On the use of Machine Learning Methods for Speech and Voicing Classification

Philip Harding, Ben Milner

University of East Anglia, Norwich, UK
p.harding@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work examines the effectiveness of machine learning (ML) classifiers on the problems of voice activity detection and voicing classification. A wide range of ML classifiers are considered and include parametric, probabilistic and non-probabilistic, artificial neural networks and regression. Evaluations are carried out in both stationary and non-stationary noise types at signal-to-noise ratios down to 0dB. In comparison to conventional methods the ML methods are found to be significantly more robust with multilayer perceptrons, Gaussian mixture models and Rotation Forest giving consistently best performance.

Index Terms: voice activity detection, mfcc, machine learning

1. Introduction

Voice activity detection (VAD) and voicing classification (VC) are well-known problems in speech processing, with numerous applications in noise estimation, speech recognition and speech coding. Many different methods have been proposed to solve these problems and they operate typically by measuring properties of the time-domain waveform such as zero crossing rate, spectral energy and spectral distortion [1]. These features are not robust to noise and so conventional methods either adapt threshold values according to noise levels or require the signal to be cleaned using speech enhancement methods [2]. More recently, machine learning (ML) techniques have been applied to these classification problems whereby features are extracted from the audio and either Gaussian mixture models (GMMs) [3] or support vector machines (SVMs) [4, 5] applied to classify the audio as non-speech, voiced speech or unvoiced speech.

The aim of this work is to make a comparative study of a range of ML classifiers on the two tasks of voice activity detection and voicing classification and to compare their performance against conventional techniques. A broad range of classifiers are considered and include parametric, probabilistic and non-probabilistic, artificial neural networks and regression. Some of these classifiers have previously been applied to speech processing applications while other classifiers chosen have not. Importantly, for the tasks of VAD and VC, the tests use speech that has been contaminated by noise as would be encountered in real situations. As computing processing power and storage increases it is useful to consider whether such machine learning classifiers have application to VAD and VC. Such methods also have advantages in that they learn classification boundaries from training data rather than requiring thresholds or constants to be determined as with conventional methods.

Before a classifier can be trained a feature vector must be chosen and this is discussed in Section 2. The classifiers are described in Section 3 along with a justification for their inclusion in this study. Experimental results for VAD and VC are presented and analysed in Section 4.

2. Feature selection

To apply machine learning techniques to VAD and VC a speech feature vector must be decided upon. Many different speech features have been proposed and used in speech recognition and speech coding applications, with one of the most successful being the MFCC vector. Previous investigations into VC also found MFCCs to be effective [3] and so for this work the MFCC vector will be used as the basic feature.

Static MFCC vectors, \( \mathbf{x} \), (comprising coefficients \( C_1 \) to \( C_{12} \)) are extracted from speech sampled at 8kHz using 23 overlapping mel-spaced triangular filterbank channels at a rate of 100 frames per second in accordance with the ETSI XAFE standard [2]. In addition to the basic MFCC vector, the zero'th coefficient can also be included as this gives a measure of energy which is useful in classification. Temporal derivatives can also be augmented to the static features and give additional information regarding rates of change of the vectors. Investigation into the effects of these combinations is presented in Section 4.2.

3. Classifiers

This section gives a brief description of each of the classifiers used in the comparison, with the aim of describing the algorithm at a high level so that basic principles and difference and similarities with other classifiers are highlighted. Implementations from the WEKA API [6] were used with the exception of the GMM classifier which used an inhouse implementation. It is not possible to evaluate every possible classifier but the criteria decided upon for inclusion was reasonably broad ranging so as to make a useful comparison of different methods.

3.1. Gaussian mixture model (GMM)

The GMM is a parametric probabilistic classifier that models the distribution of multivariate input data using a mixture of \( K \) Gaussian distributions. GMMs have been shown to be effective at modelling the distribution of MFCC vectors in many applications such as speech recognition and synthesis and notably voice activity detection [3]. This makes them a good baseline for comparing performance against other ML classifiers.

During a training stage, feature vectors, \( \mathbf{x} \), are pooled according to their class, \( c \), to give class-specific vector pools, where for example, \( c \in \{ \text{nonspeech, voiced, unvoiced} \} \). Expectation maximisation (EM) clustering is applied to each vector pool to create a GMM for each class, \( \phi^c \) [3]. An unseen vector, \( \mathbf{x} \), is classified according to the GMM with highest probability. The discrete cosine transform (DCT) used in the MFCC feature extraction process removes correlation in the log filterbank domain. However, whilst this is true for static features, augmenting the feature vector with temporal derivatives does reintroduce correlation and so a full covariance matrix is retained to take into account these cross-correlations within the feature vector space.
3.2. Support vector machine (SVM)

SVMs are a non-probabilistic binary linear classifier [7]. SVMs require feature vectors, \( \mathbf{x} \), to be linearly separable based upon their class. Where feature vectors are not linearly separable a kernel function is selected that best maps the vectors into a new feature space where the vectors are linearly separable. The dividing hyperplane that provides the largest possible margin between the transformed data is then calculated and stored as a set of support vectors. New vectors are classified by calculating which side of the dividing hyperplane they fall. SVMs clearly rely on the appropriate selection of kernel function and this work considers the standard polynomial kernel in the WEKA API [6].

3.3. Multilayer perceptron (MLP)

MLPs are an extension of the linear perceptron and a form of artificial neural network [7]. They can be viewed as being related to SVMs, differing mainly in the method of class separation [8]. Like SVMs, MLPs aim to find the maximum margin between vectors based on the class. Instead of using a kernel function to transform the feature space, MLPs use multiple linear perceptrons to separate non-linearly separable vectors on their class in the existing feature space. Unseen vectors are classified by calculating the decision region in which they fall based on the arrangement and weighting of the linear perceptrons. From visual inspection, feature vectors from the datasets described in Section 4.1 were clearly found to be non-linearly separable. This suggests that MLPs should perform well due to their ability to form complex decision regions.

3.4. C4.5 decision trees

The C4.5 algorithm builds decision trees and can be considered a form of statistical classifier [7]. The decision tree is built by calculating the information gain that results from splitting a training data set on the class for each coefficient in the MFCC vector. The individual MFCC that gives the largest information gain is chosen as the split for the current node. The MFCC chosen to split at each subsequent node is determined in the same way until all vectors in the subset are labelled with the same class. Unseen vectors are classified by the decision rules determined by the split on each tree node (MFCC) until a class is determined. For VAD in clean conditions the most discriminative MFCC is likely to be \( C_0 \) due to large differences in energy between speech and non-speech. \( C_1 \) (spectral slope) is also likely to be effective in determining speech from non-speech.

3.5. Tree ensembles (Rotation Forest)

Ensemble classifiers are multiple classifier systems that use a number of models to obtain better performance. This work uses the Rotation Forest classifier which comprises a number of decision trees [9]. Each decision tree is trained on a random subset of training data with principle component analysis (PCA) applied to each subset. All principle components are retained to preserve the information in the variance of the data whilst decorrelating the feature space. Although static coefficients are already decorrelated by the DCT the correlation introduced into the feature vector by the temporal derivatives should be removed by PCA.

3.6. Naïve Bayesian networks

Bayesian networks are probabilistic classifiers that model the joint distribution of feature vectors with a set of conditional prob-

<p>| Table 1: Effect of MFCC feature type on voice activity detection accuracy at an SNR of 10dB in white noise |
|-----------------|-------|-------|-------|</p>
<table>
<thead>
<tr>
<th>MFCC</th>
<th>SVM</th>
<th>MLP</th>
<th>LogRes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C(1 - 12) )</td>
<td>0.81</td>
<td>0.88</td>
<td>0.79</td>
</tr>
<tr>
<td>( C(0 - 12) )</td>
<td>0.85</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>( C + \Delta C )</td>
<td>0.85</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>( C + \Delta C + \Delta C )</td>
<td>0.88</td>
<td>0.97</td>
<td>0.88</td>
</tr>
</tbody>
</table>

abilities using a directed acyclic graph [7]. A fully connected graph suggests that all coefficients in the vector are dependent on one another. An alternative to the fully connected Bayesian network is naïve Bayes which makes the assumption that each individual MFCC is dependent only on the class, vastly simplifying the complexity of the model. This assumption allows each MFCC to be modelled using an independent unimodal Gaussian distribution. The method is reasonably similar to the GMM classifier (Section 3.1) when \( K = 1 \) and assuming diagonal covariance. In informal testing, performance of naïve Bayes was found to be equivalent to that of a full Bayesian network and so the naïve Bayes classifier is used in this work.

3.7. Logistic regression

Logistic regression is a form of binomial regression analysis with no assumption made as to the distribution of the data [7]. The log outcomes of the class are modelled as a linear combination of the feature vectors, with the best fit calculated using maximum likelihood estimation. New vectors are classified by calculating the log odds of the vector belonging to a particular class using the regression model built in training.

4. Results

This section presents results and analysis of voice activity detection and voicing classification across the set of classifiers and also compares accuracy against conventional systems. Results are first presented to determine the optimal feature vector.

4.1. Dataset preparation

The speech used in the experiments was taken from the WSJ-CAM0 dataset [10] and downsampled to 8kHz. For testing in stationary noise, white noise was added, while for non-stationary noise, street noise from the NOIZEUS dataset was added [11]. The modified Intermediate Reference System (IRS) filter used in ITU-T P.862 was then applied to simulate the frequency response of a telephone handset before MFCC features were extracted.

Classifiers were trained on 20 male and 20 female talkers and tested on 5 male and 5 female talkers that were previously unseen. Ten utterances were selected per talker to give a total of 400 utterances (≈1200 sec) for training and a further 100 (≈300 sec) for testing. The test set comprised 36% silence, 22% unvoiced and 42% voiced speech. Reference VAD data was obtained using an energy threshold applied to noise-free speech. A pitch track was then calculated using PRAAT [12] and combined with the energy thresholding to give labels of non-speech, unvoiced and voiced. The test set was subsequently hand corrected where necessary.

4.2. Feature selection

To determine the optimal MFCC vector a preliminary voice activity detection test was performed. Results are presented in
Table 1 using speech that has been contaminated with white noise at an SNR of 10dB. For comparison three different classifiers are used – SVM, MLP and logistic regression.

Across the three different classifiers, results show that adding C0 and including velocity (∆C) and acceleration (∆ΔC) temporal derivatives all increase performance. The gain made by each addition varies across the classifiers but the overall result is an absolute increase in accuracy of between 7% and 9%. Therefore, for the remainder of experiments the MFCC vector comprises C0 to C12 with velocity and acceleration augmented.

4.3. Voice activity detection

This section examines classifier performance on voice activity detection. Each classifier is evaluated in receiver operating characteristic (ROC) space in terms of the true positive rate (speech detected correctly as speech) and false positive rate (non-speech detected as speech). Figure 1 shows performance of the classifiers in white noise and street noise at SNRs of 20dB, 10dB and 0dB. To provide baseline results, the performance of two industry standard VADs, namely the G.729 and ETSI XAFE, are also included [1, 2].

Comparing performance between the two noise types shows accuracy to be worse in the non-stationary street noise where sounds from cars, sirens, etc. introduce misclassifications. In terms of SNR, at higher SNRs classifier accuracies are reasonably close in the ROC space but as SNRs fall the classifier performances disperse with a shift towards the higher error region (bottom-right) of the ROC space. For both noise types the most accurate classifiers are Rotation Forest, GMM and the MLP and the worst performing are SVM and logistic regression. However, all of these classifiers outperform the two baseline VADs which are seen to perform poorly even in the relatively high SNR of 20dB. As SNRs fall their performance degrades rapidly, showing their sensitivity to noise.

The three best performing classifiers all share the ability to deal with feature vectors that contain coefficients that are correlated, which distinguishes them from the other classifiers. Even though the DCT employed in MFCC feature extraction should remove the correlation within the log filterbank, augmenting the vector with velocity and acceleration derivatives reintroduces some correlation. It is postulated that this may cause varying levels of difficulty for many of the classifiers tested, with the exception of Rotation Forest, GMM and MLP which gives rise to their superior performance. This suggests that applying a further transform to the feature vector, for example PCA, would decorrelate the coefficients of the entire feature vector and potentially improve the performance of other classifiers.

4.4. Voicing classification

Voicing classification extends VAD into the three class problem of determining between non-speech, unvoiced speech and voiced speech. For some classifiers only a binary decision is possible – for example SVM and logistic regression. In these cases two instances of the classifier were used, with the first being a VAD and the second applied to speech frames to classify between voiced and unvoiced speech, therefore allowing a three class output. The results of voicing classification in white noise are displayed in Table 2 and in street noise in Table 3. The tables show the classification accuracy for non-speech (NS), unvoiced (UV) and voiced (V) frames. A measure of the overall accuracy (OVL) is also shown and is computed from the total number of frames correctly classified. Results are presented at SNRs of 20dB, 10dB and 0dB. To serve as a baseline, the voicing classification accuracy from the ETSI XAFE standard is included [2].

The results follow a similar pattern to the VAD results, with voicing classification accuracy worse in the non-stationary street noise and reducing as SNRs fall. In terms of the accuracy of individual classifiers, as was observed for VAD, the Rotation Forest and MLP have highest overall classification performance. This is likely to be for the reasons discussed for VAD and related to the ability of these classifiers to deal with correlated data. Overall voicing classification accuracy for the ETSI XAFE baseline tends to be poor and gives lowest performance for the majority of test conditions.

Figure 1: Performance of voice activity detection in white noise and street noise at SNRs of 20dB, 10dB and 0dB in ROC space
Considering now the accuracy of identifying the individual voicing categories, unvoiced speech is clearly the most difficult to identify correctly. As SNRs fall, the machine learning methods rapidly become ineffective at identifying unvoiced speech as there are relatively few distinguishing features between noise and unvoiced speech. This leads to the majority of unvoiced frames being incorrectly classified as non-speech. Unvoiced classification is further affected by the 4kHz bandwidth of the speech and the application of the IRS filter to simulate the telephony channel. Both of these reduce high frequency energy which is an important cue for unvoiced speech. Conversely, the ETSI XAFE method is seen to retain a high score for unvoiced classification. However this is at the expense of correctly identifying non-speech frames and is explained by the increasing noise levels causing the ETSI XAFE method to classify non-speech as unvoiced speech.

Machine learning classification accuracy for voiced and non-speech frames, although deteriorating as SNRs reduce, is, however, more robust to noise than for unvoiced classification. This classification problem is more simple as voiced frames tend to be of higher energy than unvoiced frames and will therefore have a higher local SNR. Voiced frame energy is focused in lower frequency regions which are retained during feature extraction thereby providing useful discriminative information in the feature vectors.

5. Conclusions

Evaluations have shown that machine learning classifiers can be applied successfully to voice activity detection and voicing classification tasks. Their performance is highly robust to noise and significantly better than the two conventional methods tested. MLPs, Rotation Forest and GMMs were found to give consistently best performance and share the ability to deal with correlation in the feature vector. Accuracy of the other methods may be improved through decorrelation of the feature vector (such as with PCA) and this is the subject of further investigation.

### 6. References


