# ACCENT CLASSIFICATION USING SUPPORT VECTOR MACHINES

*Carol Pedersen*, Joachim Diederich*^*
*School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Australia
^Department of Computer Science, The American University of Sharjah, Sharjah, United Arab Emirates
{carol , joachimd}@itee.uq.edu.au

## ABSTRACT

*Accent is the pattern of pronunciation which can identify a speaker's linguistic, social or cultural background. It is an important source of inter-speaker variability, and a particular problem for automated speech recognition. Current approaches to identification of speaker accent may require specialised linguistic knowledge or analysis of the particular speech contrasts, and often extensive pre-processing. An accent classification system using time-based segments of Mel Frequency Cepstral Coefficients and with Support Vector Machines is studied for a small corpus of two accents of English. Over one- to four-second samples from three topics, accuracy is up to 75% to 97.5% with very high recall and precision. Its use with mis-matched content is at best 85%, with a tendency towards majority-class classification if the accent groups are significantly imbalanced.*

## 1. INTRODUCTION

The effectiveness of an Automatic Speech Recognition System (ASR) is greatly reduced when the particular accent or dialect in the speech samples on which it is trained differs from the accent or dialect of the end-user [1]. Therefore, correct identification of a speaker's accent, and the subsequent use of the appropriately trained system, can be used to improve the accuracy of the ASR application. Current approaches to classifying accent often involve training several ASRs on different varieties of accented speech, and choosing the best performer as the indicator of the accent [2][3][4]. This is labor-intensive and requires specialised phonetic knowledge to transcribe and label the data. Training ASRs requires very large amounts of data, which is generally not available for accented speech, especially for some less studied or less populous accents. Other methods usually involve some prior knowledge or training on specific linguistic features [4]. The accuracy of such systems greatly depends on the method used, the accents investigated and the restrictions placed on the input speech samples, and ranges in the order of 65% to 98.5% [3][5][6].

Mel Frequency Cepstral Coefficients (MFCCs) provide an efficient means of representing the frequency components of the speech waveform, and are the most widely used feature in state-of-the-art speech recognition systems. The industry standard speech recognizer front end includes calculation of the 13 absolute MFCCs and their first and second-order derivatives (a total of 39 MFCCs). For use in an ASR the MFCCs are usually combined into phoneme units, however this requires further segmentation and identification using a pre-trained system. Since phonemes in continuous speech are approximately 60-70ms in average duration [7] and the actual identity of the units is not of concern, it may be possible to use time-based segments rather than phoneme-based segments for the simple classification task. The optimum duration of these segments would be an important part of the investigation.

Support Vector Machines (SVMs) are a class of algorithms which are well-suited to learning classification and regression tasks. They usually combine a linear learning machine (LLM) with a kernel which makes it possible for the LLM to work in a highly-dimensional feature space, since only inner products of data points are used rather than the input features themselves. The margin between classes is maximised in order to find the best possible separator, and further optimised in the presence of noisy data by the introduction of slack variables.

SVMs have been designed for high-dimensional input spaces. Speech provides the opportunity for working with a very large number of features. Very large numbers of samples of accented speech are not generally available, and the numbers of samples from the different accent groups may be imbalanced, hence investigating the performance of SVMs is an important task in these contexts. A small number of samples increases the chance of overfitting, and as a result, the performance of the SVM has to be tightly controlled [8].

This paper presents an analysis of an accent classification system using SVMs with MFCC features in time-based segments as inputs. The length of speech sample required for good performance, as well as the duration of the temporal segments is investigated for three samples of differently accented speech.

## 2. SPEECH DATA AND FEATURE EXTRACTION

A corpus of accented speech was collected from 40 male and female adult subjects in two groups, Arabic (n=27) and Indian (n=13) accents of English. Subjects read a

single page of English text on each of three topics. The speech samples were recorded using a close-talk head-mounted microphone, onto computer as 16bit WAV files at 16KHz sample rate. All recordings were made in the same location under identical conditions in order to minimise channel effects.

Three sections of speech samples were chosen for initial analysis, one from each topic, and each 10 seconds long. Samples were trimmed to 50ms before the start of the relevant section in order to minimise the effect of potential edge-related effects on parameters. Analysis was conducted on samples of between 1 and 10 seconds in duration, in 1 second steps, all starting at the same "zero" point.

The samples were processed to obtain energy and 12 basic MFCCs, their velocity and acceleration parameters (first and second order derivatives). The method included cepstral mean subtraction and energy normalisation in order to minimise any recording differences. The processing resulted in 39 features for each 10ms frame for the duration of the speech sample, giving 3900 features for each second of sample duration.

Because 10ms is a very short time relative to the length of many phonemes, each feature was averaged across a number of frames in order to obtain values for larger time segments. The procedure was repeated for segments of 10ms (that is, no averaging) to 150ms.

## 3. EXPERIMENTS

The sequence of averaged MFCCs for a particular sample was used as the feature vector for the particular subject. Leave-one (speaker)-out cross-validation (LOO) was used for performance evaluation, focusing on accuracy, precision and recall parameters and ROC curve analysis.

Recall and precision are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \textit{Equation 1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \textit{Equation 2}$$

where TP, FP and FN are the number of true positives, false positives and false negatives respectively.

### 3.1. Matching content

The three topics were initially analysed separately. Experiments were repeated for each of the sample duration and segment size combinations. Various kernel designs (linear, polynomial, RBF) were investigated for the binary classification task.

The number of MFCCs per segment was also reduced from 39 to 13 (that is, excluding the first and second order derivatives) and the analysis repeated for all duration-segment combinations.

### 3.2. Non-matching content

A series of LOOs was performed using training samples from one topic set and testing samples from each of the other two topics, in order to test the effect of sample content mismatch.

A further series of tests was performed by adding extra non-matching samples to the training set of the third topic, 1 second sample, 100ms segment case, and conducting LOOs. Samples were sourced either from within the same topic or from the other two topics. Both the number of samples and the proportion of samples from the two accent groups were varied in order to study the effect of sample number and group imbalance.

## 4. RESULTS

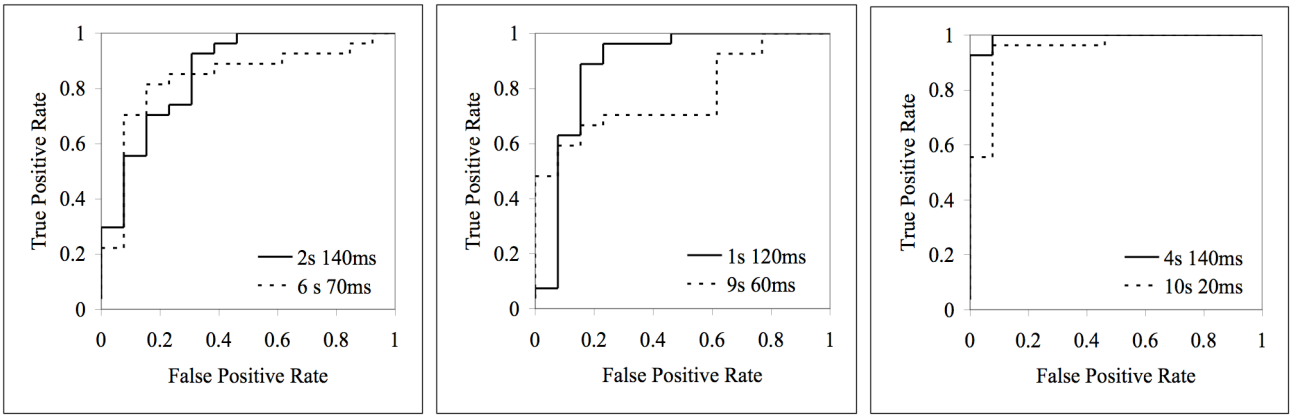
### 4.1. Matching content

Classification results varied by topic, sample length and segment size. The results for 13 and 39 features per segment were almost identical in all evaluation parameters, therefore results for 13 MFCC features will be presented. Best results were obtained using a linear SVM, for the third topic and 4 seconds sample duration or less (Table 1). Classification accuracy ranged from 75% to 97.5% at best, with very high precision and recall. Accuracy, recall and precision fell as sample duration increased from these peak results. Accuracy was slightly higher (mean 2.5 percentage points) for longer segment durations. Recall did not change with segment duration, and precision increased by an average of 1.1, 3.2 and 5.6 percentage points for the first, second and third topics respectively, as sample duration increased from 10ms to 150ms.

**Table 1.** Performance - matching content case

| Topic number | Accuracy (%) | Recall (%) | Precision (%) | Sample Duration (s) | Segment Duration (ms) |
|---|---|---|---|---|---|
| 1 | 75 | 92.59 | 75.76 | 2 | 140 |
| 2 | 87.5 | 96.3 | 86.67 | 1 | 30,40,60-80, 120-150 |
| 3 | 97.5 | 100 | 96.43 | 1 | 130 |
| | | | | 4 | 60, 80-110, 140 |

Selected ROC curves are presented in Figure 1 for examples of best and worst cases (by accuracy, precision and recall) for each topic. Area under the ROC curve is shown in Table 2.




(a) Topic 1 (b) Topic 2 (c) Topic 3

**Figure 1.** ROC curves, matching content

**Table 2.** Area Under ROC Curve - matching content

| Topic number | Sample Duration (s) | Segment Duration (ms) | Area Under ROC Curve |
|---|---|---|---|
| 1 | 2 | 140 | 0.8604 |
| | 6 | 70 | 0.8348 |
| 2 | 1 | 120 | 0.8832 |
| | 9 | 60 | 0.7778 |
| 3 | 4 | 140 | 0.9943 |
| | 10 | 20 | 0.9516 |

### 4.2. Non-matching content

The effect of a mismatch between training and testing samples varied substantially between different training-testing combinations. Best results are achieved for training on topic 3 and testing on topic 2, with up to 85% accuracy, 82% recall and 87% precision. In contrast, training on topic 1 and testing on topics 2 or 3 resulted in 50%-70% accuracy, 97% recall and 67% precision, with almost all errors being misclassification of Indian samples as Arabic. These results varied little with increasing sample duration and segment size. Training on topic 2 and testing on topic 3 produced improved recall with longer samples and smaller frame sizes, but a drop in precision in both cases.

Adding more 1-second samples from topic 3 to the training set for topic 3 (1s samples, 100ms segments in all cases) had a negative effect on accuracy, precision and recall (Fig. 2). When all 1-second samples were included in the training set (a total of 400 patterns), accuracy was 68.25%, recall 86.3% and precision 72.14%. Area under the ROC curve in this case was 0.6728, compared with 0.9829 when only the first sample was used for training.

The ratio of Arabic- to Indian-accented samples was approximately 2:1 in all cases where there was only one sample for each subject in the training set. The effect of varying this imbalance, by adding extra samples to the minority group and then adding more of the majority group until the ratio was again 2:1, can be seen in Figure 3. Two topics were used in this case. Results are similar when samples from all three topics are used. Recall fell markedly as group numbers became more equal; accuracy and precision fell less. Error analysis showed that the type of error changed as the proportion of the groups changed. When group numbers were almost equal, errors were more equally distributed between the two accent groups, whereas when group numbers were more unequal, errors became almost exclusively those of misclassifying the minority group as the majority (Indian-accented as Arabic). Area under the ROC curve was best when the group numbers were most equal (0.6530).

## 5. CONCLUSION AND FUTURE WORK

The performance of the SVM classifier using time-based segments of averaged MFCCs as features was very high, with up to 97.5% accuracy, with a sample length of up to only 4 seconds. This compares favorably with a human listener study [9] conducted using the same samples, which yielded accuracy of 92.5% (range 80%-100%) after an average of 7.7seconds. Interestingly, error analysis revealed that SVMs mostly make mistakes on the Indian-accented samples while humans make almost all their mistakes on the Arabic-accented samples.

Classification accuracy appears to be dependant on the content of the speech sample under investigation, as shown by the different results for the various topics. When the content of a test sample is different from that on which the classifier is trained, accuracy can still be up to 85% but is often worse. Adding extra, non-matching samples in order to improve the feature-pattern ratio does not improve performance, and in fact may degrade performance further. This is likely to be due to the diversity of sounds across the samples (due to diverse speech content), being greater than the difference in sound realisation between the accent groups, as represented by MFCCs.

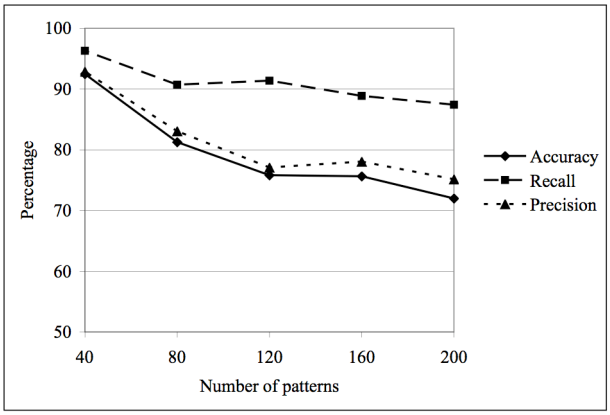



**Figure 2.** Effect of adding non-matching samples from the same topic

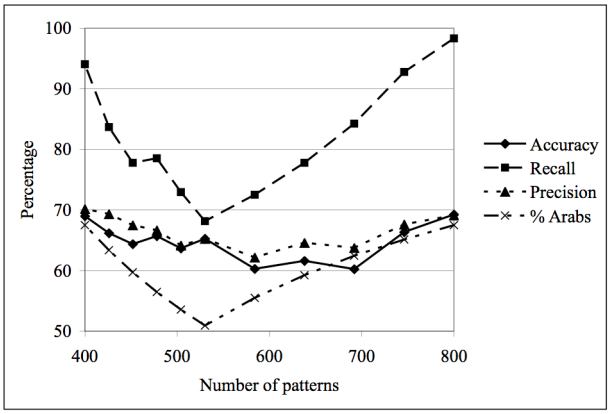



**Figure 3.** Effect of varying proportion of samples: two topic case

Many speech sounds are shared by different accents, and the nature of the variations that do occur can often be subtle and sparse. If strong contrasts in the speech sounds between the accents do actually occur in a short enough time (that is, over a few seconds, thereby avoiding excessive variation in content) the SVM-based classifier can be very effective in distinguishing between the accents, even without linguistic pre-processing or explicit identification of the individual contrasting speech sounds.

Not all MFCC-based features may be important for classifier performance, as was shown by the redundancy of the first and second order derivatives. Future work will focus on additional methods for feature selection, with the goal of minimising the number of features required, and extending the ability of the classifier to handle mis-matched data. Testing on other corpora is also an important priority. Emphasis will also be on knowledge initialisation of the SVMs by the use of domain knowledge to create virtual data sets in order to enhance classifier accuracy.

## REFERENCES

[1] M. Caballero, A. Moreno, A. Nogueiras, "Multidialectal Acoustic Modeling: a Comparative Study" in *Proc ITRW on Multilingual Speech and Language Processing,* Stellenbosch, South Africa, paper 001, April 2006

[2] K. Kumpf, R.W. King, "Automatic accent classification of foreign accentuated Australian English speech" in *Proc ICSLP 1996,* Philadelphia, PA, pp 1740-1743, October 1996.

[3] C. Teixeira, I.M. Trancoso, A. Serralheiro, "Accent Identification" in *Proc ICSLP 1996,* Philadelphia, PA, pp 1784-1787, October 1996.

[4] P. Angkititrakul, J.L.H. Hansen, "Use of trajectory models for automatic accent classification" in *Proc INTERSPEECH-2003/Eurospeech-2003,* Geneva, Switzerland, pp 1353-1356, September 2003.

[5] J. Frid, "Automatic classification of accent and dialect type: results from southern Swedish" in *Fonetic 2002 - TMH QPSR*, vol 43, pp. 89-92

[6] T. Chen, C. Huang, E. Chang, J. Wang, "Automatic Accent Identification Using Gaussian Mixture Models" in *Proc ASRU-01*, Madonna di Campiglio Trento, Italy, pp, 343-346, December 2001

[7] Y. Gong, W.C. Treurniet, "Duration of Phones as Function of Utterance Length and its use in Automatic Speech Recognition in *Proc Eurospeech-93*, Berlin, Germany, pp. 315-318, September 1993

[8] P. Golland, W.E.L. Grimson, M.E. Shenton, R. Kikinis, "Small sample size learning for shape analsis of anatomical structures" in *Proc. MICCAI-00,* Pittsburgh, PA, pp 72-82, October 2000.

[9] C. Pedersen, J. Diederich, "Listener Discrimination of Accent" in *Proc Human and Machine Speech Workshop, HCSNet Summerfest '06*, Sydney, Australia, p107, November - December 2006.