THE EFFECTS OF INTER AND INTRA SPEAKER VARIABILITY ON PATHOLOGICAL VOICE QUALITY ASSESSMENT

Juan I. Godino-Llorente\textsuperscript{1}, Tim Ritchings\textsuperscript{2}, Carl Berry\textsuperscript{2}

\textsuperscript{1} Dept. Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Campus Universitario, N II Km 33.6, 28871 Madrid, Spain.
\textsuperscript{2} School of Computing, Science and Engineering, University of Salford, The Crescent, Salford, M5 4WT, UK

Abstract: This paper describes some methodological issues to be considered while facing the task of the objective assessment of voice quality from patients with laryngeal cancer. Earlier research works showed that the automatic assessment of voice quality could be addressed by means of short-term and long-term time-domain, and frequency-domain parameters extracted from electroglotographic (EGG) signals, and using Artificial Neural Networks (ANN) such as Multi-layer Perceptron (MLP) \cite{1}. However, despite the good results, further research has showed that the choice of cross-validation techniques used for the pattern recognition can greatly influence the ability of the system to learn and to generalise. In particular, this paper is concerned with the effects of intra and inter speaker variability during cross-validation and hence on the reliability of pathological voice quality assessment. For this study, a database of male subjects steadily phonating the vowel /i/ was used, and the quality of their voices was independently assessed by a speech and language therapist (SALT) according to their 7-point ranking of subjective voice quality. Although it is found that by carefully selecting the datasets used to train and validate the ANN to minimise intra speaker variability reduces the classification accuracy, most of the time the ANN only misclassifies by only one point.

Keywords: Voice quality, classification, neural networks, cross-validation.

I. INTRODUCTION

The effectiveness and importance of the acoustic and EGG analysis of pathological voices have been proven by many experimental researches. The starting point of this work is that carried out in \cite{1}. This work proposed an Artificial Neural Network (ANN) based framework to evaluate the voice quality into a 7 point scale using short term and long term parameters extracted from the EGG signal with an accuracy over 90\%. Such work used 50 speakers whose EGG signal were recorded before the treatment. However, due to the limited number of patients, the training and validation datasets used to develop the ANN used multiple frames taken from the signals recorded for some of the patients. As a result the ANN learnt both the intra and inter speaker variations in the data. This could lead to artificially high classifications with small datasets, with the system effectively recognising a speaker in the dataset, rather than assessing voice quality from the parameters derived from signal recorded from different speakers.

This study has reconsidered the training and validation of ANNs used for voice quality assessment in the light of these intra and inter speaker variations.

II. CROSS VALIDATION

Pattern recognition techniques by themselves do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data should be removed before training begins. Then when training is done, the data that was removed could be used to test the performance of the model on “new” data. This is the basic idea for cross validation.

The holdout method is the simplest kind of cross validation. The data set is separated into two sets: the training and the validation set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the unseen data in the validation set. The advantage of this method is that it takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the validation set, and thus the evaluation may be
significantly different depending on how the division is made.

**K-fold cross validation** improves the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the validation set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a validation and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

**Leave-one-out** cross validation is to split the P patterns into a training set of size (P-1) and a validation set of size 1 and average the squared error on the left-out pattern over the P possible ways of obtaining such a partition. This is called leave-one-out (LOO) cross-validation. The advantage is that all the data can be used for training -none have to be held back in a separate validation set. The evaluation given by leave-one-out cross validation error is good, but it is computationally expensive.

For this work the k-fold cross validation method has been used, because the leave-one-out was considered very time consuming.

### III. DATA PROCESSING

The procedures used in extracting the parameters in this paper are broadly similar to those used in [1] which contains more detail than we will look at here, the main changes are due to the nature of the two different systems. In the earlier study there was a large amount of manual judgement and adjustment at various stages to obtain the best set of extracted parameters, because the long term aim for this work is to be used in a system used by non-technical users it was necessary to fully automate the extraction procedures, thus losing some accuracy in the process. The signal was first stationarised to remove drift, split into 50 ms. frames (Hanning windows overlapped by 25 ms.) then the autocorrelation was taken to remove some of the noise components. In the next stage silent frames were removed by comparing zero point crossing and short term amplitudes with that of a sample of silence (recorded at the same time as the data samples). Following on from this voiced and unvoiced frames were separated using a cepstral based approach described in [2], here we are looking for a pronounced peak indicating the presence of a fundamental frequency. This was originally done by a user but in the current work we attempt to detect this peak with a fairly simple peak finding algorithm, in any such attempt a trade off has to be made and after much experimentation the system errrs on the side of rejecting frames as to be sure that all passed frames do actually contain speech.

After the Power Spectrum Density (PSD) is calculated for each frame the frames are pooled to create the Fundamental Harmonic Normalisation (FHN) from which we can extract some of the long term parameters, again some user interaction was previously required here but this has been replaced with a peak finding algorithm. Once both the PSD and FHN have been calculated they are both fitted with Gaussian Mixture Models (GMM) in order to reduce the number of parameters needed to describe the signals. This proved difficult to automate, especially with the more damaged voices, and the algorithm still has a tendency to try to fit Gaussians to peaks that prove not to be harmonics. Once the GMMS are fitted the parameters are extracted as in the previous work [1]. The parameters extracted are 15 short term parameters of the mean, standard deviation and amplitude of the Gaussians fitted to the fundamental frequency and the first 4 harmonics (if they exist) and 4 other short term parameters, those of the value of the fundamental frequency in each frame (Fo), the noise threshold (No), the FHN Noise Energy (FHNNE) -based on the Normalised Noise Energy (NNE) [4], but derived from the FHN spectrum- and the Residual Harmonic Error (RHE). Along with these are extracted 3 long term parameters, those of the mean fundamental frequency for a sample (MFO), a measure of the jitter of the fundamental frequency between frame (JFO) and the percentage of voiced versus unvoiced frames (V+). The data extracted from the speech data and used for the ANN classification tests comprised of 3 long-term parameters (MFO, JFO, V+) and 17 short-term parameters. Full details of the data processing and extraction of these parameters can be found in [3].

### IV. THE CLASSIFIER

A widely used architecture has been used for this purpose: a three layered feedforward perceptron (MLP). The Learning algorithm used is backpropagation with adaptive momentum [4]. The training was carried out along 400 epochs. The activation function used at each node is sigmoidal, and the number of neurons of the input layer is 20, the same number as the parameters extracted. This input data was normalised to between [-1,1] before being supplied to the net. The output layer has 7 neurons that are activated for every single class.
V. DATABASE USED

The data used to develop the system was captured off-line under clinical conditions at the Christie and Withington Hospitals in Manchester, using an Electrolaryngograph PCLX system [6]. This system is used to capture electrical impedance (EGG) signals using pads placed either side of the neck synchronously with acoustic signals captured using a microphone. Both EGG and acoustic data channels were captured synchronously at 20 kHz for up to 3 seconds while the subject phonated the vowel /i/ as steadily as possible. Although speech data was recorded for both male and female patients, the largest pathological group was male, so it is these speech signals that were used in the study. For each patient the SALT made a subjective voice quality assessment using a 7-point ranking.

![Distribution of the voice disorders database](image)

Fig. 1: Distribution of the voice disorders database

The database contains about 50 males voices recorded just before treatment, six months after treatment and one year after treatment (150 files), showing in general an improvement of the voice quality. Fig 1 represents the number of files for every class.

In the earlier study only the pre-treatment registers were used and the training and validation datasets were developed extracting the 30% of the frames for validation and 70% for training. So both datasets contained information about all the speakers stored in the database.

In this research, the pre and post-treatment registers were used. This approach ensure that the same speaker belongs to different categories, depending on the stage of the treatment. It enforces the ANN to learn the speaker independent features, and so minimise the effects of the intra-speaker variations.

The results have been obtained cross validating using the k-fold cross method. It is less time-consuming than the leave-one-out, but provides a good idea of the ability of the system to classify according to our criteria. 25 different datasets have been developed for every MLP size. The training datasets contain frames that belong to 7 speakers, whereas the validation dataset contains frames belonging to 3 different speakers. The pre-treatment and post-treatment recordings were mixed together in order to ensure that the system is not able to recognise speaker dependent features. This approach ensures that the ANN is forced to classify according the quality of the voice, keeping aside the features inherent to the speaker, due to the fact that the same speaker belongs to different categories depending on the stage of the treatment.

VI. RESULTS

Fig 2 shows the results obtained. In the left column are shown the frame and file (the whole recording) accuracy of the system using the EGG signal parameterised as explained above. The right column shows the results using the same parameterisation approach over the glottogram extracted from the acoustic data by means of pitch asynchronous inverse filtering techniques [7]. The file accuracy has been obtained by aggregating the assessments for every frame in the file. The results have no significant variation on the MLP size, showing a better behaviour with EEG signal than with the glottogram extracted by inverse filtering.

As was expected results are worse than in the earlier study, but it can be seen that when the ANN misclassifies a speaker it generally does so by only one point on the SALT’s 7 point scale.

VII. CONCLUSIONS

The modest scores (~40%) could either be due to the ability to discriminate the features extracted, or due to the MLP not being able to separate the prototypes correctly. However it is shown that most of the time, the classifier misclassifies by only one class. When interpreting these scores it has to be kept in mind that the SALT classifications were made by perceptual evaluation, and sometimes the experts do not agree on the evaluation of a voice sample. It is well known that there is an intra and inter-evaluator judgement variability, due to the fact that it depends a lot on their own expertise and subjective criteria about how a normal voice should be.

Despite of the modest scores, this system is able to provide an objective approach to the assessment of voice quality. For the future work, it should be tested with a larger database to improve the accuracy of the system, and it has to be tested using on-line with a Clinic.

VIII. ACKNOWLEDGEMENTS

This work was supported under the grants: EPSRC GR/L51546 (UK), and PR2002-0239 from the Secretaría de Estado de Educación y Universidades (Spain).
REFERENCES


Fig. 2: Results using EGG signal (left column) and glotogram waveform (right column) extracted from acoustic data by means of inverse filtering. It is represented the frame and file accuracy. The performance matrix shows the number of files that have been classified into each class.