Review on Existing Methods for Features Extraction and Classification in Speech Signal Analysis

[1]Supriya Kadam, [2]Mitul Patel

[1]Student, [2]Associate Professor
[1]Bio-Medical Engineering,
[1]Government Engineering CollegeGandhinagar -Sector 28, Gujarat, India

Abstract :Speech production is the process by which spoken words are selected to be produced with the help of larynx as source and vocal tract as an articulator. Acoustic characteristics like fundamental frequency, formants, jitter and shimmer of the speech signal are used to diagnose the condition of the vocal tract. This paper reviews general methods used for feature extraction and classification of the speech signal into normal and pathological speech.

IndexTerms :**Speech signal, Fundamental frequency, Jitter, Classifiers**

I. Introduction

The physical generation of a discourse sound starts with an inhalation of air. During the inhalation, the diaphragm contracts, which increments the volume of lungs. The inhaled air moves into the lungs as a result of decreased air pressure in the lungs, possessing the low-pressure space in lungs and creating uniform air pressure. When the air is exhaled, the diaphragm reposes, lessening the volume of the lungs and expanding the air pressure which forces the air to move out of the lungs.

The vocal tract shapes and alters the flow of air that moves through it while leaving the lungs. This air moves out through the trachea, the glottis between vocal folds, pass through the larynx and oral cavity. These organs are manipulated in speech production.

Physiological disorder in speech causes changes in the acoustic characteristic of speech signal [1]. These acoustic characteristics are fundamental frequency, formants, jitter, and shimmer. These are used as features to differentiate between normal and abnormal speech using various classifiers. Recorded Speech signal of vowel sound is used for the study.

II. Generalised Block Diagram

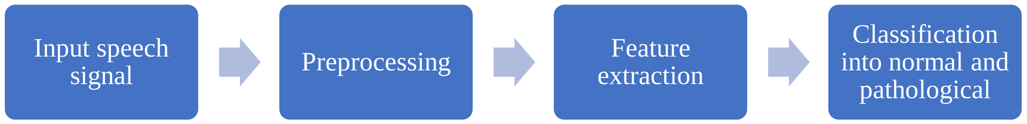

Figure. 1 Block diagram

Figure.1 shows the block diagram of classification of the speech signal into normal and pathological. The input signal is sustained vowel sound recorded for 2-3 seconds.

The input signal is pre-processed and features are extracted using various methods like auto-correlation, cepstrum method, data reduction method and Simplified Inverse Filtering Method.These extracted features are used in the classification of the speech signal.

The different types of classifiers are fuzzy logic, Linear discriminant analysis and principal component analysis with minimum distance classifier, systole activated neural network and Gaussian mixture model super vector kernel-support vector machine classifier.

III. Methods For Feature Extraction

Fundamental Frequency (f0) is characterized by the number of times per moment the vocal folds come together during phonation. Methods used for the acquisition of Fundamental frequency are

A.Auto Correlation Method

It is simply the serial-correlation of a signal with itself. It explains the relationship between the observations among the same variable over different periods of time.

It is a numerical device for finding reoccurring patterns, like the existence of a periodic signal or distinguish the fundamental frequency in a signal utilized by its harmonic frequencies. It is a technique which works on center clipping in time domain [2]. It works according to the following formula:

$$\phi_{\langle l \rangle}(p) = \frac{1}{M} \sum_{q=0}^{M'-1} [x[q+1]\omega(q)][x(q+p+1)\omega(q+p)]$$

$0 \le p \le P_0 \ -1$ (1.1)

In “Equation1”,$P_0$is the value of autocorrelation points to be computed, l is the marker of the beginning sample of the frame, w(q) is an applicable window for examination, M is the segment length being analyzed, M' is the value of signalsamples used in the computation of $\phi_{\langle l \rangle}(p)$.

B. CepstrumMethod

The cepstrum is described as the inverse Fourier transform of the logarithm of the magnitude[3]. And is given by the formula:

$$C_\tau = \left|F^{-1}\left(\log^1\left|F\left(x(t)\right)\right|^2\right)\right|^2 \qquad (1.2)$$

In “Equation1.2”,C_τ is the cepstrum value and $x(t)$ is input signal

The fundamental frequency is estimated using the autocorrelation method,

$$\hat{f} = \frac{1}{\tau_{max}} \ , C_{\tau_{max}} = \tau^{max}\, C_\tau \qquad (1.3)$$

In “Equation1.3”,$\hat{f}$ is fundamental frequency.

C. Data Reduction Method (DARD)

According to N J Millerin this pitch synchronous method, the speech signal is passed through low pass filter[4].Using zero-crossing technique, excursion cycles are identified. Principle excursion cycles are isolated using energy and pitch period limit. Error rectification strategy is used to eliminate the discontinuity of pitch markers.

D. Inverse Filtering Method (SIF)

It is the method stated by J. D. Markel[5]. The speech signal is passed through low pass filter further downsampling it to 5:1 ratio. Linear Predictive Coding Analysisis used and the coefficient of 4th order inverse filter is obtained as :

A_z= 1+ $\sum_{i=1}^{N} [a^i\, z^{-1}]$ (1.4)

In “Equation1.4”,a^i is coefficient of inverse filter, N is the value ofundetermined filter coefficients. Pitch period can be obtained by inserting the autocorrelation function near the peak of the autocorrelation function.

Meantime for classification and accuracy of each method is shown in the following table [3].

Table 1 Comparison of Feature extraction methods.

	Algorithm	**Meantime for classification**	**Accuracy**
1	Cepstrum Method	0.0753 s	84.09 %
2	Autocorrelation Method	0.0727 s	75.00 %
3	Data Reduction Method	0.0799 s	72.72 %
4	Simplified Inverse Filtering Method	0.0963 s	70.45 %

N. Sripriya, S. Poornima, R. Shivaranjani and P. Thangaraju presented feature extraction technique in 2017[6].Jitter is defined as the change in fundamental frequency from one cycle to another cycle.

Glottal closures are better representative of speech signal source. The instance when the amplitude of the signal is highest is known as glottal closure. These closure instants are used to calculate jitter. In classification 85 % accuracy was obtained[6].

$$\text{Jitter} = \frac{1}{N-1}\sum_{j}^{N-1} |T_j - T_{j+1}| \quad (1.5)$$

In “Equation1.5”,N is number of periods and T_j is extracted periods.

IV. Classification Methods

In order to classify speech signal into normal and pathological various classification methods are used.

Examination and diagnosis of vocal fold pathology using a principal component analysis with minimum distance classifier and linear discriminant analysis are given by Jennifer C Saldanha[7]. Cepstral coefficients of the signal are utilized as features to differentiate between ordinary speech from the abnormal speech in this study. The changeability of the glottal waveform can be effortlessly recognized from cepstral parameters in sustained vowels.

PCA is an unsupervised learning method which reduces the dimensionality of data and retains the alteration existingin the original data. It computes a linear transformation which plots data from a high dimensional space to low dimensional sub-space. The aim of LDA is to reducedimensions while preserving the class discriminatory information[7]. To formulate criteria for class separability, between class scatter and within class scatter are utilized in LDA.
The categorization done by LDA shows better result then PCA+MDC as shown in table[7].

Table 2 Comparison of classifiers.

	Results			
	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	LDA classifier	93.14	94	88
2	PCA+MDC	83.42	80.66	100

Fuzzy logic based disorder evaluation procedure is given by Daria Panek[8]. HereKernel Principal Component Analysis,Principal Component Analysis, and Auto-associative Neural Network are fed with a vector made up of 28 acoustic parameters.

Signals were grouped into two categories - healthy and pathological with the use of s-shaped membership function of fuzzy logic. The level of fuzzy membership of normal and pathological voice signals in their respectivecategories was a measure to evaluate the participation of the highlights of a specific class.

The classification of the pathological and normal speech signal using the systole actuated neural network is proposed by MPaulraj[9]. Time domain features are extracted from the energy of the speech signal. An artificial neuron network (ANN) is a numerical tool based on the anatomy and physiology of biological neural networks [10].
Backpropagation is a learning method which updates weights of its network by the means of successive iterations. Two actuation functions specifically, binary activation function and proposed systoleactivation function are created and prepared by the backpropagation algorithm.
The systole actuation function takes less time for training compared to the conventional method another advantage is real-time system development.

Poryasalehi in 2015 presented a feature extraction and classification method based on lifting scheme[11]. This work is depended on Parameterization of wavelet where they relies on various parameters, creating different wavelet systems. Lifting scheme is a method for performing discrete wavelet transform and designing wavelets using lifting steps[12].
In is method adaptive wavelet by lifting scheme and common wavelet were used in decomposition of speech signal up to seven steps. So eight features were acquired belonging abnormal and normal class which were further trained and tested in SVM classifier. Results of classification are compared using other wavelets as shown in the table[11].

Table 3 Comparison of Wavelets and their accuracy

	Wavelet Type	Classification Results
1	Optimization Wavelet by Lifting Scheme(when parameter length is 6)	98.30%
2	Optimization Wavelet by Lifting Scheme(when parameter length is 4)	97.52%
3	Daubechies-4	90.91%
4	Daubechies-10	92.96%
5	Coiflet-4	93.10%
6	Coiflet-5	91.50%
7	Symlet-4	93.66%
8	Symlet-8	92.54%

Comparison GMM-SVM classifier and GMM classifier for the diagnosis of speech pathology is given by Xiang Wang [13].

A GMM describes the distribution of random variable x,

$$g(x)= \sum_{n=1}^{N} \lambda_i \, \varphi(x; m_i, \Sigma_i)(2.1)$$

In “Equation2.1”,g(x) is a GMM, Σ_i, λ_i, and m_i are the diagonal covariance, mixture weight and mean of the GMM. Each component density $\varphi(x; m_i, \Sigma_i)$is an n-dimensional Gaussian function.

Support vector machine is supervised learning algorithm which classifies data using kernel function. Mapping into a high-dimensional feature space is done using kernel function K(x,y)[14],

$$F(x) =\sum_{i=1}^{L} a_i t_i \, K(x, x_i) + d \quad (2.2)$$

$$K(x,y) = b(x)^T b(y) \qquad (2.3)$$

In “Equation2.2” and In“Equation2.3”,”, d is a constant, ti is the ideal output of the class (-1 or 1), b(x) is high dimension spacexi is support vectors trained by the optimization algorithm.

On the basis of performance, the accuracy of GMM classifier is 89.2 ± 4.11 while it is 93.2 ± 3.11 in GMM-SVM Classifiers. The equal error rate (EER)of GMM is 8.0% which is reduced to 4.8% in GMM-SVM. The graph in Figure 2 shows the decrease in equal rate error[13].

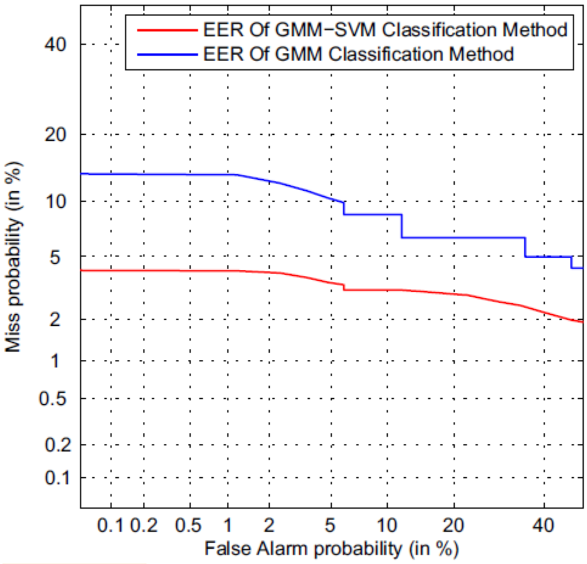

Figure 2. Comparison of equal error rate

V. Conclusion

Various methods for feature extraction including auto-correlation, cepstrum method, data reduction method and Simplified Inverse Filtering Method with their accuracy and response time have been reviewed in this paper. Different classifiers including fuzzy logic, lifting scheme framework, principal component analysis with minimum distance classifier (PCA+MDC) and linear discriminant analysis, systole activated neural network and Gaussian mixture model super vector kernel-support vector machine (GMM-SVM) classifier have been reviewed along with their advantages. These features with appropriate classifiers can be used to diagnose the condition of vocal tract. This makes the procedure less time consuming and cost effective.

VI. References

[1] A. Al nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. Malki, T. Mesallam, and M. Farahat, "Voice Pathology Detection and Classification using Auto-correlation and entropy features in Different Frequency Regions," *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.

[2] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Trans. Acoust.*, vol. 25, no. 1, pp. 24–33, 1977.

[3] K. Kolhatkar, M. Kolte, and J. Lele, "Implementation of pitch detection algorithms for pathological voices," *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 1, 2017.

[4] N. J. Miller, "Pitch detection by data reduction," *Acoust. Speech Signal Process. IEEE Trans.*, vol. 23, no. 1, pp. 72–79, 1975.

[5] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, no. 5, pp. 367–377, 1972.

[6] N. Sripriya, S. Poornima, R. Shivaranjani, and P. Thangaraju, "Non-intrusive technique for pathological voice classification using jitter and shimmer," *Int. Conf. Comput. Commun. Signal Process. Spec. Focus IoT, ICCCSP 2017*, 2017.

[7] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using PCA and LDA," *2013 Int. Conf. Intell. Syst. Signal Process. ISSP 2013*, pp. 140–144, 2013.

[8] D. Panek, A. Skalski, and J. Gajda, "Voice pathology detection by fuzzy logic," *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, vol. 2015–July, pp. 289–293, 2015.

[9] M. P. Paulraj, S. Yaacob, and M. Hariharan, "Diagnosis of vocal fold pathology using time-domain features and systole activated neural network," *Proc. 2009 5th Int. Colloq. Signal Process. Its Appl. CSPA 2009*, pp. 29–32, 2009.

[10] R. Uhrig, "Introduction to Artificial Neural Networks," *Int. Conf. Ind. Electron. Control. Instrum.*, vol. 19, no. 12, pp. 33–37, 1995.

[11] P. Salehi, "Using Patient speech signal for vocal ford disorders detection based on lifting scheme," *Int. Conf. knowlegde based Eng. Innov.*, pp. 561–568, 2015.

[12] A. M. Gavrovska, M. P. Paskaš, and I. S. Reljin, "Wavelet denoising within the lifting scheme framework," *Telfor J.*,

vol. 4, no. 2, pp. 101–106, 2012.

[13] X. Wang, J. Zhang, and Y. Yan, “Discrimination between pathological and normal voices using GMM-SVM approach,” *J. Voice*, vol. 25, no. 1, pp. 38–43, 2011.

[14] S. R. Gunn, “Support Vector Machines for Classification and Regression,” *Image Speech Intell. Syst. Tech. Rep.*, vol. 14, no. May, pp. 230–67, 1998.