Accent in Speech Samples: Support Vector Machines for Classification and Rule Extraction

Carol Pedersen[1] and Joachim Diederich[1,2]

[1] School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Australia
[2] American University of Sharjah, Sharjah, UAE

1 Introduction

1.1 Motivation and Significance

Accent is the pattern of pronunciation which can identify a person's linguistic, social or cultural background. It is an important source of inter-speaker variability and a particular problem for automated speech recognition. This study aims to investigate the effectiveness of rule extraction from support vector machines for speech accent classification. The presence of a speaker's accent in the speech signal has significant implications for the accuracy of speech recognition because 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 [4] [14]. The correct identification of a speaker's accent, and the subsequent use of the appropriately trained system, can be used to improve the efficiency and accuracy of the ASR application. If used in automated telephone helplines, analysing accent and then directing callers to the appropriately-accented response system may improve customer comfort and understanding. The increasing use of speech recognition technology in modern applications by people with a wide variety of linguistic and cultural backgrounds, means that addressing accent-related variability in speech is an important area of ongoing research. Rule extraction in this context can aid in the refinement of the design of a successful classifier, by discovering the contribution of the various input features, as well as by facilitating the comparison of the results with other machine learning methods.

1.2 Overview

Current approaches to the identification of speaker accent usually require specialized linguistic knowledge or analysis of the particular speech contrasts

C. Pedersen and J. Diederich: *Accent in Speech Samples: Support Vector Machines for Classification and Rule Extraction*, Studies in Computational Intelligence (SCI) **80**, 205–226 (2008)
www.springerlink.com
© Springer-Verlag Berlin Heidelberg 2008

between the accents, and often extensive pre-processing of large amounts of data. In contrast, this chapter presents an accent classification system using time-based segments consisting of Mel Frequency Cepstrum Coefficients as features and utilizing rule extraction from SVMs (support vector machines). It is applied to a small corpus of two accents of English. Its performance is compared to two other machine learning techniques, and rule extraction is performed using a combination of SVMs and a rule-based learner.

2 Accent Recognition

2.1 Accent

Each person speaks in his or her own idiosyncratic way, but groups of people of a similar geographical or sociological background can be considered to share various common patterns in their speech. These include, but are not limited to:

- Pronunciation or acoustic features: the use of particular vowel and consonant sounds and how these change when they are combined in words and groups of words, as well as stress, tempo, rhythmic, and intonational factors [30]
- Grammar and vocabulary: morphology and syntax, vocabulary and idiom [30]

The combination of these patterns contributes to a sense of "accent" or "dialect" and can vary according to geographical origin, sex, social class, age, education, and whether the language being spoken is one's first or has been subsequently learned. Accent is usually considered to include only the effects of pronunciation or acoustic features, whereas dialect includes accent as well as grammar and vocabulary differences.

Accents are systematic and repeatable [27] [30]. Accents occur in most if not all languages with a sufficiently large number of speakers. However, accents are not set, "standard" entities, and considerable variation between people occurs. Accent is usually established early in life [30] but may be altered by, for example, living in another country for significant lengths of time, or significant speech training. A person's accent may also change in the short term, depending on to whom he or she is talking [30]. That is, humans are able to adjust their accent (usually along cultural or social lines rather than regional/geographic lines) in order to improve understanding and/or social acceptance [30].

Despite the variability in the realization of the spoken message due to different accents, and also differences in listener characteristics (such as listener attention and familiarity with the accent and speaker), both utterance understanding and accent identity are usually preserved [27]. Tatham and Morton [27, p114] note that "Speakers and listeners recognise the utterance

'behind' the accent. That is, they can identify... the same utterance spoken in different accents, and can readily disassociate the utterance from the accent."

Accent recognition by listeners can even occur in the absence of an identifiable message, that is, in unfamiliar languages [3] and in artificially manipulated speech (e.g. playing recordings backwards, removing features [22] [23] [29]).

Modern English is an important world language, serving as an international lingua franca in business, education, international relations and the media [7]. Many, if not most, speakers of English worldwide have English as their second or even third language, so advances in computerized speech technology will increasingly have to deal with a wide variety of accents if they are to be successful on the world stage. Similarly, speech technology based on other languages will benefit from research into accent and other sources of variation in human speech.

2.2 Automatic Speech Recognition

ASR is the recognition of human speech by use of computer analysis. An input speech signal is compared to a stored model of the various elements of spoken language (usually phonemes and their combinations) and the most likely sequence of words is produced. It is used in many applications, such as data entry, voice dialing, caller routing and translation assistance. Accuracy is highly dependant on the application domain and the training data used to build the stored acoustic and language models.

The effectiveness of an Automatic Speech Recognition System 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 [14]. Increasing the accuracy of ASRs on accented speech can be done in a number of ways. At its simplest, the accented speech can be passed through a number of ASRs trained on different accents, and the output is then evaluated for the most likely utterance. If there are a large number of accents, this "multiple processing" can be time consuming and expensive.

Alternatively, the recognizer itself can be made to identify multiple accented "versions" of the various phonemes and words. This leads to very large numbers of alternative representations within the recognizer, of words, phonemes or other units relevant to the operation of the ASR. This can actually reduce decoding accuracy [14] because of greater numbers of confusions and overlapping entities.

A third approach is to classify speech into its appropriate accent, and then to pass it through the appropriately accent-trained recognizer. The classifier would not necessarily need to fully decode the speech sample, but would only need to classify speech into the appropriate accent group.

If the aim is simply to identify an utterance as having one accent or another in, for example, a telephone call or recording, the usual approach involves training several ASRs on different varieties of accented speech, and choosing

the best performer as the indicator of the accent [17] [28] [1]. The training of ASRs is labor-intensive and requires specialized phonetic knowledge to transcribe and label the data. Training ASRs also 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 [1]. 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% [28] [9] [14].

For a simple classification task, the use of a fully developed ASR may not be required if the differences between particular accents can be learned by a supervised machine learning system. It may not even require knowledge of the specific linguistic differences between the accents of concern if the classifier can successfully learn from real speech examples. However, discovering the contribution of the various input features - e.g. through rule extraction - can aid in the refinement of the design of a successful classifier.

2.3 Mel Frequency Cepstrum Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) provide an efficient means of representing the frequency characteristics of the speech waveform, and are the most widely used feature in state-of-the-art speech recognition systems.

The standard speech recognizer front end includes calculation of the 13 absolute MFCCs and their first- and second-order derivatives (a total of 39 MFCCs). MFCC extraction is carried out on a speech sample using the following steps [19] [8]:

1. Compensation for the unequal sensitivity of human hearing across frequency.
2. Spectral analysis using a Fourier transform on 20–30 ms Hamming-shaped windows (frames) of speech every 10 ms.
3. Mel-scale filtering using a bank of triangular windows which become more compact at lower frequencies, in accordance with the sensitivity of human hearing.
4. Log compression of the mel-filterbank channels to model the relationship between the intensity of sound and its perceived loudness.
5. Discrete cosine transforms to produce the cepstral coefficients.
6. Cepstral mean subtraction and energy normalisation to reduce channel effects.
7. Extraction of derivatives (first- and second-order).

Following their widespread use in the speech domain, MFCCs have also been used successfully in the classification of music samples [13].

The first MFCC (MFCC(0)) shows a close correlation to the geometric energy in the (mel filtered) speech signal. MFCC(1) represents spectral slope, but beyond that it is less clear what the individual MFCCs are representing

in terms of how they relate to perceived aspects of speech and sound. Simple inversion of the MFCC extraction process does not generate a speech signal [20], since phase and fundamental frequency information are discarded. Relating individual MFCC values to individual aspects of human perception (such as pitch or particular phonemes) appears extremely difficult. This limits the use of rule extraction because rule sets may be useful during the experimental stage only.

MFCCs on a single time frame basis may not be particularly useful for accent classification, since they may encode a time too short to represent meaningful information. For use in an ASR, the MFCCs are usually combined into phoneme units; however this requires further segmentation of the speech sample and identification of phoneme units using a pre-trained system. Pure phoneme recognition rate is rather low [18] so further processing is required to produce a sequence of recognized speech, using scoring probabilities for phoneme combinations in the particular language and domain.

Since phonemes in continuous speech are approximately 60–70 ms in average duration [12] and the actual identity of the units is not of concern for simple accent classification, 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.

3 Rule Extraction from Support Vector Machines for Accent

3.1 Support Vector Machines

Support vector machines (SVMs) are a class of algorithms which are well-suited to learning classification and regression tasks. They have been used successfully on a wide variety of tasks, including text and image classification [15] [5] as well as bio-medical applications [11]. SVMs utilize kernels to work in a high-dimensional feature space, since only inner products of data points are used rather than the input features themselves. In classification tasks, the margin between the two classes is maximized in order to find the best possible separator, and is further optimized 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 [10].

3.2 Rule Extraction

There are three classes of users of speech recognition systems: (1) the engineer who explores features sets and designs the system, (2) the application expert who installs and maintains a speech recognition system (e.g. for directory assistance), and (3) the end user. The application expert is interested in the performance of the system but not necessarily in speech features while the end user is interested in a fast, reliable and accurate service and does not require detailed knowledge of the system. In this application, it is the engineer who employs machine learning and tests its performance who is interested in explanation. In this context, it is important to know *why certain input features lead to acceptable results* and while other feature sets fail.

As Craven and Shavlik (1994) [6] observe, "a (learning) system may discover salient features in the input data whose importance was not previously recognised." If a support vector machine has learned interesting relationships, these are encoded incomprehensibly as alphas and support vectors and hence cannot easily serve the generation of scientific theories. Rule extraction algorithms significantly enhance the capabilities of SVMs to explore data to the benefit of the user.

Support vector machines do not easily lend themselves to the discovery of explanations or rules that represent classification decisions. Unlike rule-based or decision tree systems, the output of the SVM is a numeric value and does not include additional information such as pattern elements or their combinations, which could be useful in explaining the knowledge obtained in the training process. Such explanations are important for the acceptance of SVM results by researchers and developers performing machine learning experimentation, and for the contribution SVMs can make to the knowledge in the domain in which they are operating. Rule extraction from SVMs is, therefore, an important advance for both the usefulness and verification of SVMs.

As indicated earlier, the relevant user in this case is the engineer/researcher who applies support vector machines for speech recognition and not the end user. That is, rule extraction from SVMs supports *experimentation* and *testing*, in particular the identification of features and feature sets that contribute to classification. A range of alternative methods are available, for instance sensitivity analysis. However, rule extraction from SVMs represents a convenient way to capture *the totality of knowledge learned by the SVM* (at least, this is the objective). The rule-based representation facilitates the comparison of SVM learning results with other machine learning techniques. Hence, the experimenter can select the best machine learning method for deployment.

Rule extraction is easily realized if SVMs are used in combination with other symbolic learning systems such as decision trees, when pattern labels predicted by an SVM (the "black box") are used as input labels for the second system (the "white box") in order to represent what the SVM has learned. The rules extracted may be few and simple, with high accuracy and fidelity [2],

however, this method faces severe limitations if the "white box" learner cannot accept high dimensional input patterns (see Chap. 1). In addition, small rule sets may not fully explain the decision-making process of the SVMs, i.e. the totality of knowledge learned by the SVM. Going from the high dimensional SVM to a lower dimensional learning system represents a loss of information and as a result, reduced fidelity and explanatory capability.

3.3 Objectives

If it is accepted that the intention in the utterance "behind" the accent is unaltered despite differences in realization due to accent differences - and this would be especially true in the case of "read" speech - then by comparing the manifest speech patterns for particular utterances, an appropriately trained classifier may be able to accurately identify the accent of the speaker from features derived from the speech signal, *even without explicit decoding of the intended message itself.* Knowledge of the particular acoustic or phonetic contrasts between various accents may not be necessary if the classifier is able to "learn" from examples rather than operate using the coding of known accent-related differences. This would make the extension of the system to previously unstudied accents a simpler and less time-consuming task. In addition, if accent-related differences were already known or were discovered, they could later be used to further enhance the effectiveness of the system. Support Vector Machines provide an ideal example of a classifier which is able to work with the high-dimensional inputs provided by speech. Rule extraction from SVMs in this context is useful in advancing further classifier design and for an explanation of the knowledge obtained

This chapter 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. Rule extraction is undertaken in order to identify the features which contribute to classification.

4 Methodology

4.1 Speech Data and Feature Extraction

A corpus of accented speech was collected from 40 male and female subjects in two groups, Arabic (n = 27) and Indian (n = 13) accents of English. Subjects were aged between 20 and 56 years (mean 27.8 y) and had a high to very high level of spoken English proficiency. Fully informed consent was obtained. Subjects read a single page of English text on each of three topics. Read speech was chosen to provide a uniform sample space, and because it is easier to elicit than spontaneous or conversational speech. The speech samples were

recorded using a unidirectional dynamic close-talk head-mounted microphone, via a mixer and USB audio interface onto computer as mono WAV signed 16 bit PCM (uncompressed) files at 16 KHz sample rate. All recordings were made in the same location under identical conditions in order to minimize channel effects.

Three sections of speech samples were chosen for initial analysis, one from each topic, and each 10 s long. Samples were trimmed to 50 ms before the start of the relevant section in order to minimize the effect of potential edge-related effects on parameters. Analysis was conducted on samples of between 1 and 10 s in duration, in 1-s 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 normalization in order to minimize any recording differences. A frame shift of 10 ms was used (that is, MFCCs were calculated every 10 ms) and a Hamming window of width 25 ms was used. There was therefore some overlap between adjacent frames, however the shape of the window means that most of the energy was in the center of the segment. The processing resulted in 39 features for each frame for the duration of the speech sample, giving 3,900 features for each second of sample duration.

Because a shift of 10 ms 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 segment "lengths" of 10 ms (that is, no averaging) to 150 ms.

4.2 Machine Learning Experiments

The sequence of averaged MFCCs for a particular sample was used as the input feature vector for the particular subject for SVM training and testing. The samples were not divided into separate training and testing groups due to the small number of samples; instead, leave-one (speaker)-out cross-validation (LOO; built-in to SVM^{light}, one of the tools use in this study [16]) was used for performance evaluation, focusing on accuracy, precision and recall parameters. In addition, ROC curve analysis was used.

Recall and precision are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

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

The three topics were initially analyzed separately. Experiments were repeated for each of the sample duration and segment size combinations.

Various SVM 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 MFCC derivatives) and the analysis repeated for all duration-segment combinations.

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. This was repeated for each combination of topic, sample length and segment size.

A further series of tests was performed by adding extra "non-matching" samples to the training set of the third topic (1-s sample, 100 ms segment case), and conducting LOOs. The "non-matching" samples had the same duration and segment size but were not from the same part of the speech sample as the original training set.

A subset of the 13-MFCC duration-segment combinations from each of the three topics was analyzed using a Decision Tree Learner (J48) [31] [24] and a Rule-based classifier (JRip) [31] in order to provide a comparison with the SVM results.

4.3 Rule Extraction and Evaluation

A variation of the pedagogical rule extraction method [21] [2] was used for rule extraction from the SVMs for the same subset of the duration-segment combinations for each topic that was used with the non-SVM machine learning methods.

In each topic, sample length and segment size combination, the model produced by SVM analysis was used to reclassify the original input patterns. The predicted class labels were then applied to the patterns to create a synthetic dataset which was used to train a rule-based classifier (JRip). The rules produced by the rule-based classifier were then examined in terms of accuracy and ROC curve analysis, and were compared to the performance on the original data, both in relation to the SVM and the rule-based classifier. Individual rules were examined and the elements of the rules compared with both the original JRip analysis and the original J48 analysis.

5 Results

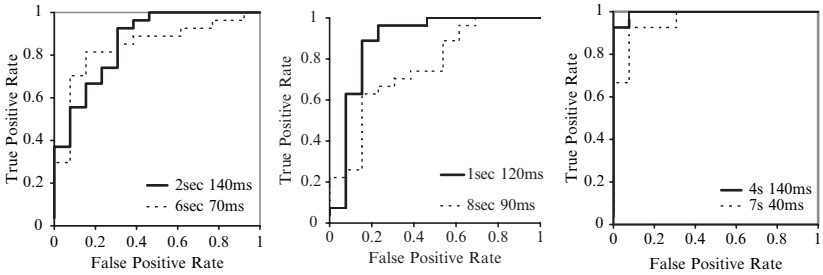
5.1 Machine Learning Experiments

Classification results varied by topic, sample length and segment size. The results for 13 and 39 features per segment were almost identical, therefore results for 13 MFCC features per segment will be presented.

Best results were obtained using a linear SVM, for the third topic and 4-s sample duration or less (Table 1). Best classification accuracy ranged from

Table 1. Performance of SVM – accuracy, duration, recall – for best cases

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
3	97.5	100	96.43	4	60, 80–110, 140



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

Fig. 1. ROC curves

Table 2. Area under ROC curve

Topic number	Sample duration (s)	Segment duration (ms)	Area under ROC curve
1[a]	2	140	0.8604
1[b]	6	70	0.8348
2[a]	1	120	0.8832
2[b]	9	60	0.7778
3[a]	4	140	0.9943
3[b]	10	20	0.9516

[a] Best case
[b] Poor case

75% to 97.5%, 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% 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% points for the first, second and third topics respectively, as sample duration increased from 10 to 150 ms.

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

The effect of a mismatch between training and testing samples varied substantially between different training-testing combinations. Best results were

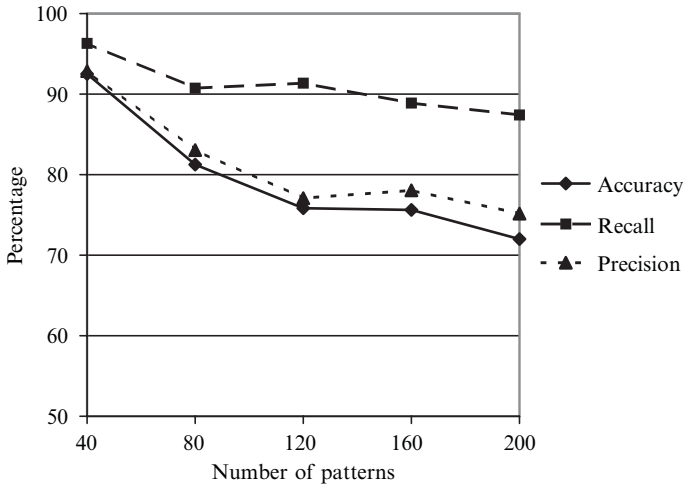


Fig. 2. Effect of adding non-matching samples from the same topic

achieved for SVM 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-s samples from topic 3 to the training set for topic 3 (1-s samples, 100 ms segments in all cases) had a negative effect on accuracy, precision and recall (Fig. 2). When all 1-s 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.

Decision Tree (J48) and rule-based (JRip) analyses were conducted on samples of 1, 2 and 4 s duration, with segments of 40 to 150 ms and 13 MFCCs per segment, in order to provide a comparison with SVM results, for most of the best cases in Table 1. Due to the small number of patterns, 40-fold cross-validation was done to most closely correlate to the SVM LOOs, and a comparison of the accuracy for each classifier (by topic) is shown in Table 3. The mean accuracy was calculated for each sample duration-text combination and these are shown in Fig. 3. J48 and JRip accuracy varied according to segment size much more than did SVM accuracy. Taking the topic 1, 2-s group as a typical example, SVM accuracy varied by up to 7.5% points across segment sizes from 40 to 150 ms, whereas J48 and JRip accuracy varied by up to 42.5 and 30% points respectively. J48 trees and JRip rules are shown in Tables 4 and 5 for the topic 1, 2-s cases.

Table 3. Comparison of LOO accuracy for various machine learning methods

Topic number	JRip accuracy (%)		J48 accuracy (%)		SVM accuracy (%)	
	Mean	SD	Mean	SD	Mean	SD
1	67.2	8.3	65.6	10.7	66.6	4.0
2	71.5	8.8	73.2	8.8	82.9	3.2
3	81.9	7.3	83.5	8.6	94.6	2.0

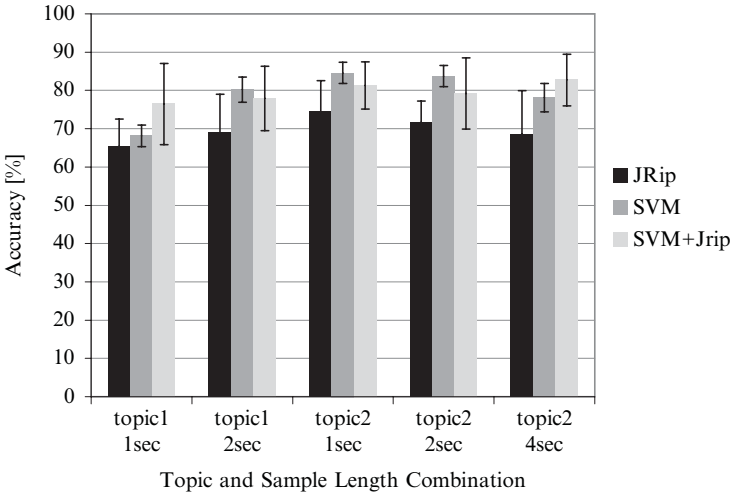


Fig. 3. Accuracy for various machine learning techniques

Table 4. Trees from J48 analysis on topic 1, 2-s samples

Segment size (ms)	Tree

40	`time6mfcc2 <= -4.595: Arabic`
	`time6mfcc2 > -4.595`
	`|  time3mfcc12 <= 1.622: Indian`
	`|  time3mfcc12 > 1.622: Arabic`
	`Number of Leaves :  3`
	`Size of the tree :  5`

50	`time5mfcc2 <= -3.177: Arabic`
	`time5mfcc2 > -3.177`
	`|  time3mfcc13 <= 0.005: Indian`
	`|  time3mfcc13 > 0.005: Arabic`
	`Number of Leaves :  3`
	`Size of the tree :  5`

Table 4. (Continued)

Segment size (ms)	Tree
60	time4mfcc2 <= -5.963: Arabic time4mfcc2 > -5.963 | time2mfcc4 <= 3.721: Arabic | time2mfcc4 > 3.721: Indian Number of Leaves : 3 Size of the tree : 5
70	time4mfcc13 <= -2.145: Indian time4mfcc13 > -2.145 | time2mfcc2 <= 4.046: Arabic | time2mfcc2 > 4.046 | | time24mfcc5 <= -0.078: Indian | | time24mfcc5 > -0.078: Arabic Number of Leaves : 4 Size of the tree : 7
80	time2mfcc8 <= -4.328: Indian time2mfcc8 > -4.328 | time3mfcc12 <= 4.537: Arabic | time3mfcc12 > 4.537: Indian Number of Leaves : 3 Size of the tree : 5
90	time3mfcc2 <= -4.493: Arabic time3mfcc2 > -4.493 | time2mfcc8 <= -0.409: Indian | time2mfcc8 > -0.409 | | time3mfcc1 <= -1.387: Indian | | time3mfcc1 > -1.387: Arabic Number of Leaves : 4 Size of the tree : 7
100	time3mfcc13 <= -1.73: Indian time3mfcc13 > -1.73 | time17mfcc7 <= -1.242 | | time1mfcc10 <= -2.199: Arabic | | time1mfcc10 > -2.199: Indian | time17mfcc7 > -1.242: Arabic Number of Leaves : 4 Size of the tree : 7
110	time9mfcc9 <= -6.577: Indian time9mfcc9 > -6.577 | time4mfcc2 <= -0.788: Arabic | time4mfcc2 > -0.788 | | time1mfcc2 <= 1.778: Arabic | | time1mfcc2 > 1.778: Indian Number of Leaves : 4 Size of the tree : 7

(*continued*)

Table 4. (Continued)

Segment size (ms)	Tree
120	time1mfcc6 <= 7.001 | time14mfcc7 <= -3.048 | | time4mfcc2 <= -2.548: Arabic | | time4mfcc2 > -2.548: Indian | time14mfcc7 > -3.048: Arabic time1mfcc6 > 7.001: Indian Number of Leaves : 4 Size of the tree : 7
130	time1mfcc6 <= 2.334 | time1mfcc7 <= -2.372 | | time1mfcc4 <= 4.738: Arabic | | time1mfcc4 > 4.738: Indian | time1mfcc7 > -2.372: Arabic time1mfcc6 > 2.334 | time10mfcc9 <= 3.71: Indian | time10mfcc9 > 3.71: Arabic Number of Leaves : 5 Size of the tree : 9
140	time1mfcc6 <= 2.285 | time2mfcc9 <= 4.912: Arabic | time2mfcc9 > 4.912: Indian time1mfcc6 > 2.285 | time3mfcc9 <= -4.871: Arabic | time3mfcc9 > -4.871: Indian Number of Leaves : 4 Size of the tree : 7
150	time1mfcc6 <= 0.311: Arabic time1mfcc6 > 0.311 | time2mfcc2 <= -3.38 | | time6mfcc8 <= -2.558: Indian | | time6mfcc8 > -2.558: Arabic | time2mfcc2 > -3.38: Indian Number of Leaves : 4 Size of the tree : 7

5.2 Evaluation of the Rule Extraction Results

Analysis was conducted on samples of 1, 2 and 4 s duration, with segments of 40–150 ms and 13MFCCs per segment, for all three topics.

The SVM-predicted class labels were found to be the same as the original class labels for all segment durations of the topic 3, 2 and 4-s duration cases, and for over 50% of the cases in topic 1, 4-s and topic 3, 1-s duration. Therefore, further analyses of accuracy and AUC were only conducted on topic 1

Table 5. Rules extracted using JRip on original data, topic 1, 2-s samples

```
40 ms (2 rules)
(time6mfcc2 >= -2.312) => class=Indian
 => class=Arabic

50 ms (2 rules)
(time5mfcc2 >= -2.905) => class=Indian
 => class=Arabic

60 ms (3 rules)
(time28mfcc7 <= -1.046) and (time5mfcc2 <= -0.249) => class=Indian
(time4mfcc1 <= -1.587) => class=Indian
 => class=Arabic

70 ms (2 rules)
(time24mfcc7 <= -1.006) and (time4mfcc9 >= 2.148) => class=Indian
 => class=Arabic

80 ms (2 rules)
(time3mfcc2 >= -4.391) and (time11mfcc7 >= -1.175) => class=Indian
 => class=Arabic

90 ms (2 rules)
(time3mfcc2 >= -3.551) => class=Indian
 => class=Arabic

100 ms (3 rules)
(time17mfcc7 <= -1.242) and (time1mfcc10 >= -1.917) => class=Indian
(time3mfcc13 <= -1.73) => class=Indian
 => class=Arabic

110 ms (3 rules)
(time17mfcc9 >= 3.455) and (time1mfcc1 <= 1.029) => class=Indian
(time3mfcc13 <= -2.838) => class=Indian
 => class=Arabic

120 ms (3 rules)
(time14mfcc7 <= -3.057) => class=Indian
(time1mfcc6 >= 7.082) => class=Indian
 => class=Arabic

130 ms (2 rules)
(time1mfcc6 >= 2.346) => class=Indian
 => class=Arabic

140 ms (2 rules)
(time2mfcc2 >= -3.543) and (time1mfcc6 >= -0.125) => class=Indian
 => class=Arabic

150 ms (2 rules)
(time2mfcc2 >= -3.843) and (time1mfcc6 >= 0.426) => class=Indian
 => class=Arabic
```

(1 and 2-s samples), and topic 2, (1, 2 and 4-s samples). Rule content was analyzed for all data.

Accuracy and Area Under Roc Curve

Mean results are shown in Table 6. Paired samples t-tests were performed to compare the various conditions. Accuracy and AUC were significantly worse for the original JRip analysis compared to both the SVM and JRip on data labeled with SVM-predictions (hereafter called "JRip Improved") ($p < 0.001$). There was no significant difference in accuracy between the original SVM analysis and JRip Improved analysis. AUC was significantly better for original SVM analysis than JRip Improved analysis ($p = 0.001$), which was in turn significantly better than original JRip analysis (all $p < 0.001$).

Mean Accuracy and AUC for the five topic-segment duration combinations are shown in Figs. 4 and 5. Where SVM accuracy was poorest (topic 1, 1-s samples), JRip Improved analysis had significantly greater accuracy than both SVM and JRip (paired t-test, $p = 0.021$ and 0.024 respectively) but AUC was not significantly different (at around 0.6). Only in the topic 2, 4-s sample case was the AUC for JRip Improved analysis significantly different from AUC for

Table 6. Accuracy and AUC for different learning systems

Learning system	Mean accuracy (%)	Area under ROC curve
JRip on original labels	69.79	0.5965
SVM on original labels	78.95	0.7316
JRip improved	79.5	0.6778

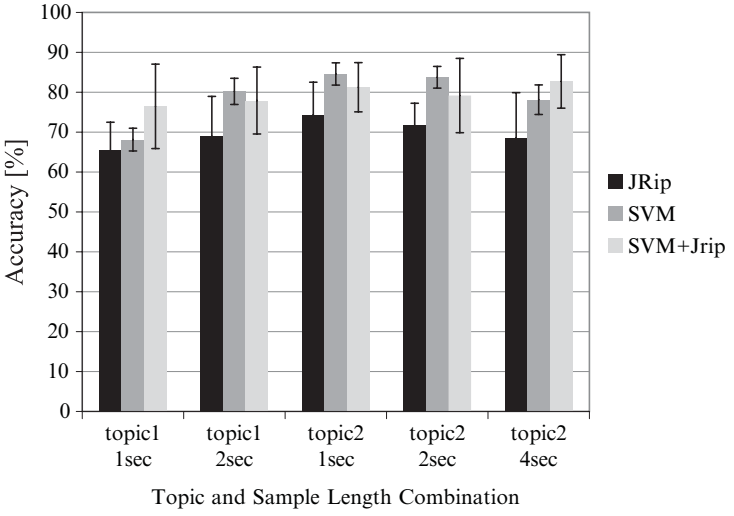


Fig. 4. Accuracy, various machine learning systems

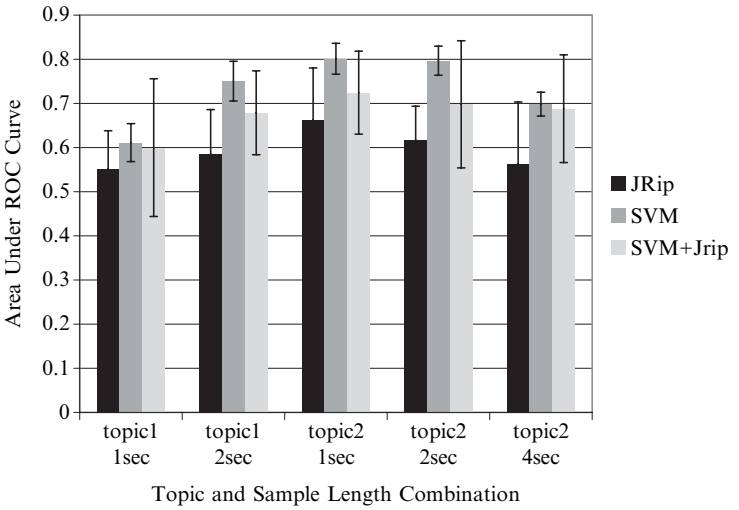


Fig. 5. Area under ROC curve, various machine learning systems

original JRip analysis (pairwise t-test, p = 0.008), whereas SVM AUC was significantly better than original JRip analysis for all but the topic 1, 1-s case.

Rules

Individual rules were examined for each topic, sample length and segment size, for J48, JRip and JRip Improved analysis.

MFCCs in the antecedents of the rules were identified by a segment time period and an MFCC number (e.g. time6mfcc2, signifying MFCC number 2 in the 6^{th} segment from the beginning). The MFCC number represents the same aspect of the speech signal regardless of the segment it is in, whereas each segment time period covered a different section of the speech signal (both in duration and location), depending on the segment size being analyzed.

In the JRip Improved analysis, there were 166 MFCCs mentioned in the antecedents of the 108 rule sets (all three topics and three sample lengths). MFCC number 3 was mentioned most often (15% of rules, in 25 rule sets), followed by MFCCs 1 and 2 (each 13.25% of rules) and MFCC 5 (12.65%), MFCC 7 (10.8%) and MFCC 6 (10.2%). The distribution of these MFCCs across the various topics varied greatly. Counting those MFCCs occurring in greater than 10% of the rule antecedents within a particular topic, MFCCs 1, 3 and 7 accounted for 83.3% of the antecedents for topic 3; MFCCs 2, 3, 5 and 6 accounted for 80.6% of the antecedents for topic 2, and MFCCs 2, 6, 7, 9 and 13 accounted for 76.8% of the antecedents for topic 1.

Sample rules are shown in Table 7 for JRip Improved analysis. In JRip Improved analysis, there were 108 rule sets (three topics, three sample lengths, 12 segment sizes). 68.5% of rule sets had only two rules, and 29.6% had three

Table 7. Rules for topic 2, 1-s samples extracted using Jrip on data labelled with SVM predictions (JRip improved)

```
40 ms segments (2 rules)
(time4mfcc5 >= 2.198) => class=Indian
 => class=Arabic

50 ms segments (2 rules)
(time3mfcc6 >= -0.076) and (time2mfcc2 >= 6.887) => class=Indian
 => class=Arabic

60 ms segments (3 rules)
(time4mfcc3 <= -5.189) => class=Indian
(time15mfcc11 >= 4.709) => class=Indian
 => class=Arabic

70 ms segments (2 rules)
(time2mfcc6 >= 1.696) => class=Indian
 => class=Arabic

80 ms segments (2 rules)
(time3mfcc3 <= -3.96) => class=Indian
 => class=Arabic

90 ms segments (2 rules)
(time2mfcc5 >= 2.514) => class=Indian
 => class=Arabic

100 ms segments (3 rules)
(time2mfcc5 >= 2.708) => class=Indian
(time1mfcc12 >= 2.972) => class=Indian
 => class=Arabic

110 ms segments (2 rules)
(time1mfcc6 >= 2.352) and (time1mfcc12 >= -2.99) => class=Indian
 => class=Arabic

120 ms segments (2 rules)
(time2mfcc6 >= -3.298) and (time1mfcc6 >= 1.41) => class=Indian
 => class=Arabic

130 ms segments (2 rules)
(time1mfcc6 >= 1.581) and (time2mfcc5 >= -3.744) => class=Indian
 => class=Arabic

140 ms segments (2 rules) 8
(time2mfcc3 <= -5.81) => class=Indian
 => class=Arabic

150 ms segments (3 rules)
(time2mfcc3 <= -4.848) => class=Indian
(time3mfcc5 >= 4.652) => class=Indian
 => class=Arabic
```

rules. Two rule sets had only one rule, that is, everything was classified as Arabic. Of the 246 rules, only 29 had more than one condition.

57% of the 60 topic-segment length combinations (topic 1, 1 and 2-s segments; topic 2, 1, 2 and 4-s segments) had at least one common MFCC mentioned in the antecedent of a rule (or in a tree) for both J48 and JRip analysis (on original data). The accuracy of JRip on these cases was significantly better than where there was no MFCC in common in the rule antecedent (73.3% vs. 65.2%, t-test $p = 0.001$). Exactly 50% of the combinations had at least one common MFCC mentioned in the antecedent in the rules of JRip and JRip Improved analysis. There was no significant difference in accuracy of JRip Improved analysis between these two groups. (77.99% vs. 77.01%). Exactly 50% of the combinations had at least one common MFCC in the antecedent of a rule in J48 and JRip Improved analysis. The accuracy of JRip Improved was significantly better if there was a common MFCC in the rule antecedent in both J48 and JRip Improved analysis (82.5% vs. 76.5%, $p = 0.0024$).

Out of the 30 MFCCs that were in common between J48 and JRip Improved analysis, only 19 were already in common between J48 and JRip (original). There was no significant difference in JRip Improved accuracy between these two groups. There were 34 rule sets where there was an MFCC in common between JRip and J48 analysis; 15 of these did not have that MFCC in common when the rules and trees for those cases in J48 and JRip Improved analysis were compared. In addition, 11 new common MFCCs had arisen between J48 and JRip Improved analysis.

6 Discussion

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 s. This compares favorably with a human listener study [25] conducted using the same samples, which yielded accuracy of 92.5% (range 80–100%) after an average of 7.7 s. Interestingly, error analysis revealed that SVMs mostly made mistakes on the Indian-accented samples while humans made almost all their mistakes on the Arabic-accented samples.

SVM accuracy also compared favorably with JRip and J48 classifier accuracy. SVM accuracy was 7.3% and 7.9% points higher on average than JRip and J48 accuracy respectively. There was much less variability in SVM accuracy than for the other classifiers, across the various segment durations. This means that the choice of segment size was much less critical for the good performance of the SVM system. SVM accuracy was, overall, slightly higher for longer segment durations than shorter ones.

Classification accuracy with SVMs appears to be dependant on the content of the speech sample under investigation, as shown by the different results for the various topics. Also, 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 [26]. 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 realization between the accent groups, as represented by MFCCs.

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.

SVMs do not easily provide rules or explanations for the classifications that they make, but in this study the rules provided by the "white box" learner also do not easily translate to knowledge about accent differences in the speech stream. Apart from the first MFCC (usually termed MFCC(0), but here termed "mfcc1"), which shows a close correlation to log Energy of a speech signal, and the second MFCC which represents spectral slope, it becomes increasingly unclear as to what exactly the higher individual MFCCs represent, in terms of actual speech sounds, despite their demonstrated usefulness in speech recognition and accent classification. In addition, because of the small number of patterns and their high dimensionality, rules extracted by the "white box" learner can vary greatly from one fold to the next, depending on which patterns are excluded at learning time. Nevertheless, rule extraction gives some indication which features of the high dimensional input space contribute to classification.

As mentioned above, there was no significant difference in accuracy between the original SVM analysis and JRip Improved analysis. However, AUC was significantly better for the original SVM analysis than JRip Improved analysis, which was in turn significantly better than original JRip analysis. It is worth noting that rule extraction resulted in the JRip Improved analysis which is a success in itself. The more conservative AUC evaluation confirmed the overall superiority of the SVM.

Not all MFCC-based features may be important for classifier performance, as was shown by the redundancy of the first and second order derivatives. The prominence of certain MFCCs in the rulesets of the various "white box" learners is also an indication that not all MFCCs are equally important for accurate classification. However, the contribution of individual MFCCs may not be fully captured in propositional rules such as those presented in this study. Investigation of more expressive rule languages may capture relations between features that are not represented in the rules presented above, but which are nevertheless important for the good performance of the SVM learner.

Future work will focus on additional methods for feature selection, with the goal of minimizing the number of features required, and extending the ability of the classifier to handle miss-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. Angkititrakul P, Hansen JLH (2003) Use of trajectory models for automatic accent classification. In: Proc INTERSPEECH-2003/Eurospeech-2003, Geneva, Switzerland, pp. 1353–1356, September 2003.
2. Barakat N, Diederich J (2005) Eclectic Rule-Extraction from Support Vector Machines. Int J Computational Intelligence 2(1):59–62.
3. Bond ZS, Stockmal V, Markus D, (2003) Sentence Durations and Accentedness Judgments. J Acoust Soc Am 113(4):2330–2331.
4. Caballero M, Moreno A, Nogueiras A (2006) Multidialectal Acoustic Modeling: a Comparative Study. In: Proc ITRW on Multilingual Speech and Language Processing, Stellenbosch, South Africa, paper 001, April 2006.
5. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification: Vapnik-Chervonenkis (VC) learning theory and its applications. In: IEEE Transcactions on Neural Networks, vol. 10, no. 5, pp. 1055–1064, September 1999.
6. Craven MW, Shavlik JW (1994) Using Sampling and Queries to Extract Rules from Trained Neural Networks. In: Cohen WW, Hirsh H (eds) Machine Learning: Proceedings of the Eleventh International Conference. Morgan Kaufmann San Francisco pp. 37–45.
7. Crystal D (1997) English as a global language. Cambridge University Press, Cambridge New York.
8. Davis SB, Mermelstein P (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In: IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 4, August 1980.
9. Frid J (2002) Automatic classification of accent and dialect type: results from southern Swedish. In: Fonetic 2002 – TMH QPSR, vol. 43, pp. 89–92.
10. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10): 906–914.
11. Golland P, Grimson WEL, Shenton ME, Kikinis R (2000) Small sample size learning for shape analysis of anatomical structures. In: Proc. MICCAI-00, Pittsburgh, PA, pp. 72–82, October 2000.
12. Gong Y, Treurniet WC (1993) 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.
13. Guo G, Li SZ (2003) Content-Based Audio Classification and Retrieval by Support Vector Machines. IEEE Transactions on Neural Networks 14(1):209–215.
14. Huang C, Chen T, Chang E (2004) Accent Issues in Large Vocabulary Continuous Speech Recognition. Int J Speech Technology 7:141–153.

15. Joachims T (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: ECML-98, 10th European Conference on Machine Learning, Heidelberg, Germany, pp. 137–142, April 1998.
16. Joachims T (1999) Making Large-Scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in Kernel Methods – Support Vector Learning, MIT Press.
17. Kumpf K, King RW (1996) Automatic accent classification of foreign accented. Australian English speech. In: Proc ICSLP 1996, Philadelphia, PA, pp. 1740–1743, October 1996.
18. Lin X, Simske S (2004) Phoneme-less heirarchichal accent classification. In: Matthews MB (ed) Signals, Systems and Computers 2004; Conference Record of the Thirty-Eighth Asilomar Conference on. vol. 2:1801–1804.
19. Milner B (2002) A Comparison of Front-End Configurations for Robust speech Recognition. In: Proc. ICASSP 2002, Orlando Florida May 2002.
20. Milner B, Shao X, (2007) Prediction of Fundamental Frequency and Voicing from Mel-Frequency Cepstral Coefficients for Unconstrained Speech Reconstruction. IEEE Transactions on Audio, Speech and Language Processing 15(1): 24–33.
21. Mitsdorffer R, Diederich J, Tan CNW (2002) Rule Extraction from Technology IPOs in the US Stock Market. In: 9th International Conference on Neural Information Processing. 4th Asia-Pacific Conference on Simulated Evolution And Learning. 2002 International Conference on Fuzzy Systems and Knowledge Discover. Orchid Country Club, Singapore, 18 November-22 November 2002.
22. Munro MJ (1995) Non-segmental factors in foreign accent: Ratings of filtered speech. Studies in Second Language Acquisition 17:17–34.
23. Munro MJ, Derwing TM, Burgess CS (2003) The Detection of Foreign Accent in Backwards Speech. In: Sole M-J, Recasens De, Romero J (eds) Proceedings of the 15th International Congress of Phonetic Sciences, (Barcelona). Causal Productions Australia. pp. 535–538.
24. Quinlan JR (2007) Data Mining Tools See5 and C5.0, Rulequest Research (2007) http://rulequest.com/see5-info.html.
25. Pedersen C, Diederich J (2006) Listener Discrimination of Accent. In: Proc Human and Machine Speech Workshop, HCSNet Summerfest '06, Sydney, Australia, p107, November–December 2006.
26. Pedersen C, Diederich J (2007) Accent Classification Using Support Vector Machines. In: Lee R, Chowdhury MU, Ray S, Lee T (eds) Proceedings 6th IEEE/ACIS International Conference on Computer and Information Science. Melbourne Australia, July 2007, pp. 444–449.
27. Tatham M, Morton K (2005) Developments in Speech Synthesis. Wiley, Chichester.
28. Teixeira C, Trancoso IM, Serralheiro A (1996) Accent Identification. In: Proc ICSLP 1996, Philadelphia, PA, pp. 1784–1787, October 1996.
29. van Els T, de Bot K (1987) The Role of Intonation in Foreign Accent. The Modern Language Journal 71(2):147–155.
30. Wells JC (1982) Accents of English: An Introduction. Cambridge University Press Cambridge New York.
31. Witten IH, Frank E (2005) "Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco.