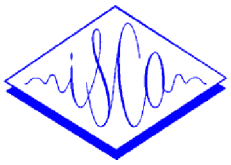


Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

AUTOMATIC ASSESSMENT OF VOICE QUALITY USING FUNDAMENTAL HARMONIC NORMALISED SPECTRA AND GAUSSIAN MIXTURES

MA McGillion[1], RT Ritchings[1], CJ Moore[2]

[1]Department of Computation, UMIST, Manchester, UK, (mm@fs1.co.umist.ac.uk)
[2]North Western Medical Physics, Christie Hospital, Manchester, UK

Abstract

Classification of speech data from male volunteers (normal) and patients recovering from cancer of the larynx (abnormal) is discussed. Analysis of normals and abnormals has shown that there is a significant distinction in the fundamental frequency and harmonic envelope between these groups during constant phonation of vowel sounds. This work proposes a method of deriving the Fundamental-Harmonic Normalised (FHN) spectrum from the speech data and fitting a mixture of Gaussians to model the distribution of power within the FHN spectrum. The aim of this work is to provide a set of features for subsequent classification using an Artificial Neural Network (ANN).

Introduction

An increasingly important factor in prescribing treatment for cancer of the larynx is the quality of voice retained post-therapy. Current techniques for analysis of voice quality following treatment of cancer of the larynx are slow, mainly subjective, and based on limited numbers of retrospective studies. A method of accurately measuring voice quality in cancer patients with respect to a standard normal voice quality is required to enable speech and language therapists (SALTS) to provide an objective assessment during clinical evaluation.

Earlier work has shown that a Multi-Layer Perceptron (MLP) trained using raw power spectral data can accurately classify speech signals as normal or abnormal [1]. The aim of this research is to derive an improved set of features to better model the frequency/power distribution whilst also reducing the dimensionality of the data. A further constraint is that the feature set must provide equal or better classification accuracy of voice quality for patients recovering from cancer of the larynx.

Data Capture

Data was captured under clinical conditions at the Christie and Withington Hospitals in Manchester. The tool used to capture speech sequences was the Electrolaryngograph PCLX system [2]. This system is used to capture electrical impedance signals using impedance pads placed either side of the neck synchronously with acoustic signals using a microphone. The Electrolaryngograph provides four-channel 16-bit analogue-to-digital conversion and two-channel 16-bit digital-to-analogue conversion. A TI TMS320C25 50MHz DSP chip carries out digital signal processing functions, e.g. sampling, filtering, quantisation. Impedance data channels were captured synchronously at 20kHz for up to 3 seconds while the subject phonated the vowel /i/ as steadily as possible.

Feature Extraction

Determining the most suitable features for analysis of voice quality is a non-trivial process and many have been proposed [3,4,5]. Features such as jitter, shimmer, normalised-noise energy and other measures have been used in an attempt to provide an overall evaluation of vocal function.

Following discussions with a SALT it was concluded that their expert knowledge was related to subtle variations in the frequency structure in a patient's stylised speech. It has already been shown that the frequency structure of speech recordings can be used to classify speech quality [6].

A method of separating the spectral envelope (containing the harmonic and formant frequencies) from the distribution of the fundamental frequency and its harmonics has now been developed. First, the impedance time series is transformed to stationarity prior to further processing. The auto-covariance of a 1000 point frame is multiplied with

a Hanning window to suppress variance at increasing lag and then transformed into the frequency domain using the Fast Fourier Transform (FFT).

The Power Spectral Density (PSD) is normalised relative to the fundamental frequency and its harmonics to produce the Fundamental Harmonic Normalised (FHN) spectrum. An example of the FHN spectrum is shown in figure 1.

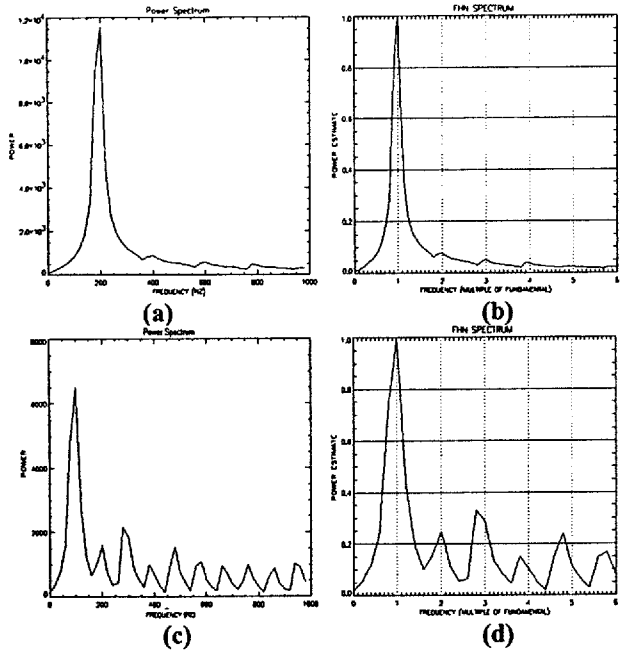
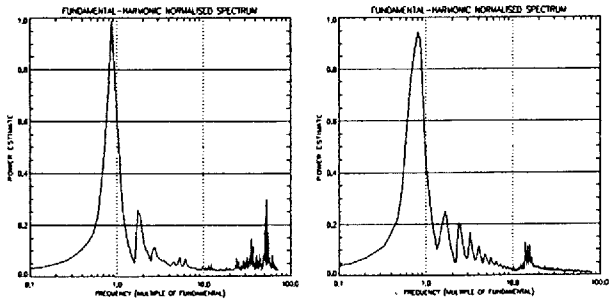


Figure 1. (a)(c) Original power spectrum estimate, (b) (d) Normalised FHN spectrum

The FHN spectra provides a normalised representation of the power spectrum useful for intra-pooling to reveal changes in the patients fundamental-harmonic structure over time. Further, through inter-pooling, the FHN spectra have been shown to provide a good reference standard of normal voice quality in both males and females [7]. An example of the FHN spectra from abnormal and normal subjects is shown in figure 2a and 2b respectively.



(a) (b)

Figure 2. (a) Abnormal male (stage T2), /i/, (b) Normal male, /i/.

Note the well-defined harmonic structure in the normal spectra up to the 6th harmonic. This is in contrast to the abnormal spectra where only the 1st and 2nd harmonic are well defined. Also, there is evidence of increased noise at higher harmonics in the abnormal spectra that may be a result of the vocal tract response to poor vocal fold function.

The fundamental frequency for each frame is determined by performing an iterative search for the most 'significant' peak in the FHN spectra between 60Hz and 300Hz. First, the candidate peaks are found. Then a weight is applied to each candidate so that the most significant candidates are those with maximum gradient and power:

$$\left.\begin{aligned} \overline{magnitude} &= W_{mag}\, x_i \\ \overline{slope} &= W_{slope}(x_i - x_{i+1}) \end{aligned}\right\} W_{mag} + W_{slope} = 1 \quad (1)$$

The weights, W_{mag} and W_{slope} in (1) are determined through empirical analysis and are set to overcome noisy features often present in the frequency spectra. At present, these weights are set to 0.1 and 0.9 respectively. Once the location of the fundamental frequency is determined, amplitude is then normalised relative to the power of the fundamental frequency and the discrete frequency scale is transformed into the discrete harmonic scale (DHS) [7].

Fitting Gaussian Mixtures

A recent approach by Zolfaghari and Robinson [8] to feature extraction from the highly dimensional PSD illustrated the use of Gaussian mixtures to model the PSD. In [8] the formant frequencies from the PSD were successfully modelled using mixtures of Gaussians with application to low bit-rate speech coding.

The FHN spectrum is essentially a histogram that can be interpreted as a probability density function composed of a mixture of Gaussians. A normal (Gaussian) mixture distribution is assumed and has the following definition:

$$f(x) = \sum_{m=1}^{K} c_m \frac{1}{2\pi\sigma_m} e^{-\frac{1}{2}\left[\frac{x-\mu_m}{\sigma_m}\right]^2} \quad (2)$$

where c_m is the mixture weight for each of the M mixtures, μ_m is the mean and σ_m is the standard deviation. The mean, variance and weights of the

Gaussian mixture distributions are initially calculated as follows:

$$\mu_m = \frac{1}{\sum P(x_m)} \sum_{n=1}^{N} nP(x_{mn}) \qquad (3)$$

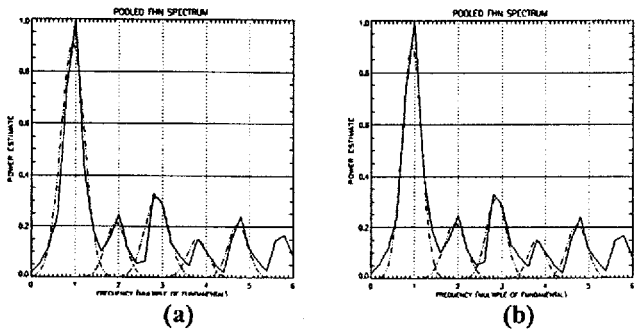
$$\sigma_m^2 = \frac{1}{\sum P(x_m)} \sum_{n=1}^{N} P(x_{mn})(n-\mu_m) \qquad (4)$$

$$c_m = \frac{P(\mu_m)}{\max(f(x))} \qquad (5)$$

where $P(x_m)$ is the amplitude of the signal at frequency x_m within the m_{th} Gaussian, and N is the number of discrete frequencies under consideration. In order to provide a better fit to the data, only amplitudes above a threshold are considered:

$$t_m = \mu_m + \sigma_m \qquad (6)$$

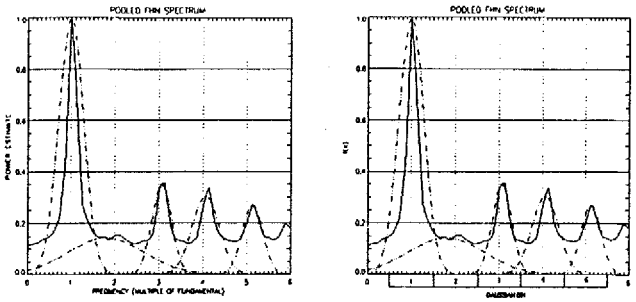
Two methods of fitting the Gaussian mixture have been tested. In the first method, the variance (4) and mixture weight (5) is determined from those points in the region exceeding the threshold value (6). Although this method enables a reasonable fit to the data, analysis of the Gaussians determined that this could be improved. In the second method, the variance (4) and mixture weight (5) were adapted to minimise the mean square error (in a least squares sense) between the Gaussian and the data. A plot of the results of these methods can be seen in figure 3.



(a) (b)

Figure 3. Fitting the GMM (a) method 1, (b) method 2.

The DHS from the fundamental frequency up to the 4th harmonic is divided into 5 bins of equal length, where each bin is centred on the harmonic. This can be seen in figure 4.



(a) (b)

Figure 4. (a) Gaussian mixtures superimposed on FHN spectra, (b) harmonic regions over which Gaussians are defined.

An analysis of the Gaussian distributions in figure 4 is shown in table 1.

μ_m	σ_m^2	c_m
1.000	0.08056	0.71144
1.800	0.74551	0.30959
3.067	0.05004	0.19752
4.067	0.09723	0.23783
5.133	0.05506	0.16034

Table 1. *mean* μ_m, *variance* σ_m^2, *amplitude* c_m, of the Gaussian distributions of figure 3.

Conclusions

Earlier work has shown that a Multi-Layer Perceptron (MLP) trained using raw power spectral data can accurately classify speech signals as normal or abnormal [1]. In this work, FHN spectra were derived from 38 male patients recovering from cancer of the larynx and 38 male volunteers who were considered as having normal voice quality. Gaussian mixtures were fitted to the FHN spectra using method 1 and method 2. Visual analysis of the resulting spectrum indicates that method 2 provides a superior representation of the FHN spectra by minimising the error between the Gaussian distribution and the spectral data.

Fitting Gaussians by minimising the error function will provide the foundation for the development of a new ANN based classification system. Modelling the FHN spectra using Gaussian mixtures provides a normalised representation of the speech signal whilst also reducing the high dimensional PSD to a few coefficients.

The next stage of this work will concentrate on the development of an ANN based system for classification of abnormal voice quality across a range of clinical disorders trained using features such as those in table 1. The goal of this work is the development of a practical on-line system that will be utilised by SALTs during clinical

examinations to provide an objective measure of voice quality.

Acknowledgements

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

References

[1] McGillion MA, Ritchings RT, Moore CJ, et al, 1998, A Neural Network-Based Approach to Objective Voice Quality Assessment, *Procs 18th Int'l Conf. on Expert Systems ES'98*, Cambridge, UK.

[2] Fourcin AJ, Abberton E, Miller D, Howell D, 1995, Laryngograph: Speech pattern element tools for therapy, training and assessment. *European Journal of Disorders of Comunication*, **30:2**, 101-115.

[3] Aref A, Dworkin J, Syamala D, Denton L, Fontanesi J, 1997, Objective evaluation of the quality of voice following radiation therapy for T_1 glottic cancer. *Radiotherapy and Oncology*. **45**, 149-153.

[4] Gavidia-Ceballos L, Hansen JHL, 1996, Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *IEEE Trans. on Biomedical Engineering*, **43:4**, 373-383.

[5] Tadeusiewicz R, Wszolek W, Modrzejewski M, 1998, The evaluation of speech deformation treated for larynx cancer using neural network and pattern recognition methods. *Proceedings of the Int'l Conf. on Engineering Applications of Neural Networks*, EANN'98, Gibraltar.

[6] McGillion MA, Moore CJ, Ritchings RT, 1999, Objective Voice Quality Assessment Using Multi-Layer Perceptrons, *Procs 4th Int'l Workshop Neural Networks in Applications NN'99*, Magdeburg, Germany.

[7] Moore CJ, Slevin N, Winstanley S, 1999, Objective characterisation of Normal Vowel Phonation in Male and Female Populations by Fundamental-Harmonic Spectral Normalisation. *Int'l Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications,* University of Firenze, Italy.

[8] Zolfaghari P, Robinson AJ, 1996, Formant Analysis using Mixtures of Gaussians, *Procs 4th Intl' Conf. on Spoken Language Processing,* Philadelphia, USA.