1} \\ \mathbf{S}_{2,1} \\ \mathbf{S}_{2,2} \\ ... \\ \mathbf{S}_{K,n_K} \end{pmatrix}_{N \times 1326} \qquad \mathbf{Y} = \begin{pmatrix} 1 \\ ... \\ 1 \\ 2 \\ 2 \\ ... \\ K \end{pmatrix}_{N \times 1} \tag{1}$$

where $\mathbf{S}_{k,j}$ denotes the speaker model of the j-th speaker in the k-th accent, $N = \sum n_k$. The usage of variable selection is to select a set of columns in $\mathbf{X}$ such that the generalization error is minimized. Two variable selection methods, ANOVA and SVM-RFE, are attempted respectively to select accent relevant features, with which classifiers, GMM or SVM, are trained to identify the accent of a test speaker model.

3. From SVM to SVM-RFE

3.1. Multiclass SVM

SVM [7] is one of the most popular supervised classifiers on a wide range of data sets due to its superior performance. It looks for an optimal hyperplane as a decision function in a high-dimensional space by simultaneously minimizing the empirical classification error and maximizing the geometric margin. For a two-class problem, suppose input vectors $\mathbf{x}$ are mapped into a higher dimensional space by a mapping function $\varphi(\mathbf{.}) : \mathbf{z} = \varphi(\mathbf{x})$, where $\mathbf{z}$ denotes vectors in the high dimensional space. A optimal hyperplane can be found to classify the samples from two classes:

$$f(\mathbf{z}) = \mathbf{w}^T\mathbf{z} + b, \tag{2}$$

where the weight vector $\mathbf{w}$ is a linear combination of training patterns, and $\mathbf{b}$ 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 support vectors. The mapping function is implicitly defined by a kernel function $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$. In this paper, the linear and the RBF kernels are explored.

The two-class SVM classifier can be extended to the multiclass case by combining several binary SVMs with the one-versus-one (OVO) strategy or the one-versus-all (OVA) strategy.

3.2. SVM-RFE

When the number of variables is significantly more than the number of samples, e.g. speakers per accent, the generalization ability of the classifier is restrained. Eliminating some noisy variables prevents overfitting, as we will show in Sec. 4. Recursive feature elimination based on SVM (SVM - RFE) [4, 8], derived from the classical SVM, is a variable ranking algorithm to evaluate the contribution to the classification error of each variable. In Eq. 2 the squared weight w_i^2 of the i-th variable of the weight vector $\mathbf{w}$ is a scale measuring the width of the margin that this variable contributes to the hyperplane. The larger the margin is, the more informative this variable will be. A variable is eliminated if its removal influences the classification hyperplane the least. This procedure is repeated until a desired number of variables are retained. To limit the computational cost, sometimes without losing much accuracy, SVM-RFE allows to eliminate a few variables simultaneously.

In [5], a multiclass SVM-RFE is proposed. Suppose T linear binary SVMs are obtained by the OVO or the OVA strategy, the ranking score of each variable is computed as:

$$c_i = \frac{\overline{v}_i}{\sigma_{v_i}}, \tag{3}$$

where $\overline{v}_i$ and σ_{v_i} are mean and standard deviation of $v_{ij} = (w_{ji})^2$, and w_{ji} is the weight value associated with the i-th feature from the j-th linear binary SVM.

Although both ANOVA and SVM-RFE are *wrapper* variable selection techniques, their flavors are different: ANOVA aims to measure the ratio between the variation within classes and the variation between classes, under the assumption of a Gaussian distribution for each variable in one class. The importance of variables is indicated by their p-values under the H_0 hypothesis that the means of a variable are equal over different accents. The ANOVA selection is a *forward selection* procedure: one variable is added into the selected feature set in one step in terms of their ranked p-values. On the contrary, SVM-RFE is a *backward elimination* procedure which progressively eliminates variables from the remaining variable set. By now there is still no decisive conclusion about which one is superior to the other, although it is often agreed that forward selection is computationally more efficient while backward selection assesses the importance of variables in the context of other variables [9].

4. Experiments

4.1. Database and human perception

All experiments are performed on the CoGeN database, containing 101 male and 73 female speakers. Speakers were asked to read five paragraphs of standard Dutch, which yielded about 5 minutes of speech per speaker. Speakers are grouped in accents corresponding to the five Flemish provinces, namely Antwerp (A, 42 speakers), Brabant (B, 26 speakers), Limburg (L, 34 speakers), Oost-Vlaanderen (O, 36 speakers) and West-Vlaanderen (W, 36 speakers). Dialect studies show that the Flemish provinces correspond fairly well with dialect regions.
On this database, we set up an informal human perception experiment. Ten native, but untrained members of our research team were asked to identify the accent of speakers after listening to a few 30-second speech segments. The maximum accuracy was about 63%, and the average over ten listeners was 45%, implying the native accent identification on this database is a difficult task even for human beings.
All experiments are performed by the leave-one (speaker)-out (LOO) strategy: one speaker is left for evaluation and the remaining 173 speakers compose the training set. Because of the LOO, in variable selection the selected features may differ from one training set to another. In case of the necessity of tuning hyper-parameters, for instance, utilizing SVM as the classifier, 10-fold cross-validation is adopted over the training set.

4.2. Classifiers for the native AID

In our previous work [1], we used ANOVA as a feature selection tool and GMM as a classifier. In modeling, a speaker model is estimated using all speech available from each speaker and then an accent specific GMM is trained with the speaker vectors from the same accent. In this experiment, leaving the feature selection method untouched, we compare the classification capabilities of GMM and SVM [1]. For GMM, the number of mixtures is set to 1, and a diagonal covariance is used. We have verified that increasing the number of mixtures does not improve the performance due to the limited number of speakers. With SVM, two kernels, namely the RBF kernel and the linear kernel, are studied. In decomposition of binary SVMs, there is no significant difference between the OVA and the OVO strategies. Here we only show the result of the OVA. At a pre-defined number of selected features, the classifiers are trained using the same variable set. The accuracy curves of the classifiers at different numbers of variables are shown in Fig. 2 when we use all speech of a test speaker to achieve his/her speaker model. The best accuracies for all classifiers are achieved in the range of 40-100 features, indicating that 90% of the features in the original speaker vectors are uninformative. Overall, the SVM classifiers outperform the GMM classifiers with the same number of features. Also the performance of SVM degrades much slower with increasing number of features than in the case of GMM classification. Nevertheless, variable selection remains an essential ingredient for success, also for the SVM classifiers. Finally, there is only a marginal difference between the linear and RBF kernel, with a small preference for the latter.

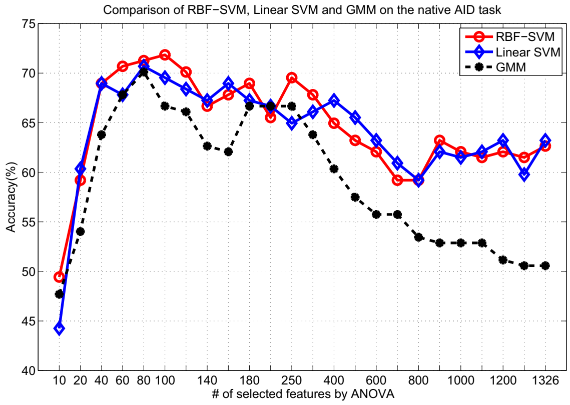

Figure 2: Comparison of RBF-SVM, linear SVM and GMM classifiers on the native AID.

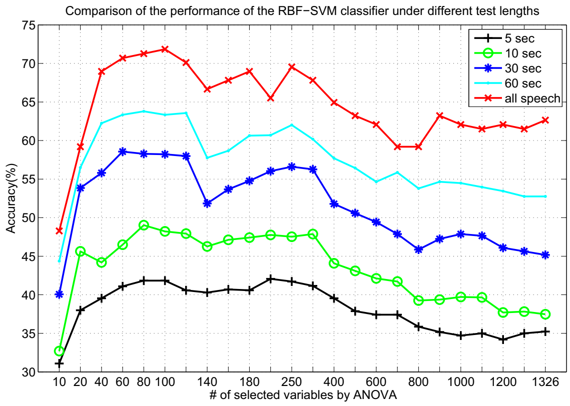

Figure 3: Comparison of the performance of the RBF-SVM under different testing lengths.

4.3. Different test lengths

The test speaker model in Sec. 4.2 is estimated by all speech of a test speaker. In this experiment we investigate the performance of the RBF-SVM classifier under different test lengths. The test lengths we choose are 5, 10, 30 and 60 seconds. A speech signal with a certain length is randomly segmented from the speech of the test speaker. The features at a certain number are still decided by the ranking of ANOVA. The experiment is repeated 10 times and the average accuracies are shown in Fig. 3. As can be seen, the performance of the AID increases as the test length grows monotonously. 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.

4.4. ANOVA vs. SVM-RFE

Fig. 2 and 3 illustrate that the variable selection helps a classifier to be less influenced by indiscriminative variables. In the next experiment, two variable selection methods, ANOVA and SVM-RFE, are compared. Due to the great computational cost brought by the SVM-RFE if we start with the complete variable set, we use ANOVA as a pre-selection process: 500 variables

[1] We also performed experiments with a k Nearest Neighbor classifier. However its performance was not at all competitive, probably due to the small number of speakers per class.

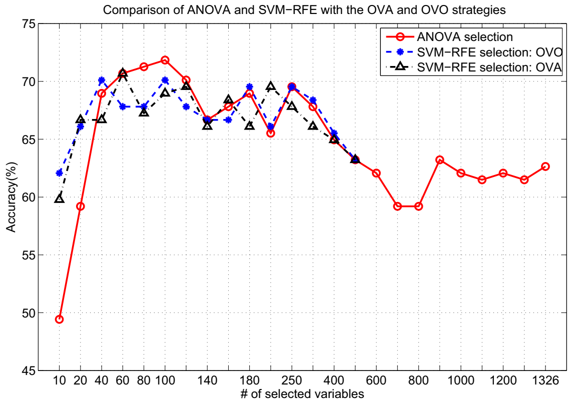

Figure 4: Comparison of ANOVA, the OVO and OVA based SVM-RFE strategies on the native AID.

are kept as an initialization after more than half of the variables whose p-values of ANOVA are large are thrown away. With these 500 variables, the SVM-RFE is iterated by the following rule. If the size of variable set is larger than 215, 5 variables are eliminated at the same time; otherwise only 1 variable is eliminated. After each elimination, the decision function is recalculated. The accuracies of ANOVA, the OVO based SVM-RFE and the OVA based SVM-RFE at different numbers of variables are shown in Fig. 4, with the RBF-SVM as the classifier. Note that in this figure, the selected variables by the three methods may differ even for the same test speaker. As illustrated, at a moderate number of variables, there is no big difference between ANOVA and the SVM-RFE; however, when only a small number of features are selected, SVM-RFE is much superior to the ANOVA variable selection.

4.5. Discussions

In Sec. 4.2, we have shown that all classifiers benefit from the variable selection procedure. With the same variable set, the SVMs outperform GMM, especially in high dimensions. This verifies SVM's insensitivity to the number of dimensions. On the other hand, in low dimensions, the RBF-SVM also gives the best result while the performance of the linear SVM is even worse than the GMM. This might be a consequence of the complexity of our task: a linear hyperplane with few features is not enough to separate the subtle difference among accents satisfactorily. When around 100 variables are selected, both the SVMs and the GMM reach their maximal accuracies. We observe that the single Gaussian classifier obtains comparable performance as the SVMs do. A possible reason for the excellent performance of the GMM is that the one Gaussian distribution fits perfectly the assumption of ANOVA. While the variables are ranked and selected based on the assumption, it is not surprising that the single Gaussian shows good classification ability.

Concerning the comparison of the variable selection methods, we have seen that the SVM-RFE works much better than the ANOVA in low dimensions. This phenomenon can be explained like this: the p-values of variables are the only measurement for the ANOVA to decide the order of selected variables; it does not care about the dependency and correlation between the potential candidates and the current variable set in the procedure of forward selection. Oppositely, the SVM-RFE concerns about the whole decision hyperplane and eliminates the least influencing one, which implicitly analyzes the correlation between features. For example, the second order of Mel cepstrum of phoneme I_2, whose p-value is extremely small, seems to be highly discriminative for the accents "A", "B" against the accents "L", "O" and "W". In our speaker modeling, the three correlated I_2s from three HMM states are highly ranked, implying they have a high chance to be selected even when the desired number of variables is limited. By this selection, in low dimensions, the variable set selected by ANOVA often contains correlated variables. But adding one or more I_2s to the selected variable set does not increase the discrimination, as long as the set already holds one I_2. As a backward elimination method, SVM-RFE would throw the correlated features away in the early elimination and only keep the most significant one among the three.

5. Conclusions

In this paper, we construct vector-based speaker models for the native AID task and compare the performance of SVM and GMM classifiers. We find that the SVM is superior to the other uniformly. Although variable selection reduces the confusion brought by accent irrelevant variables, the insensitivity of the SVM classifier to the *curse of dimensionality* makes it suitable for the native AID in high dimensions. Variable selection plays a crucial role to boost the discrimination between accents regardless of different lengths of test segments. We also investigate two different variable selection techniques in either forward or backward way and observe that only using very few variables selected by the multiclass SVM-RFE is able to achieve quite excellent performance; but it does not outperform the ANOVA selection in the neighborhood of their optimal variable sets.

6. References

[1] TY. Wu, D. Van Compernolle, J. Duchateau, Q. Yang, and J-P. Martens, "Improving the discrimination between native accents when recorded over different channels," in *InterSpeech*, Lisbon, Portugal, Sept 2005, pp. 2821–2824.

[2] P.A. Torres-Carresquillo, D.A. Reynolds, and J.R. Deller, "Language identification using gaussian mixture model tokenization," in *Proc. ICASSP*, Orlando, USA, May 2002, pp. 1375–1378.

[3] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[4] I. Guyon, J. Weston, S. Barnhill, and N. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[5] K-B. Duan, J.C. Rajapakse, and M.F. Nguyen, "One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification," in *EvoBIO*, 2007, pp. 228–234.

[6] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Comm.*, vol. 24, no. 1, pp. 5–17, 1998.

[7] V. Vapnik, *Statistical Learning Theory*, Wiley Interscience, 1998.

[8] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *Journal of Machine Learning Research*, vol. 3, pp. 1357–1370, 2003.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.