IDENTIFICATION OF VOICE PATHOLOGY USING AUTOMATED SPEECH ANALYSIS

C. Maguire¹, P. de Chazal¹, R.B. Reilly¹, P.D. Lacy²
¹Department of Electronic and Electrical Engineering, University College Dublin, Ireland
²Royal Victoria Eye and Ear Hospital, Dublin, Ireland

Abstract: The classification performance of an automatic classifier of voice pathology for the detection of normal and pathologic voice types is presented. The proposed classification system is non-intrusive and fully automated. Speech files of sustained phonation of the vowel sound /a/ in the ‘Disordered Voice Database Model 4337’ provided 631 subjects of both genders (58 normal, 573 pathologic). This database includes features extracted by the Multi Dimensional Voice Program (MDVP). Mel frequency cepstral coefficients (MFCC) were extracted for all of the speech files. Discrete Fourier transform (DFT) features, Log DFT and Cepstral features were also extracted. Cross-fold validation was used to measure the classifier performance. Linear discriminant analysis was employed as the classifier model. The MDVP feature set of shimmer and signal-to-noise ratios are shown to have similar classification performance to the Log DFT and the MFCC features.

Keywords: Voice Pathology, speech analysis, Linear Discriminant Analysis.

I. INTRODUCTION

A wide variety of vocal fold pathologies are found in patients with vocal disorders. These pathologies can be found in varying degrees of severity and development. They can be classed as physical, neuromuscular, traumatic and psychogenic and all directly affect the quality of the voice. At present a number of diagnostic tools are available to the otolaryngologists and speech pathologists such as videostroboscopy [1] and videokymography. However these current methods are time and personnel intensive and lack objectivity.

Research has been reported on the development of reliable and simple methods to aid in early detection, diagnosis, assessment and treatment of laryngeal disorders. This research has lead to the development of feature extraction from acoustic signals to aid diagnosis. Much focus has been centred on perturbation analysis measures such as jitter and shimmer and on signal-to-noise ratios of voiced speech, which reflect the internal functioning of the voice. Through this research it has been shown that these features can discriminate between normal and pathologic speakers [2], [3], [4], [5].

The aim of this research was to investigate the performance of a voice pathology classifier categorising sustained phonation of the vowel sound /a/ from a large labelled database into either a normal or pathologic class. The goal of this project was to produce a stand-alone classifier that would be non-intrusive and objective.

II. METHODOLOGY

Each stage of the flow chart of a voice pathology classifier in Figure 1 is discussed below.

Figure 1. Flow Chart of the Processes involved in a Voice Pathology Classifier

Acquisition: The labelled voice pathology database “Disordered Voice Database Model 4337” [6] acquired at the Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory and distributed by Kay Elemetrics was used in this study. A detailed description of the database can be found at [6], [7].

Digitised voice recordings of the sustained phonation of the vowel sound /a/ were used for training and testing the classifier. The database contains 631 recordings of which gender information is available for 389 recordings. In this study we divided the available data into three datasets in order to investigate the influence of gender on classification performance:

1. A mixed gender dataset containing 631 subjects (58 normal, 573 pathologic)
2. A male dataset containing 164 subjects (22 normal, 142 pathologic)
3. A female dataset comprising 225 subjects (36 normal, 189 pathologic)

Feature Extraction: The Multi Dimensional Voice Program (MDVP) [8] was used as a feature extractor. For each sustained phonation of the vowel sound /a/ in the database there are 33 associated MDVP features. These 33 features can be divided into six subsets. Each subset is a grouping of features that describe specific properties of the phonation. Namely: 1) the fundamental frequency, $F_0$, 2) jitter (short-term, cycle-to-cycle, perturbation in the fundamental frequency of the voice), 3) shimmer (short-term, cycle-to-cycle, perturbation in the amplitude of the voice), 4) Signal-to-noise ratios, S/N 5) count and 6)
duration features. Some recordings contained missing MDVP feature values. In these cases missing features were replaced by the average value of that MDVP feature. This ensured that the replaced features would not aid in the classification. The histogram was examined for each feature and where appropriate a log transformation was applied. This forced all the features to have an approximately Gaussian distribution.

The Mel Frequency Cepstral Coefficients (MFCC) features are commonly used in Automatic Speech Recognition (ASR) and also Automatic Speaker Recognition systems [9]. These coefficients are formed by taking the Discrete Fourier Transform (DFT) of the speech signal. Then a linearly spaced filterbank in the Mel frequency domain that translates to a log spaced filterbank in the Frequency domain is applied to the spectrum of the signal. The Mel scale is based on the non-linear human perception of sounds. Finally the signal is log transformed and the inverse discrete Fourier transform or the discrete cosine transform is applied. The MFCC were extracted from the speech signal using the Hidden Markov Model toolkit (HTK) that is commonly used in speech research [10]. A first order pre-emphasis filter using a coefficient of 0.97 was utilised here so that the measured spectrum has a similar dynamic range across the entire frequency band. The signal was then separated into 20ms frames using a Hamming window with an overlap of 10ms between each frame. HTK employs the DCT to transform the outputs of the filterbanks to the cepstral domain. MFCC were calculated for each frame and then averaged across all frames in a recording. Thus each speech recording is represented by the averaged MFCC for that particular speech recording. These averaged MFCC were used as features for the classifier.

The MFCC were divided into subsets in order to investigate the system performance using subsets of features. Different MFCC feature sets were extracted from the speech recordings with a varying number of filter channels and also a varying number of MFCC.

**Classifier:** Linear discriminants (LD) [11] partition the feature space into the different classes using a set of hyper-planes. The parameters of this classifier model were fitted to the available training data by using the method of maximum likelihood. Using this method the calculation required for training is achieved by direct calculation and is extremely fast relative to other classifier building techniques such as neural networks. This model assumes that the feature data has a Gaussian distribution for each class. In response to input features, linear discriminants provide a probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate.

**Estimating the classifier performance:** The cross-validation scheme [12] was used for estimating the classifier performance. The variance of the performance estimates was decreased by averaging results from multiple runs of cross validation where a different random split of the training data into folds is used for each run. In this study ten repetitions of ten-fold cross-validation were used to estimate classifier performance figures. For each run of cross fold validation the total normal population and a randomly selected group of abnormals equal in size to the normal population was utilised. This results in a more realistic reflection of the predictive ability of the system.

In this study the performance of the classifier is quoted using the class sensitivities, predictivities and the overall accuracy. The sensitivity of the classifier to a particular voice class is the fraction of speech files in the class that are correctly classified. The specificity is the sensitivity calculation applied to the normal class. The positive/negative predictivity is the fraction of speech files detected as abnormal/normal that are correctly classified. The overall accuracy is the fraction of the total number of subjects’ voices that are classified correctly.

**III. RESULTS**

All MDVP features were log-transformed so that the resulting histograms more closely approximated Gaussian distributions. Classification results were obtained for the MDVP, MFCC, DFT, Log DFT and Cepstral features as well as the combination of these features for mixed genders together and for each gender individually. The number of filterbank channels and coefficients used in the MFCC was examined. Through testing it was seen that utilisation of 15 filterbank channels and 15 coefficients resulted in satisfactory system performance.
The duration features of the MDVP were not included as intuitively there was no link between the duration of the recording and any pathology. The predictive ability of the count features was found to be poor and so this group was disregarded for the rest of the study. The classification performance of different feature sets is shown in Table 1. The feature set of shimmer and signal-to-noise ratios performs at a much lower level though all of the different gender classifiers, 75.97% and 78.14% respectively. The reduced set of MDVP features using frequencies between 0 and 385 Hz. Utilisation of the DFT magnitude and Log DFT features with all three gender classification systems achieve consistently high results of 81.18, 83.15, 79.42% and 81.53, 84.81, 81.79% respectively.

The Cepstral feature set did not perform as well as the MFCC feature set resulting in an accuracy of 77.19%, 80.11% and 71.95% for the mixed, male and female gender classifiers. This illustrates that by incorporating the human auditory system’s non-linear perception of the audio spectrum through application of the Mel scale improves the performance of the system.

Through the use of the first five MFCC it is possible to achieve the same classification rates as achieved using all 15 MFCC. This trend is consistent with research reported by [13] where the authors observed that only the first few MFCC were required for automatic speaker recognition systems. The test set accuracies for the system employing the MFCC perform well in the mixed gender and male gender classifiers, 82.65 and 88.40%, but the accuracy was lower for the female speech recordings, 73.95%. The MFCC are based on homomorphic analysis whose function is to deconvolute the speech signal, i.e. to separate the excitation and impulse response of a linear time-invariant system. The coefficients at the beginning of the MFCC and Cepstrum represent the impulse response of a linear system that combines the effects of the glottal wave shape, the vocal tract impulse response and the radiation impulse response [14]. For this reason these features should yield information about the health conditions.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Gender</th>
<th>Test set (%)</th>
<th>Gender</th>
<th>Test set (%)</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDVP (F0, Jitter, Shimmer, S/N)</td>
<td>Mixed</td>
<td>84.74</td>
<td>84.83</td>
<td>84.64</td>
<td>Male</td>
</tr>
<tr>
<td>MDVP (Shimmer, S/N)</td>
<td>Mixed</td>
<td>87.16</td>
<td>83.26</td>
<td>80.8</td>
<td>Male</td>
</tr>
<tr>
<td>MFCC (1:15)</td>
<td>Mixed</td>
<td>82.65</td>
<td>83.62</td>
<td>81.68</td>
<td>Male</td>
</tr>
<tr>
<td>MFCC (1:5)</td>
<td>Mixed</td>
<td>83.35</td>
<td>83.45</td>
<td>83.25</td>
<td>Male</td>
</tr>
<tr>
<td>DFT Magn (1:8)</td>
<td>Mixed</td>
<td>81.18</td>
<td>87.59</td>
<td>74.69</td>
<td>Male</td>
</tr>
<tr>
<td>Log DFT (1:8)</td>
<td>Mixed</td>
<td>81.53</td>
<td>83.1</td>
<td>79.93</td>
<td>Male</td>
</tr>
<tr>
<td>Cepstrum (1:8)</td>
<td>Mixed</td>
<td>77.19</td>
<td>74.66</td>
<td>79.76</td>
<td>Male</td>
</tr>
<tr>
<td>MDVP (F0, Jitter, Shimmer, S/N) &amp; MFCC (1:15)</td>
<td>Mixed</td>
<td>85.69</td>
<td>86.03</td>
<td>85.34</td>
<td>Male</td>
</tr>
<tr>
<td>Log DFT (1:8) &amp; MDVP (Shimmer, S/N)</td>
<td>Mixed</td>
<td>88.55</td>
<td>88.28</td>
<td>88.83</td>
<td>Male</td>
</tr>
<tr>
<td>Log DFT (1:8) &amp; MFCC (1:5)</td>
<td>Mixed</td>
<td>85.86</td>
<td>86.55</td>
<td>85.17</td>
<td>Male</td>
</tr>
<tr>
<td>DFT Magn (1:8) &amp; MFCC (1:5)</td>
<td>Mixed</td>
<td>84.82</td>
<td>87.41</td>
<td>82.2</td>
<td>Male</td>
</tr>
<tr>
<td>Cepstrum (1:8) &amp; MFCC (1:5)</td>
<td>Mixed</td>
<td>82.57</td>
<td>82.59</td>
<td>82.55</td>
<td>Male</td>
</tr>
</tbody>
</table>

The reason why only the first eight coefficients are significant for the DFT, Log DFT and cepstral coefficients is due to the fact that it is a vowel sound /a/ that is being analysed and hence most of the fundamental frequency content will be contained in the lower frequencies. Utilisation of the DFT magnitude and Log DFT features with all three gender classification systems achieve consistently high results of 81.18, 83.15, 79.42% and 81.53, 84.81, 81.79% respectively.

IV. DISCUSSION

The MDVP feature set performs well for the mixed gender classifier achieving a classification accuracy of 84.74%. However, its performance falls off when utilised in the individual gender classifiers, 75.97% and 78.14% respectively. The reduced set of MDVP features using shimmer and signal-to-noise ratios performs at a much more consistent level though all of the different gender classifiers with an accuracy of 87.10%, 90.61% and 75.97% respectively.
Various combinations of the feature sets were examined however we observed that the systems performance was not improved significantly.

A number of research groups [15], [16], [17] have reported results for detection rates for voice pathologies of 94.87%, 76% and 96.30% respectively. In [15] the Disordered voice database was employed and their results may be compared with the results obtained in this study. However results from [15] should be considered biased as the authors used the MDVP speech recording duration features “SEG” and “PER”. In the database the normal recordings are three times longer in duration than the pathologic recordings and therefore the “SEG” and “PER” features are three times as large for normal recordings than for pathologic recordings. Hence the features based on the recording duration could be used to distinguish the normals from pathologica cases with high success due to the different durations of normal and pathologic recordings.

In study [16] different databases were used and a direct comparison of results cannot be made. The database used in the present study provides a large amount of pathologic subjects that might not fairly represent the pathologies present in other studies conducted in this are or those encountered by the medical profession on a day to day basis. The predictice ability of this model could be confirmed through external validity. The latter study [17] utilises similar features to the ones used in this study however their classification performances were based on correct classification of individual frames from the speech files which implies that the training data used would consist of data very similar to the testing data.

V. CONCLUSION

The MDVP feature set containing the shimmer and signa-to-noise features offers the best classification results over each of the gender classifiers. The utilisation of the Log DFT and MFCC feature set in the classification system performs almost as well as the MDVP features. However the Log DFT and MFCC features are implemented with very little computational cost in comparison to the MDVP features.

In this study, the performance of the mixed-gender classifiers was similar to the classification performance of the single-gender classifiers. These results suggest that for this particular automatic classification system there is no advantage to be gained by utilising single-gender classifiers to detect pathologic voice.

The Support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

REFERENCES