A COMPARATIVE STUDY OF INTELLIGENT VOICE QUALITY ASSESSMENT USING IMPEDANCE AND ACOUSTIC SIGNALS

Carl Berry & Tim Ritchings

School of Computing, Science and Engineering,
University of Salford, UK
c.berry2@salford.ac.uk

Abstract: **Objective assessment techniques for classifying voice quality for patients recovering from treatment for cancer of the larynx should lead to more effective recovery than the present approach, which is very subjective and depends heavily on the experience of the individual Speech and Language Therapist (SALT). This work follows an earlier study where an Artificial Neural Network (ANN) was trained on parameters derived from electrolaryngograph electrical impedance (EGG) signals recorded while a patient was phonating /i/ as steadily as possible, and gave an indication of voice quality inline with the standard UK Speech and Language Therapist (SALT) seven point scale. The applicability of this approach to voice quality assessment of acoustic signals is described, and the results are found to compare very well with those derived from the impedance signals. It was also noted that for both the impedance and the acoustic signals, the ANNs were able to classify the very good (recovered) and the very poor (abnormal) voices well, but performed quite badly with the mid-range classifications, raising questions about the accuracy of these classifications.**

Keywords: Voice quality, classification, Artificial Neural Network, acoustic, impedance.

I. INTRODUCTION

In the UK, voice quality assessment for patients recovering from treatment for cancer of the larynx is undertaken by Speech and Language Therapists (SALT), who use a standard 7-point classification scale ranging from Lx0-Lx6, with Lx0 being a near normal (recovered) voice while Lx6 represents an abnormal, very poor quality voice. The approach taken to reach a classification is very subjective and depends to a large extent on the experience of the SALT.

This work is concerned with a series of investigations aimed at producing an intelligent computer-based system which can provide objective classifications of voice quality in patients recovering from cancer of the larynx patients in line with the UK standard 7-point classification scale.

Previous work [1,2] has demonstrated that accurate classifications could be obtained from a Multi Layer Percepton (MLP) Artificial Neural Network (ANN) which was trained on a combination short-term and long-term parameters derived from electrolaryngograph electrical impedance (EGG) signals while a patient was phonating /i/ as steadily as possible. Although, acoustic signals were recorded at the same time as the impedance signals, they were not analysed as they appeared much noisier than the EGG signals. However, classification of voice quality from the acoustic signals is advantageous, if possible, as highly specialised and expensive equipment (the electrolaryngograph) will not be necessary, and this raises the possibility of screening in a GP's practice, rather than in the secondary care centres.

A preliminary assessment of the acoustic signals is described here, and the resulting classifications that have been achieved with the ANN approach are compared with those obtained for the impedance signals.

II. TREATMENT OF VOICE SIGNALS

A. *Collection of Voice Signals*

The patient's voice data was collected by the Christie Hospital and the South Manchester hospital using an electrolaryngograph PCLX system [3]. The equipment simultaneously records the electrical impedance signal via pads placed at specific positions on the patient's neck at the same time as the acoustic voice signal using a microphone. In these studies, the patient was attempts to steadily phonate the /i/ sound. This process means that two datasets are collected, one showing the EGG and a second showing acoustic variation, allowing for a direct comparison between the two sets. In the work only the male voices were used as the number of female voices in the dataset was too small to give an accurate assessment, a feature of the dataset is that most cancer of the larynx patients are male. Voice quality was subjectively classified by a SALT for each patient using their 7-point scale. The number of patients in each of the 7 categories is shown in Table 1.

Lx0	Lx1	Lx2	Lx3	Lx4	Lx5	Lx5
22	36	25	33	26	25	11

Table. 1. Patients in each SALT category

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online)

B. Signal Pre-processing

In order to be able to extract the short and long term parameters used in the classification process, a number of pre-processing stages were applied to the impedance and acoustic datasets. Initially the signals were stationarised to remove drift, split into 50 ms frames (Hanning windows overlapping by 25 ms) and then converted to the autocorrelation form of the signal to remove some of the noise components. Once these processes were complete, the frames were examined to check if they contained silence or sound. This involved comparing the frames with a sample of silence frame recorded under the same conditions, and used zero point crossing and short term amplitudes as checks. Once the silence frames have been removed, the remaining frames were separated into voiced and unvoiced frames; voiced frames containing vocal phonation while unvoiced containing no recognisable speech. This was achieved using the cepstrum based approach as described in [4]. The Fundamental Harmonic Normalisation (FHN) as described in [5] was then calculated from Power Spectrum Density (PSD) and then this structure (typical examples are shown in Fig 2) was modelled by fitting a Gaussian Mixture Model (GMM) in order to reduce the number of parameters needed to describe the signal.

C. Parameter Extraction

A total of 22 short and long term parameters are extracted for use with classification, as detailed in [1,2]. The short term parameters consist of 15 parameters relating to the mean, standard deviation and peak of the gaussians used to describe the fundamental frequency and first four harmonics in the frame (if they can be detected); the value of the fundamental frequency in each frame (F_0), the noise threshold value (N_0), the FHN Noise Energy

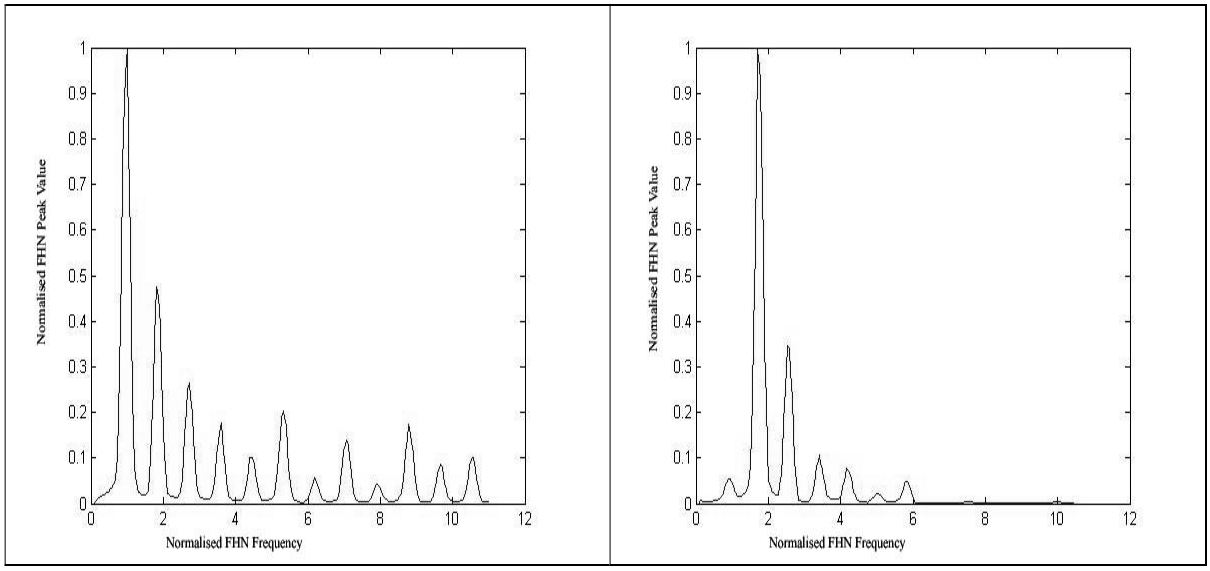

Figure 2a FHN plot of good quality impedance signal

Figure 2b FHN plot of good quality acoustic signal

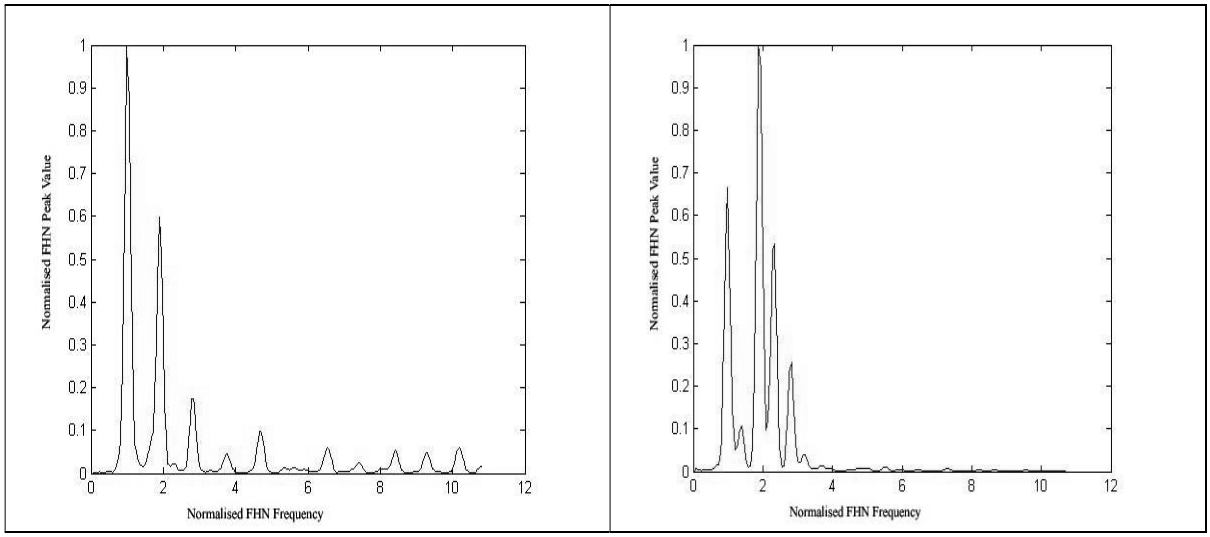

Figure 2c FHN plot of poor quality impedance signal

Figure 2d FHN plot of poor quality acoustic signal

(FHNNE) and the Residual Harmonic Energy (RHE). The 3 long-term parameters were extracted from the speaker's whole voiced speech. These included the mean fundamental frequency across all frames (MF_0), a measure of jitter of the fundamental frequency between frames (J_0) and the ratio of voiced to unvoiced frames.

D. The classification technique

Once the parameters have been extracted, a 3 layer feed-forward ANN with a sigmoidal activation function in the hidden layer, and using backpropagation of errors, is used for classification purposes. The ANN had 22 inputs, one for each of the short-term and long-term parameters derived from the voice signals, and 7 outputs, corresponding to the SALT categories. The "leave one out" cross validation strategy was used as it generally regarded as one of the most accurate methods and by leaving out a single patient's voice sample we can ensure to avoid inter versus intra speaker effects [2].

3. COMPARISON OF IMPEDANCE AND ACOUSTIC SIGNALS

A. Observed differences between the signal types.

When the FHN spectra of the impedance and acoustic signals were examined visually, a number of differences can be observed. Figs. 2a and 2b show the spectra derived for of a good quality pathological voice (Lx0) for the impedance and acoustic signals respectively. A clear difference it that for the impedance signal, the largest peak belongs to the fundamental frequency, whilst in the acoustic signal it is the 1st harmonic. This is normal and corresponds to the pattern that would be expected from a human voice. However, even in this good quality voice the acoustic signal only shows six harmonics, as compared for eleven for the impedance signal. It should also be noted that the peak of the fundamental frequency is very small in the acoustic signal, making it difficult to detect with the techniques used for the impedance signals.

This figure also shows typical FHN impedance (Fig 2c) and acoustic (Fig 2d) spectra for a poor quality voice (Lx5). Again, the impedance signal shows many more harmonic structures. This reduction of harmonics also has an impact on the number of parameters that are available for use with the classification algorithms, and in the case of fig 2b, for example, it would only be possible to extract short term parameters for the fundamental frequency and the first 2 harmonics meaning that the model would be missing 6 short term parameters relating to the 4th and 5th harmonics. In some cases the situation is even worse, and in the extremely bad voices or very poor frames, it is not unusual to find the fundamental frequency and a single harmonic as the only recognisable structures.

Finally, it may be seen from Fig 2d that the acoustic signal suffers from far more noise between the harmonics than the impedance signal, making it much more difficult to fit the Gaussian Mixture Models. This also causes the centres of the harmonic structures to be shifted away from their correct positions in the FHN.

B. Classification differences between the signal types.

Classifications were made for both the impedance and acoustic signals, using the same parameters and training and verification procedures. In the cases where harmonics were not found, these parameters were set to zero. The resulting classifications are shown in Table 2.

As the acoustic signals are generally noisier than the impedance signals, with a poorer quality output, leading to typically fewer parameters, it was expected that the acoustic classifications would be less accurate that those obtained for the impedance signals. Surprisingly, this turns out to not be the case, and it can be seen in the Table that there is very little difference between the final classifications achieved for the two types of voice signal.

It should also be noted that for both types of signal, the ANNs give the best classifications for the worst voices (Lx5-6), obtained good results for the best quality voices (Lx0-1), but had difficulty correctly classifying the mid-range of voices (Lx2-4).

As results from other approaches to voice quality classification have found differences between the computer-based classifications and the SALT assessments in the upper middle categories [5], it was decided to repeat the training and classification of both the

	Impedance signal predicted class %								**Acoustic signal predicted class %**							
		0	**1**	**2**	**3**	**4**	**5**	**6**		**0**	**1**	**2**	**3**	**4**	**5**	**6**
Class	**Lx0**	50	17	13	8	12	0	0	**Lx0**	45	27	18	9	0	0	0
	Lx1	25	48	15	5	7	0	0	**Lx1**	8	47	8	19	8	8	0
	Lx2	2	12	12	27	30	8	10	**Lx2**	4	20	16	40	12	8	0
	Lx3	5	7	10	28	40	8	2	**Lx3**	3	21	21	21	15	15	3
	Lx4	13	5	8	27	37	7	3	**LX4**	8	15	8	30	35	4	0
	Lx5	0	0	8	13	20	43	15	**Lx5**	0	12	4	24	4	32	24
	Lx6	0	0	0	15	20	30	35	**Lx6**	0	0	0	0	9	36	55
	Impedance Results : Overall accuracy = 36.2%								**Acoustic Results** : Overall accuracy = 35.7%							

Table 2. Percentage of correctly classified voices on SALT 7-point scale for impedance and acoustic data.

impedance and acoustic using only three nodes in the ANN output layer, corresponding to "good" (Lx0-1), "medium" (Lx2-4) and "bad" (Lx5-6) classifications of voice quality. The results that were obtained are presented in Table 3.

	Impedance %	**Acoustic %**
Good (Lx0-1)	64	63
Medium (Lx2-4)	26	32
Bad (Lx5-6)	83	91

Table 3. Percentage of correctly classified voices on 3-point scale for impedance and acoustic signals

Again, it should be noted that similar classifications were obtained for the impedance and acoustic signals, and that the ANNs give the best classifications for the "bad" voices.

V. CONCLUSIONS.

A comparative study of voice quality assessment of patients recovering from cancer of the larynx has been made using impedance and acoustic signals. Following earlier work, a collection of short-term and long-term parameters were extracted from each type of signal and input to a ANN, which was successfully trained to match the SALT's assessment of the patient's voice quality using their 7-point scale.

The impedance signal taken gained from the electrolaryngograph is a much cleaner signal than it's equivalent acoustic version, and generally showed more harmonics and contains less noise in the signal. The acoustic signal was more difficult to work with, having fewer harmonics, and the pre-processing stages had to be carried out far more carefully and occasionally produced extra errors that didn't occur with the impedance signal, such as badly fitting Gaussian Mixture Models. However the extra parameters that can be routinely derived from the impendence did not appear to lead to more accurate classifications. This was particularly encouraging as it may allow further research to be carried out using microphones instead of the more expensive and specialised electrolaryngograph.

It was also noted that for both the impedance and the acoustic signals, the ANNs were able to classify the very good (recovered) and the very poor (abnormal) voices well, but performed quite badly with the mid-range classifications. This observation was reproduced when the signals were re-classified into a 3-point scale of "good", "medium" and "bad" voices.

The reason for the poor classifications in the mid-range categories of the 7-point scale and the "medium" category" in the 3-point scale is not yet clear. One possibility is that the SALT are more comfortable with classifying the extreme cases of abnormal and recovered voices, and are less consistent, or possibly less able, to distinguish the intermediate (recovering) voices. If this is the case, then the accuracy and usefulness of the 7-point scale for voice quality assessment would need to be examined.

Alternatively, these problems may be associated with the makeup of frames within a recovering voice, where it might be expected that some frames will be effectively normal, while other are still abnormal. The ANN training process would try and classify all these frames as being characteristic of one of the mid-range categories, as that is the SALT's overall classification of the patient's voice. This possibility is currently under investigation, and it may be necessary to classify individual frames within the voice signal, and then investigate ways of combine the results to achieve closer agreement with the SALT classifications in the mid-range categories.

REFERENCES

[1] Ritchings RT, McGillion M, Moore CJ "Pathological voice quality assessment using artificial neural networks." *Medical Engineering and Physics* 24 (2002) , pp561-564, PII S1350-4533(02)00064-4.

[2] Godino-Llorente, JI, Ritchings, RT, Berry C "The Effects of Inter and Intra Speaker Variability on Pathological Voice Quality Assessment" *3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy 2003.

[3] Fourcin, A.J., Abberton E., Miller, D, Howell D. Laryngograph : "Speech pattern element tools for therapy, training and assessment." *European Journal of Disorders of Communication* 30(2), 1996, pp.101-115

[4] Rabiner, L. and Juang, B.H. Fundamentals of speech recognition. New Jersey Prentice Hall, 1993.

[5] Moore C.J., Manickam K., Slavin N. "Voicing recovery in males following radiotherapy for larynx cancer." *4th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications,* Firenze, Italy 2005.