Voice Analysis: Various Computational Techniques

Smt. P A Alamelu
Dept. Of Physics SCTIT, Vignan Nagar, Bangalore, India
alamelupa@yahoo.co.in

Leena Govind Gahane
Dept of Physics, Anjuman College of engineering, Sadar, Nagpur, India,
leenagahane@rediffmail.com

Abstract:-This review presents a various methods of studying the human voice and categorizing it. This consists of various comparative and computational techniques, for distortions. Study involves techniques like Vocal Track Length Normalisation, Speech enhancement algorithm for single channel patterns of Indian and English languages, identification of presence of additive coloured noise, detection classification and recognition of Dysphonic patients, Neural network classifiers for speech recognition and MATLAB based back propagation Neural Network for back speech recognition.

Keywords:- *Voice tract length normalisation, Warp factor, speech enhancement, Two –dimensional continuous systems, Voice disorder detection, Voice disorder classification, Speech disorder for disordered voice, Voice XML, SALT, Speech User Application, Speech Abiding Systems.*

*****

1Introduction

Sound is a phenomenon that occurs due to the rapid vibration i.e. compression and rarefaction of air molecules. The human voice changes these variables of frequencies (pitch) and amplitudes (loudness) in cycles per second. The vocal chords compress air molecules at a rate of between 100-5000 times per second to recreate sound. Sound waves are converted from sound to electricity. Every voice generates pattern of amplitude and frequency changes giving individual a unique and recognizable speech tone.

2 Study of voice by using various parameters

2.1 Phoneme classification and Vocal track Length Normalization (VTLN)

ASR (automatic speech recognition) simplifies the interaction between human speakers and machines to accomplish assigned tasks and functions. Speaker variation is divided into- intra and inter speaker variability. Intra-speaker variability, also known as within-speaker variability, is related to difference from within speaker that causes the same speaker to speak differently out of normal habit of Utterances. On the other hand, inter –speaker variability is more on physiological variation between different speakers, although both speaker and environment variations are different. Environment condition is fixed to avoid any variation.

The ASR, which depends on efficient phoneme (smallest linguistic unit) recognition performance and quality of speech signal received. In brief, it includes Inter – Speaker variability, multi speaker frequency warping, fuzzy phoneme recognition, Vocal tract length normalisation (VTLN) and warp factor. The difference between this adaptation and normalisation lies on the ASR component that these approaches are operated on with adaptation focuses on model transformation within the ASR engine while normalisation focuses on front-end signal pre-processing closely collaborates together with feature extraction part of the ASR.

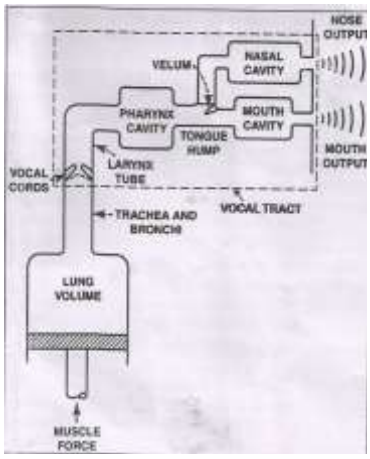

Fig.1 Human speech production system [1].

Speaker adaptation and normalisation are two approaches that handle inter speaker variation. This is rectified using VTLN and warping. Vocal tract length normalisation (VTLN) is the currently known speaker normalisation approach that compensates inter-speaker variability problem [1]. This is done using human speech production system [1]. VTLN yields significant improvement on recognition performance, the drawback is on speaker specific warp factor which normalise spectrum related to one specific speaker's speech.

Physical difference in vocal tract and larynx is corresponding to the physiological differences that are often regarded as the major physical source of inter-speaker

variability [1]. VTLN is the currently known method that performs normalisation process in the signal space by applying some appropriate, speaker –specific, and warping of the frequency scale of a filter bank. Due to physiological differences argument, warp factor is often estimated within small range factor value using maximum likelihood (ML) approach [1].

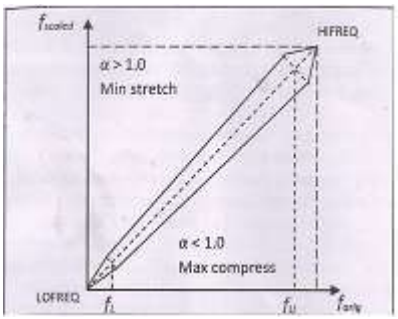

Fig. 2: Piecewise linear warping (derived from [1]).

2.2Analysis of speech Enhancement Algorithm for Single Channel Speech patterns of Indian and English Language:

Speech enhancement is the process of improving the quality and intelligibility of speech signal. All these methods are used for analysis with noisy speech signal of different languages from different speech corpora. The basic spectral subtraction algorithm gives the magnitude spectrum of clean speech by subtracting the noise spectrum from the noisy speech spectrum in the short time Fourier transform domain.

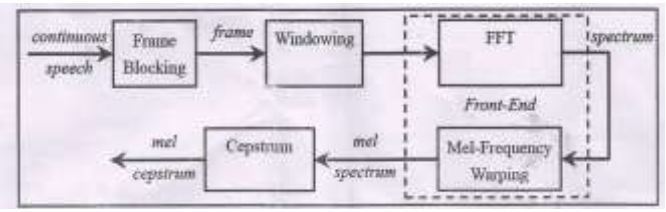

Fig.3: MFCC conversion process flow.

MFCC (Mel-Scale Frequency Cepstral Coefficients) is a Filter bank analysis where additional application of warp factor value and frequency warping range based on piecewise linear warping approach.

The various SINGLE CHANNEL SPEECH ENHANCEMENT TECHNIQUES are:

1. Basis of subjective measure parameters MCS (Modulation Domain Channel Selection) method gives better results than the other methods owing to its use of both acoustic spectrum and modulation spectrum at a given acoustic frequency.

2. Spectral subtraction method: It was explained by Berouti et al. In 1979 [2]. Reduces effect of background (additive) noise. In this noise is additive. Musical noise is overcome by introducing a speech spectral property into speech enhancement methods.

3. Wiener filtering: Utilises spectral property of both speech and noises [2]. The aim of this filter is minimisation of mean square error between the desired signal (clean signal) and the estimated output.

4. MMSE-Speech Presence Uncertainty: The minimum mean square error estimator is a very efficient algorithm that is MMSE-Speech Presence Uncertainty. MMSE-SPU algorithm is motivated by the fact that speech might not be present at all times and at all frequencies.

5. Log-MMSE: Log-spectrum based minimum mean square error is described by Ephraim and Malah after simple MMSE [2]. This algorithm assumes a Gaussian model for the complex spectral amplitudes of both speech and noise.

6. P-MMSE: The squared –error cost function given in traditional MMSE is not subjectively meaningful and emphasises spectral peak (formants) information which takes into account auditory masking effects.

7. Ideal Binary Mask (IdBM): It is based on the principle of comparing the true instantaneous SNR with a preset threshold.

SIMULATION CONDITIONS

1. Speech corpus: The clean speech patterns of each language were added to 12 different types of noise patterns taken from NOIZEX-92 database.
2. Comparison Metrics: The performances of these methods were compared using SNR, SSNR, PSNR and MSE as subjective measures and speech intelligibility index (SII) is considered as objective measures.

Conclusions: On the basis of subjective measure parameters, MCS is superior and gives better results than the other methods owing to its use of both acoustic spectrum and modulation spectrum at a given acoustic frequency.

Analysis for English database corrupted with various noises. English speech mixed 12 types of noises. Fig. 4

shows Volvo car type. Different types of noise values are shown. The noise estimators are MMSE – Minimum mean square estimator. MMSE SPU –

MMSE – Speech presence uncertainty. P-MMSE Postulate MMSE.

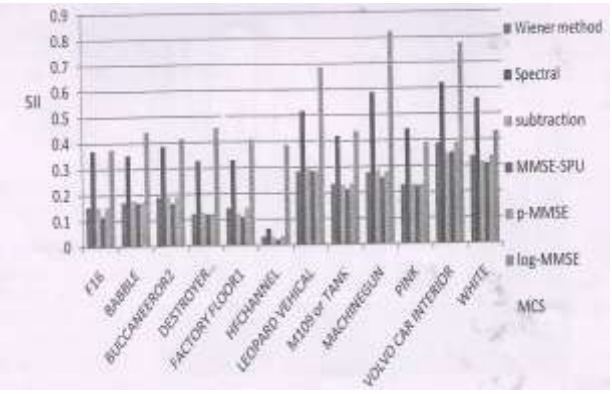

Fig.4: Speech intelligibility index for English speech database

2.3Recursive Identification of Continuous Two-Dimensional Systems in the Presence of Additive Coloured Noise

Although the continuous system identification techniques were developed before the discrete techniques , they were dominated by discrete system identification techniques due to development of digital computers in recent decades [3]. The continuous system identification has two main advantages. First, estimated parameters in the continuous model can be interpreted directly in physically meaningful terms. Second , unlike the discrete methods , sampling problems do not arise and estimated parameters are not functions of sampling interval.

This study involves detection, classification and speech recognition for Dysphonic patients. The performance of three different speech processing systems for medically disordered voice having fold disorders by using different speech features extracted from words rather than vowels. The features compared are as follows: MFCC (mel-frequency cepstral coefficients), LPCC (linear predictive campestral coefficients), PLP (perceptual linear predictive), and RASTA-PLP (relative spectral transform perceptual linear predictive. MFCC is most widely used speech feature for normal people or patients with voice disorders. RASTA Then is Detection, Classification and Speech Recognition was suggested to compensate the channel effect. The first two systems are for voice disorder detection and voice disorder classification. The third system is a digit recognition system for dysphonic patients. The first system provides a binary decision that a person has vocal fold disorders or not. The second system is for voice disorder i.e. type of disorder from the voice. The third system is a speech recognition system that recognizes digits spoken. The support vector machine, hidden markov model, and Gaussian mixture model are modelling techniques used.

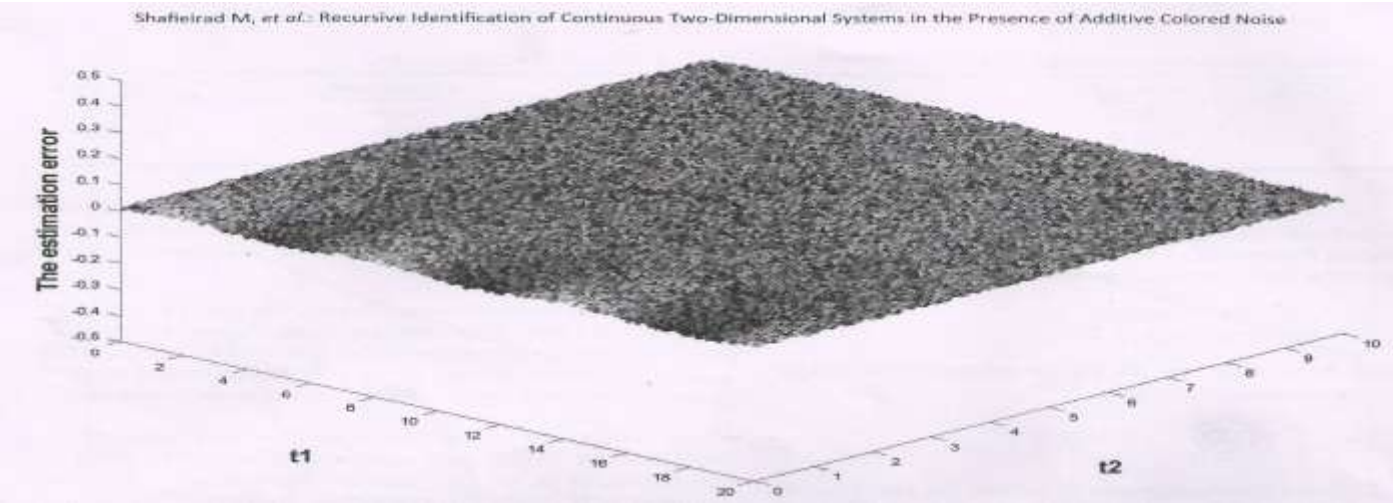

Fig.5 The error between the noise free output x and its estimation $\hat{x}$ [3]

These voice disorder and symptoms are sulcus, cyst, polyp, gred, paralysis.

An automatic speech recognition system for assessment or therapy of a voice disorder is suggested in [4]. The recognition rate of speech recognition system can be affected by the vocabulary size, the speech type (isolated words or continuous speech) , the microphone quality and its placement, and the dependency on text. MFCC , LPCC, PLP AND RASTA_PLP coefficients are extracted from the speech samples of the dysphonic patients and inputted to the pattern matching techniques of HMM and GMM.

Description about recognition systems:

1.1 Speech Corpus : For the disordered voices , the databases concentrated on fivefold disorders,namely cyst, GRED, paralysis , polyp, and sulcus.

1.2 To develop three systems for processing speech affected by vocal fold disorders, different feature extraction and modelling techniques are implemented. The feature extraction techniques MFCC, LPCC,PLP and RASTA_PLP, are used to convert the speech to a parametric representation.

For ex. LPCC is an extension of LPC and can be derived from LPC by using recursion [4], given by equation:

$$C_0 = \ln \sigma^2$$

$$C_m = \sum_{k=1}^{m} (k/m) c_k a_{m-k}, \quad m = 0,1,2,.....,p$$

The LPC features model the vocal tract properties by using all-pole model. These features represent the main vocal tract model. The major components of the MFCC extraction are as follows: frame blocking , windowing , fast Fourier transformation, mel-frequency filtering , and discrete cosine transformation.

1.3Modelling Techniques: Two modelling techniques HMM and GMM are used for voice disorder classification and digit recognition systems to construct the acoustic models of the patients and normal persons.

As in any study, error and disorder cannot be eliminated, Voice Pathology assessment System for Dysphonic patients: detection, classification and speech recognition is another part of this study.

Graph showing the comparison between best performance parameters of the features and ROC curves of different features with highest accuracies.

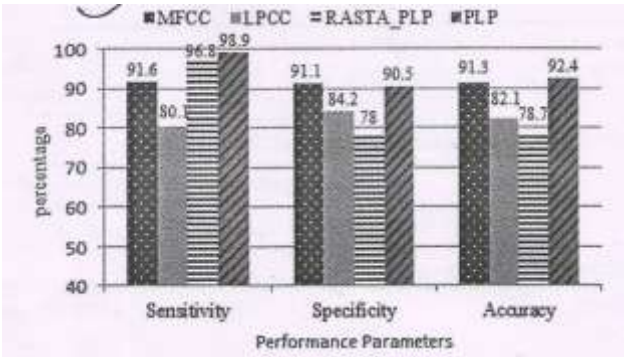

Fig.6: A comparison between best performance parameters of the features [4].

2.4Comprehensive study of Voice XML and SALT (variation and similarity) : For understanding natural languages like English , hindi etc ., two mark up languages are identified as Voice XML and SALT (speech applications language tags).These two provide overview of speech technologies via speech recognition task and speech user applications using speech interface , data and control flow. They work differently for following two reasons i) they have different goals ii) they have different web heritages Voice XML is designed for telephony applications and was developed to allow the specifications of interactive voice response applications in a mark up language that facilitates the authoring of system-driven and mixed- initiative voice dialogs over telephones and cell phones.

2.5Neural network classifiers for speech recognition

Neural nets offer massive parallelism for real-time operation and adaptation, which has the potential of helping to solve difficult speech recognition tasks. Speech recognition, however, will require different nets for different tasks. Different neural net classifiers were reviewed and it was shown that the three-layer perceptrons can form arbitrarily processing. nodes in a left-to right HMM word model. Connections are provide maximal output when there is a match between the input spectral pattern sequence and the sequence expected for the word that the net is designed to recognize.

Viterbi net differs from other neural net approaches because it implements a slightly modified version of a proven algorithm that provides good recognition performance. In addition, the Viterbi net's weights can be computed by using the forward-backward training algorithm [6]. It has a massively parallel architecture that could be used to implement many current HMM word recognizers in hardware. These neural net classifiers performed better than Gaussian classifiers for a digit classification problem. Three-layer perceptrons also performed well for a vowel classification task. A new net, called a feature map classifier, provided rapid single-trialleaming in the course of completing this task. Another new net, the Viterbi net, implemented a temporal decoding algorithm found in current word recognizers by using a parallel neural net architecture and analog.

A Viterbi net is a neural network architecture that uses analog parallel processing to implement the temporal alignment and matching score computation performed in conventional HMM recognizers. Nodes represented by large open triangles correspond to

2.6Components of an isolated word speech recgnizer

An isolated-word speech recognizer must perform four major tasks. The recognizer first includes a preprocessing section that extracts important information from the speech waveform. Typically, the preprocessor breaks the input waveform into 10-ms frames and outputs spectral patterns that represent the input waveform.

One of the most important applications of new-al network architectures is in the realization of compact real-time hardware for speech, vision, and robotics applications.

Talker-dependent isolated-word tests Algorithms based on neural nets have been proposed to address speech

recognition tasks which humans perform with little apparent effort. In this paper, neural net classifiers are described and compared with conventional classification algorithms.

3 Pattern recognition using back-propagation Neural Network on MATLAB

Unlike the traditional sequential machines where rules and formula need to be specified explicitly [12], a neural network learns its functionality by learning from the samples presented .

3.1Characteristics of artificial neural networks

Artificial neural networks have a labelled directed graph structure where nodes perform some computations. They consist of a set of nodes and a set of connections connecting pair of nodes.

3.2 Classification

Classification means assignment of each object to a specific class or group. It is of fundamental importance in a number of areas ranging from image and speech recognition to the social sciences.

3.3 Limitations of using perceptrons

If there are three input dimensions, a two class problem can be solved using a perceptron only if there is a plane that separates samples to different classes. A robust algorithm would achieve a reasonable separation between most of the samples of the two classes. Two algorithms achieve robust classification for linearly non-separable classes - pocket algorithm and least mean square algorithm.

3.4)Pocket algorithm

This algorithm identifies the weight vector with a longest unchanged run as the best solution among the weight vectors examined so far. The contents of the pocket [12] are replaced whenever a new weight vector with a longer successful run is obtained .

3.5)Adalines

Robust recognition may also be achieved by minimizing the mean square error (MSE) instead of the number of misclassifies samples. An adaptive linear element or adaline accomplishes classification by modifying weights in such a way as to minimize the MSE at every iteration training. This can be achieved using gradient descent, since MSE is a quadratic function whose derivative exists everywhere.

3.6Supervised learning using multi-layer networks

Perceptron approach can be extended to solve linearly non-separable classification problems, using layered structure of nodes. Such networks contain one or more layers of hidden nodes that isolate useful features of the input data.

3.7BACK-PROPAGATION NETWORKS

The term back propagation network is used to describe feed-forward neural networks trained using the back propagation learning method. The back propagation algorithm is the modification of least mean square algorithm. It modifies network weights to minimize the mean squared error between the actual and desired outputs of the network.

3.7.1Architecture of back propagation networks

The back propagation algorithm assumes feed-forward neural network architecture. In this architecture nodes are partitioned into layers numbered 0 to L. Here the layer number indicates the distance of a node from the input nodes.

3.7.2 Objectives of back propagation networks

We train the back-propagation network with supervised learning algorithm using a large number of input patterns, say P = 50. For each input vector xp, we have the corresponding desired output vector dp, of dimensions, say K. This collection of input output pairs constitute the training set {xp, dp}. The length of the input vector xp is equal to the number of features of the input pattern [12]. The length of output vector dp is equal to the number of outputs of given application i.e. the number of classes, as decided by the given classification problem.

4 Example the output of Time – Frequency Analysis of chanting Sanskrit divine sound "OM" Mantra is shown:

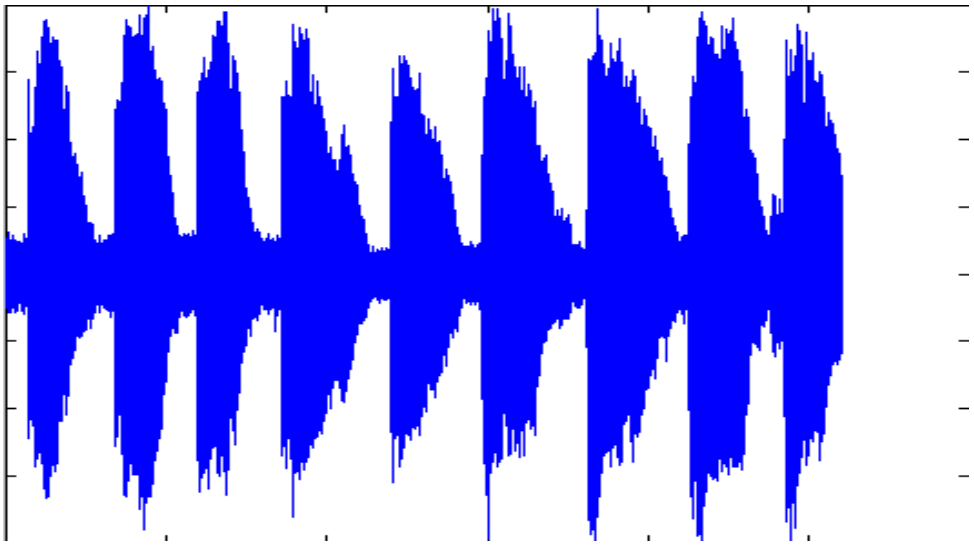

Fig 7: represents the initial chanting of OM by normal person. [10]

5 Conclusion :The objective is to analyse the voice so that the output is as close as possible to the desired output , when a samples input vector is presented to the network. To achieve this objective, the cumulative error of the network needs to be minimized. The difference between the actual output and the desired output represented by error function Err should be non-negative [12]. Further **MATLAB** based feature recognition using back propagation neural network for ASR is to explore how neural networks can be employed to recognize isolated-word speech as an alternative to the traditional methodologies. The general techniques developed here can be further extended to other applications such as sonar target recognition, missile tracking and classification of acousticsignals.

References:

[1] IETE The Institution Of Electronics & Telecommunications Engineers Technical Review|VOL 31|No 2| MAR-APR 2014
[2] IETE The Institution Of Electronics & Telecommunications Engineers Technical Review|VOL 31|N0 1| JAN-FEB 2014
[3] IETE The Institution Of Electronics & Telecommunications Journal Of Research|VOL 60|N0 1|JAN-FEB 2014
[4] IETE The Institution Of Electronics & Telecommunication Journal Of Research|VOL 60|N0 2|MAR-APR 2014
[5] International Journal Of Software Engineering ISSN 0974 Volume 2, Number 1 (2011) , 2 No.1 pp. 31-38 @ International Research Publication House http://www.irphouse.com
[6] The Lincoln laboratory Journal Volume 1, Number (1998)
[7] Artificial Intelligence – Elaine Rich & Kevin Knight IInd edition
[8] Introduction to Artificial Neural Systems –Jacek M Zarada.
[9] Artificial Intelligence –Ela Kumar
[10] IJCSNS International Journal of Computer Science and Network Security, VOL.8No.8,August 2008 – (A A Gurjar and S A Ladhake)
[11] Artificial Neural Networks – B. YegNarayana.
[12] International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (1An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 7, July 2014