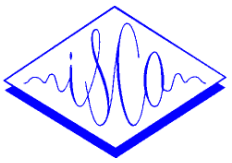

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

LPC, LPCC AND MFCC PARAMETERISATION APPLIED TO THE DETECTION OF VOICE IMPAIRMENTS

[†]Juan I. Godino-Llorente, [†]Santiago Aguilera-Navarro, [††]Pedro Gómez-Vilda
[†] LTR (Lab. de Tecnologí a de Rehabilitación), Universidad Politécnica de Madrid, Ciudad Universitaria, 28041 Madrid, Spain. Tlph: +34.91.5495700 Ext.540 Fax: +34.91.3367323
e-mail: godino@die.upm.es
[††] Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain

ABSTRACT

There is an increased risk for vocal and voice diseases due to the modern way of life. It is well known that most of the vocal and voice diseases cause changes in the acoustic voice signal. These diseases have to be diagnosed and treated during an early stage. Acoustic analysis is a non-invasive technique based on digital processing of speech signal. Acoustic analysis can be a useful tool to diagnose this kind of diseases, furthermore it presents several advantages: it is a non-invasive tool, an objective diagnostic and, also, it can be used for the evaluation of surgical and pharmacological treatments and rehabilitation processes. ENT clinicians use acoustic voice analysis to characterise pathological voices. In this paper, we study threee well known parameterisation approaches applied to the automatic detection of voice disorders. Former and actual works demonstrate that impaired voice detection can be carried out by means of supervised neural nets: MLP (Multilayer perceptron). We have focused our task in detection of impaired voices by means of neural network technology (ANN) and parameters such a LPC, LPCC and MFCC extracted from the voice signal. The performance of the neural network based detector is compared with that using acoustic parameters such a Fo, NHR, NNE, Shimmer, Jitter... as input variables. The aim of this paper is to study and compare those widely used parameterisation method in speech technology applied to the detection of impaired voices.

1. INTRODUCTION

Feature extraction of speech is one of the most important issue in the field of speech recognition. There are two dominant acoustic measurements of speech signal. One is the parametric modelling approach, which is developed to match closely the resonant structure of the human vocal tract that produces the corresponding speech sound. It is mainly derived from Linear Predictive analysis, such as LPC and LPC-based cepstrum (LPCC). The other one is the non-parametric modelling method that is basically originated from the human auditory perception system. FFT-based mel frequency cepstral coefficients are utilised for this purpose. The term “mel” is some kind of measurements of perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scales (mels) is approximately linear below 1KHz and logarithmic at higher frequency. The bandwidth of the critical band varies with the perceived frequency. It is about linear up to 1KHz and increasing logarithmically above 1KHz. The suggested formula that models their relationship is described as:

$$F_{mel} = 2595 \log_{10}\left(1 + \frac{F_{Hz}}{700}\right)$$

The advantages are that those parameters are capable of being immune to noise and that it is easy to warp frequency into a non-uniform scale, such as mel scale.

Our target is to design a LPC, LPCC and MFCC based classifiers, comparing performance of the detectors.

2. PARAMETERIZATION

Three widely used methods have been used to parameterise the speech signal: LPC, LPCC and MFCC coefficients. Such parameters are complemented with first and second order temporal derivatives. An overview of these methods is provided:

2.1 Overview of LPC and LPCC coefficients

LPC and LPCC coefficients are calculated with a method adapted from [3].

$$s[n] \approx a_1 \cdot s[n-1] + a_2 \cdot s[n-2] + \ldots + a_p \cdot s[n-p]$$

Under the assumption of stacionarity, the basic idea behind the LPC model is that a given speech sample at time *n, s[n]* can be approximated as a linear combination of the past *p* speech samples.

This assumption allows us to model the vocal tract as an all-pole system represented by its transfer function *H(z):*

$$H(z) = \frac{G}{1 - \sum_{m=1}^{p} a_m \cdot z^{-m}}$$

where a_m are the LPC coefficients.

Calculation procedure is carried out in 50% overlapped Hamming windowed frames of 20 ms size. Autocorrelation sequence is calculated first, and then used for LPC coefficient calculation by means of Levinson-Durbin algorithm. LPC coefficients are then converted into LPCC coefficients, which have been found to be a robust feature set for use in speech and instrument recognition [11]. LPCC coefficients c_m can be derived directly from the LPC coefficient set a_m by means of the next recursion formula:

$$c_o = \ln(\sigma^2)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad m > p$$

where s^2 is the gain term in the LPC model, and a_m are the LPC coefficients.

2.2. Overview of MFCC Algorithm [3]

We assume that *y[n]* denotes the input speech signal. The complete calculation process of the coefficients can be described step by step as follows:

Step 1: Transform the input speech signal from time domain to frequency domain by applying short-time Fast Fourier Transform (FFT) method.

$$Y(\Omega) = \sum_{n=0}^{F-1} y[n] \cdot w[n] \cdot e^{-j2p \cdot n \frac{m}{F}}$$

where *m = 0,1,2..F-1; F* is frame size, which is generally equal to the power of 2. *w[n]* is the Hamming window function, which is based on the fact that the signal can be regarded as stationary and uninfluenced by the others within short period of time, i.e. the frame size.

Step 2: Find the energy spectrum of each frame.

$$X(\Omega) = |Y(\Omega)|^2$$

Step 3: Calculate the energy in each mel window.

$$S_k = \sum_{j=0}^{\frac{k}{2}-1} W_k(j) \cdot X(j)$$

where $1 \le k \le M$; *M* is the number of the mel windows in mel scale, which ranges from 20 to 24 generally. $W_k(j)$: the triangular weighted function associated with the k^{th} mel window in mel scale.

Step 4: Proceeding with logarithm and cosine transforms, we can figure out the mel frequency cepstral coefficients:

$$mc_m = \sum_{k=1}^{M} \log(S_k) \cos\left[n \cdot (k - 0.5) \cdot \frac{p}{M}\right]$$

where $1 \le n \le L$; *L* is the desired order of MFCC.

2.3. Temporal Derivative [3]

An improved representation can be obtained by extending the analysis to include information about the temporal parameters derivative. Both first and second derivatives have been used. To introduce temporal order into the parameter representation, we denote the m^{th} coefficient at time *t* by $c_m(t)$:

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx m \cdot \sum_{k=-K}^{K} k \cdot c_m(t+k)$$

where *m* is an appropriate normalization constant and *(2K+1)* is the number of frames over which computation is performed.

For each frame *t*, the results of the analysis is a vector of *Q* coefficients, and appended to it two *Q* length vectors more of first and second time derivatives; that is:

$$o(t) = (c_1(t), c_2(t), ..., c_Q(t),$$
$$\Delta c_1(t), \Delta c_2(t), ..., \Delta c_Q(t),$$
$$\Delta\Delta c_1(t), \Delta\Delta c_2(t), ..., \Delta\Delta c_Q(t))$$

where *o(t)* is a vector with *3Q* components. The dimension *Q* depends on the type of parameters used: LPC, LPCC or MFCC.

3. DATABASE

The company Kay Elemetrics recorded to CD-ROM a database of 1,400 voice samples from approximately 700 subjects. The Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs originally developed this database [1].

The acoustic samples are sustained phonation and running speech samples from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders.

The speech samples were collected in a controlled environment and sampled with a 25 or 50 kHz sampling rate and 16 bit of resolution.

The database contains sustained phonation of vowels and running speech samples, but due to the non-stationary features of the speech signal, extraction of

acoustic parameters were carried out over sustained vowel phonation[1]. Phoneme /ah/ has been studied.

Data have been divided into two subsets: the first subset has been be used to train the net (70%), the second (30%), to simulate and validate the results.

4. METHODOLOGY

In the introduction we pointed out that the focus of this research was the classification between pathological and non-pathological voices. That is, detection of voice disorders. Figure 1 shows a block diagram explaining how the pre-processing front-end works. First of all, speech is filtered to avoid aliasing. After that, the signal is converted into a sequence of samples by the A/D converter; later, speech is windowed using Hamming windows. The next module is and endpoint detector in order to avoid unvoiced segments or silences. After the endpoint detection module, those parameters that will be used by the neural net to detect the presence of impairment are calculated. No pre-emphasis has been used.

The extraction of parameters module allows us to calculate, from the voice frames, those parameters (LPC, LPCC and MFCC) that will be used by the pattern recognition module (MLP) to classify. Such parameters easily allow us to model the spectral shape of sustained vowels.

The last module (MLP) is related with the classifier: the neural network architecture selected for the application was the classical Multilayer Feedforward Percepton [2] with a single hidden layer. The Learning algorithm used *is backpropagation with momentum* [5]. Such architecture is widely used in pattern classification.

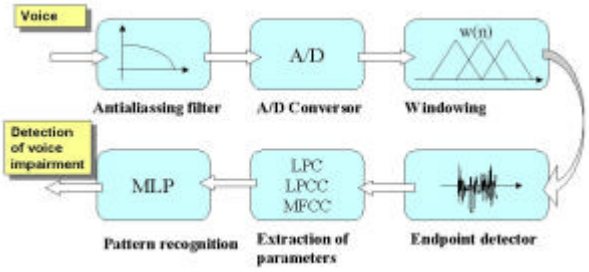

Figure 错误！不能识别的开关参数。: Block diagram of the pre-processor and the detector.

[1] This is the way that ENT clinicians use to sutudy voice disorders. They use to calculate over sustained vowels fundamental frequency and peak amplitude deviations (jitter and shimmer), normalised noise energy, harmonics to noise ratio, etc.[6].

Net weights are randomly initialized. Data were not normalized before giving to the net due to the fact that most of coefficients used fall in an interval among -1 and 1.

Training began with 20.000 epochs. The number of epochs was decreased progressively in order to find the optimum. Sum of mean squared error is controlled as parameter to stop training. The initial number of neurons n_{nhl} in the hidden layer is calculated by means of the next empirical formula [2]:

$$\frac{M}{2 \cdot N} < n_{nhl} < \frac{2 \cdot M}{N}$$

where M is the number of training samples, and N is the number of input coefficients.

The final selected number of voice samples from the database was 135 (53 normal and 72 pathological voices). Normal voice registers are, more or less, 5 seconds long; whereas pathological voice registers are 3 seconds long, due to the fact that people presenting voice impairments have many problems to sustain a vowel during more than 2 or 3 seconds.

As the pre-processing front-end divides speech signal into overlapped frames, we will dispose of one input vector per frame for training the classifier. The total amount of vectors used to train the neural net is around 25.000, each corresponding to a framed window. Nearby 50% correspond to normal voices, and 50% remaining to pathological ones.

5. PERFORMANCE EVALUATION

In order to evaluate the performance of the detector, several kinds of error have been taken account:

- **Correct Rejection**: detector found no event when indeed none was present (CR).
- **Correct detection**: detector found an event when one was present (CD).
- **False negative**: the classifier missed an event (FN)
- **False positive**: the detector found an event when none was present (FP).

It is summarised in the performance matrix shown in table 1.

		EVENT	
		ABSENT	PRESENT
DECISION	ABSENT	**CR**	**FN**
	PRESENT	**FP**	**CD**
	TOTAL	**T1=CR+FP**	**T2=FN+CD**

Table 1: Performance matrix

Results are calculated over the simulation set frames. The final decision about presence or absence of impairment is obtained by means of the normal and pathological ratio calculated as shown in the next expressions:

$$normal \quad ratio = \frac{normal \quad frames}{total \quad frames}$$

$$pathol \quad ratio = \frac{pathol \quad frames}{total \quad frames}$$

The "normal ratio" close to 1, and "pathol ratio" close to 0 corresponds to a normal voice. The "pathol ratio" close to 0, and "normal ratio" close to 1 corresponds to a pathological voice.

Also, test sensibility *(S)*, and efficiency *(E)* were evaluated:

$$S = \frac{CD}{CD + FN}; \quad E = \frac{CR + CD}{CD + FN + FP + CR}$$

6. RESULTS

Three different nets were trained: the first trained with vectors formed by LPC parameters. The second trained with LPCC; and the third using MFCC. Always combining each family of parameters with their first and second time derivatives.

6.1 Results using LPC

The starting hypothesis is the fact that most disorders present an increase of the energy in the high frequency range [6]. As LPC parameters models the spectral envelope, it could be reasonable to think that they could allow as detecting voice impairments.

As the sampling frequency f_s is 25 KHz, we began testing with $N_{lpc}=(f_s/2)+1$ coefficients [3]. The number of coefficients was decreased to find out the optimum.

LPC parameters plus their first and second derivative provides really poor classification accuracy. The classifier never reached a global error minimum. All frames were classified as pathological. The reason is clear: LPC coefficients model the vocal tract response, and most of voice impairments affects to the vocal folds, i.e. to the excitation.

6.2 Results using LPCC

Testing began with $N_{lpcc}=3 \cdot N_{lpc}/2$ coefficients [3]. Also, we have tested it with a different number of LPCC coefficients trying to find out the best number of coefficients for our purpose.

LPCC parameters plus their first and second derivative provides poor classification accuracy. The classifier never reached a global error minimum. Again, all normal frames were misclassified.

6.3 Results using MFCC

Results are shown in table 2. Best results were obtained using 12 MFCC coefficients plus their first and second time derivatives. A MLP with a single hidden layer were used. Only 10 nodes were needed.

		EVENT	
		ABSENT	PRESENT
DECISION	ABSENT	**93.273%**	**10.2218%**
	PRESENT	**6.15516%**	**90.673%**
	TOTAL	**100**	**100**
	Sensibility=0.867891, Efficiency= 0.892087		

Table 2: Frame accuracy using MFCC.

The frame accuracy within the training set was 99,994%. It decreased to approximately 91% when the simulation set was presented to the classifier.

Due to the small frame classification error obtained, all registers (100%) were rightly classified *("normal ratio" and "pathol ratio"* close to 1 and 0). It is shown in table 3.

		EVENT	
		ABSENT	PRESENT
DECISION	ABSENT	**100%**	**100%**
	PRESENT	**0%**	**0%**
	TOTAL	**100**	**100**
	Sensibility = 1, Efficiency = 1		

Table 3: Performance matrix using MFCC.

7. CONCLUSIONS

When interpreting results, we will have to keep in mind that the detector has to maximise percentage of correct detection. The greater the correct detection is, the better the detector is. In principle, we are not so worried about correct rejection. This is due to the fact that is better to forecast a disease when none exist, than forecast no disease when it exists.

Neural networks technology and MFCC coefficients seem to be promisable tools for the automatic detection of voice impairments. Anyway, we have to be wise because the database stores a collection of very significant medical cases. Conclusions have to be tested with a larger database.

Classifying can be done using one single hidden layer. MFCC parameters seem to work much better than LPC and LPCC. We can get a 100% detection accuracy using MFCC.

As demonstrated in [4] detection of voice disorders can be carried out by means of acoustic parameters such a Fo, NHR, VTI, NNE, Shimmer, Jitter, etc [1][6] as input variables. The misclassification accuracy is the same in both cases (0%) using this small database, but reliability seems to be better using MFCC because, in many cases, parameters such a Fo, Jitter, HNR, NNE are difficult to estimate by means of numerical algorithms due to the presence of pathology.

8. FUTURE WORK

Techniques such a GMM could be tested (combined with MFCC) in order to characterise the statistical patterns of these parameters.

Due to the fact that it seems to be easy to automatically detect voice disorders by means MFCC coefficients and neural networks (MLP), the next step will be to distinguish between a set of disorders. For this purpose we may use a similar scheme to the one proposed, but a new and wider database have to be used.

The training and acquisition time could be reduced pruning the input pattern space, in order to use a shorter number of coefficients.

9. REFERENCES

1. "Disordered Voice Database", Version 1.03, Kay Elemetrics Corp, 1994
2. "An Introduction to computing with neural nets" Richard P. Lipmann. IEEE ASSP Magazine April 1987
3. "Fundamentals of speech recognition" L. Rabiner. Prentice Hall. 1993.
4. "On the selection of meaningful speech parameters used by a pathologic/non Pathologic voice register classifier". Godino-Llorente Juan I., Aguilera-Navarro Santiago, Hernández-Espinosa Carlos, Fernández-Redondo Mercedes, Gómez-Vilda Pedro. *EUROSPEECH' 99*, Budapest, Hungary, 1999.
5. "Artificial neural networks for speech analysis/synthesis" Mazin G. Rahim. Chapman & Hall. 1994.
6. "Clinical measurement of speech and voice" R.J. Baken. Taylor & Francis. 1993.