ELSEVIER

Medical Engineering & Physics 24 (2002) 561–564

Medical Engineering & Physics

www.elsevier.com/locate/medengphy

Technical note

Pathological voice quality assessment using artificial neural networks

R.T. Ritchings [a,*], M. McGillion [a], C.J. Moore [b]

[a] *Department of Computer Science, University of Salford, Salford, UK*
[b] *North Western Medical Physics, Christie Hospital NHS Trust, Manchester, UK*

Received 28 May 2002; accepted 28 May 2002

Abstract

This paper describes a prototype system for the objective assessment of voice quality in patients recovering from various stages of laryngeal cancer. A large database of male subjects steadily phonating the vowel /i/ was used in the study, and the quality of their voices was independently assessed by a speech and language therapist (SALT) according to their seven-point ranking of subjective voice quality. The system extracts salient short-term and long-term time-domain and frequency-domain parameters from impedance (EGG) signals and these are used to train and test an artificial neural network (ANN). Multi-layer perceptron (MLP) ANNs were investigated using various combinations of these parameters, and the best results were obtained using a combination of short-term and long-term parameters, for which an accuracy of 92% was achieved. It is envisaged that this system could be used as an assessment tool, providing a valuable aid to the SALT during clinical evaluation of voice quality.

Keywords: Intelligent systems; Neural networks; Electroarynograph signals; Voice quality; Larynx cancer

1. Introduction

An increasingly important factor in prescribing treatment for cancer of the larynx is the quality of voice retained post-therapy. At present, speech and language therapists (SALTs) endeavour to rehabilitate a patient's voice back to normality, or as near normal as possible, quickly following treatment. They currently assess voice quality on a seven-point ranking (0=least abnormal, 6=most abnormal) based on a variety of sound parameters, some of which are well defined, such as shimmer and jitter, while others, such as whisper and creak, are descriptive or have tenuous physical correlates. As a result, the assessment is largely subjective and depends upon the experience of the SALT.

This situation will be clearly improved by the availability of an objective voice-quality system, which can provide accurate, reproducible, graded measures of a patient's voice quality to help the SALT plan the patient's rehabilitation.

Earlier work has shown that a multi-layer perceptron (MLP) trained using features derived from a normalised power spectral representation, the fundamental-harmonic normalised spectrum (FHN) [1], of stationary vowel segments can classify EGG speech signals as normal or abnormal with an accuracy of 80% [2].

Whilst this provided good classification between normal and abnormal voice quality, the feature set was limited to sub-optimal classification results, as it is well known that some pathologies are measured more easily using long-term ($>$50 ms) parameters [3]. This paper describes the refinement of the artificial neural network (ANN) approach to voice quality assessment, by introducing long-term features to the prototype classification system.

In addition, the extension of the system to provide a sub-classification of abnormal voices in line with the SALT seven-point ranking scheme is investigated, and preliminary results are presented.

* Corresponding author.
E-mail address: t.ritchings@salford.ac.uk (R.T. Ritchings).

1350-4533/02/$22.00
PII: S1350-4533(02)00064-4

2. Data capture

The data used to develop the system were captured off-line under clinical conditions at the Christie and Withington Hospitals in Manchester, using an Electrolaryngograph PCLX system [4]. 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 s while the subject phonated the vowel /i/ as steadily as possible. Other vowel sounds were also recorded, however, the /i/ vowel is most frequently used by SALTs as the onset of phonation occurs more quickly than in other vowels and is a good indication of vocal fold health.

The EGG signal provides information about vocal fold contact behaviour during voice production, as the electrical impedance varies with the opening and closing of the glottis. This signal is modulated by the resonant cavities of the vocal, oral and nasal tracts to provide the acoustic signal. As the EGG signal is much less complex than the acoustic signal, the visual appearance of this signal is used by the SALT in conjunction with their perception of the acoustic sound when making their assessment of voice quality.

Although speech data were 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. Implicit in this is that the SALT had made a voice quality assessment of the patient using their own seven-point ranking.

3. Data processing

An automated voicing analysis was performed upon each 3 s EGG and acoustic speech signals to determine if the subject had voiced during phonation. If voicing was considered to have occurred, the EGG signal was processed to extract the long-term features initially, and then the short-term features for classification of voice quality. The long-term features include Mf_0, the mean of fundamental frequency, f_0, the standard deviation of f_0 (SDf_0), and the percentage of the 3 s signal that is voiced (V+), while the short-term features include parameters related to the spectral envelope of the first few glottal harmonics, and the glottal noise.

The voicing test involved taking 50 ms frames from the signals and applying Cepstral analysis techniques [5,6] to identify the voiced frames. Each frame was then pre-emphasised by forward differencing to reduce the effects of drifting signal amplitude, and its autocovariance was multiplied by a Hanning window, prior to transformation to the frequency domain using the fast Fourier transform. [7] An estimate of f_0 for each frame, deduced from the voicing analysis, was used to derive the FHN normalised spectral representation. This process removed the large observed inter-patient variability in f_0 and its harmonics, thus allowing a more effective modelling of the spectral envelope among groups of patients. Once the FHN spectrum had been determined, Gaussians were fitted to the data around f_0 and its first few harmonics [8]. Each Gaussian, $G_{h,}$ (h=0 up to typically 8) was parameterised as:

$$G_h = (\text{position}_h, \text{width}_h \text{ and amplitude}_h)$$

An observation was made that the mixture of Gaussians gave a better 'fit' to the FHN spectrum for the less abnormal patients, and so a parameter related to the goodness of fit, called the harmonic linearity measure (HLM), was calculated for each frame. Finally, as glottal noise is considered to be an important measure of voice quality, a parameter based on the normalised noise energy (NNE) [9,10], but derived from the FHN spectrum, FHNNE, was calculated for the data.

The parameters extracted from the speech data and used for the ANN classification tests comprised three long-term parameters (Mf_0, SDf_0, V+) and 18 short-term parameters (f_0, G_1, G_2, G_3, G_4, G_5, HLM, FHNNE). Full details of the data processing and extraction of these parameters can be found in McGillion [11].

4. Data classification

In total, 77 abnormal speech signals were available for training and testing data. For each of the seven classes, 450 patterns were used for training/validation and 200 for testing. Unfortunately, as a result of the relatively small dataset, there were different numbers of patients in each class. As it is desirable to have equal numbers in each class to train an ANN adequately, additional frames were taken from some patients in classes with the fewest patients and a small percentage of extra patterns was produced by adding normally distributed noise to the short-term features that were derived from these frames.

A two-layer, seven-output MLP, as shown schematically in Fig. 1, was trained using the back-propagation training algorithm, softmax activation function, and cross-entropy error function [12]. The advantage of using the cross-entropy activation function is that the output across all seven classes sums to 1.0 and can therefore be interpreted as a probability of membership of each of the seven classes. A further constraint placed upon the MLP is that for any single class to be declared the 'winner' the output for that class must be greater than 50% (0.5). MLP structures with different numbers of hidden nodes and subsets of the 21 input parameters were investigated in order to determine the combination that provided the minimum classification error.

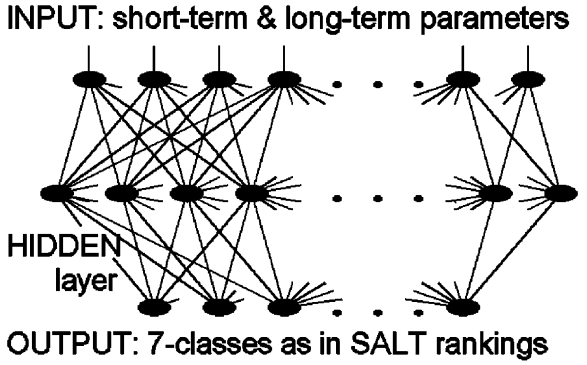

Fig. 1. Generalised MLP structure used in this study.

5. Results and discussion

Several different combinations and subsets of the 21 parameters were investigated. All the short-term features were found to contribute to the classification. The classification accuracy increased from 26.5% with [G1] alone to 67.7% with [G1, G2, G3, G4, G5]. Adding the other short-term features [FHNNE] and [HLM] increased the discrimination ability of the MLP to 72.07 and 68.64%, respectively.

Similarly, the long-term features were also found to be very important to the discrimination between the classes. These parameters [Mf_0], [SDf_0], [V+] alone were able to distinguish between the classes with accuracies of 37.57, 23.07, and 26.78%, respectively. However, it was found that it was the combination of the short-term and long-term features that provides the most accurate classifications of the abnormal signals.

The best overall ANN structure was a 20-40-7 MLP using the parameters [G_1, G_2, G_3, G_4, G_5, FHNNE, HLM, Mf_0, SDf_0, V+], and the results indicate that this MLP was able to distinguish between the seven abnormal groups with an accuracy of up to 92%.

Fig. 2 shows the output of the MLP for SALT's 20 patients pre-classified as class 3 abnormals. The output of the MLP is an estimate of the posterior probability of the membership of each class C_i ($0 \leq i \leq 6$). It should be noted that class estimates were also produced for the other SALT classes, and only one misclassification was found. This was for signal CA87EE, as seen in Fig. 2, where the highest class probability was for class 6. However, when the output probability was transformed to take into account the prior probability of each class [13], this signal was correctly assigned to group 3. Perhaps unsurprisingly, the classes at the two extremes of the scale, 0 and 6, provide the best classification results. In all cases, classes 3, 4, and 5 are the most difficult to discriminate between.

6. Conclusions

The results from this work suggest that a voice quality assessment system incorporating an ANN can be trained to provide objective sub-classifications of voice quality in line with the seven-point ranking scheme used by the SALT.

However, it should be noted that the ANN has been trained on the assessments of one SALT, which could lead to subjectively biased results. The collection of patient speech data, including voice quality rankings from several SALTs in the region, is now taking place, and will hopefully provide a larger and less biased dataset for training the system.

At the same time, work is taking place to identify and evaluate other parameters that can be derived from the speech data, in particular the acoustic data, which have been largely ignored in this study so far, in order to further improve the accuracy and reproducibility of these experimental results.

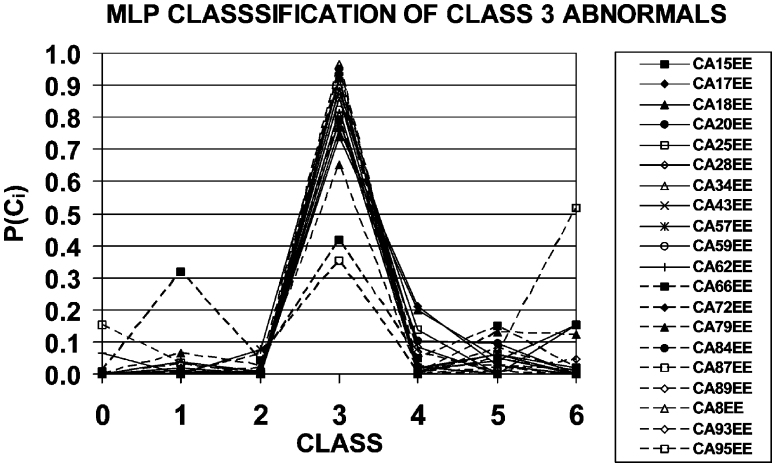

Fig. 2. The MLP estimate of class probability for the SALT's pre-classified class 3 abnormals.

Acknowledgements

The support of this work by the EPSRC award GR/L51546 is greatly appreciated.

References

[1] Moore CJ, Slevin N, Winstanley S. Characterising vowel phonation by fundamental spectral normalisation of LX-waveforms. Proceeding of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, vol. 1; 1999. p. 1–6.
[2] Ritchings RT, McGillion M, Conroy G, Moore C. Objective assessment of pathological voice quality. Proc IEEE SMC99, vol. 2; 1999. p. 340–5, ISBN: 0-7803-5683-7.
[3] Baken RJ. Electroglottography. J Voice 1992;6(2):98–100.
[4] Fourcin AJ, Abberton E, Miller D, Howell D. Laryngograph: speech pattern element tools for therapy, training and assessment. Eur J Disord Commun 1995; 30(2):101–15.
[5] O'Shaughnessy D. Speech communications: human and machine. IEEE Press, 2000.
[6] Noll A. Cepstrum pitch determination. J Acoust Soc Am 1967;41:293–309.
[7] Chatfield C. The analysis of time-series: an introduction. London: Chapman & Hall, 1996.
[8] Zolfaghari P, Robinson T. Formant analysis using mixtures of Gaussians. Proceedings of the International Conference on Speech and Language Processing, ICSLP'96, vol. 2; 1996. p. 1229–33
[9] Michaelis D, Gramss T, Strube HW. Glottal-to-noise excitation ratio—a new measure for describing pathological voices. ACOUSTICA—Acta Acustica 1997;83:700–6.
[10] Kasuya H, Ogawa S, Mashima K, Ebihara S. Normalised noise energy as an acoustic measure to evaluate pathologic voice. J Acoust Soc Am 1986;80(5):1329–34.
[11] McGillion. Automated analysis of voice quality. PhD Thesis, UMIST; 2000.
[12] Bishop CM. Neural networks for pattern recognition. Oxford University Press, 2000.
[13] Tarassenko L. A guide to neural computing applications. London: Arnold, 1998.