Feature Extraction of Voice Segments Using Cepstral Analysis for Voice Regeneration

P. S. Banerjee[1], Baisakhi Chakraborty[2] and **Jaya Banerjee[3]**

Submitted in February, 2015; Accepted in April, 2015

Abstract–Even though a lot of work has been done on areas of speech to text and vice versa or voice detection or similarity analysis of two voice samples but very less emphasis has be given to voice regeneration. General algorithms for distinct voice checking for two voice sources paved way for our endeavor in reconstructing the voice from the source voice samples provided. By utilizing these algorithms and putting further stress on the feature extraction part we tried to fabricate the source voice with different pitch and intonation patterns. This process of uniquely tracing the features and re assembling them to reproduce the target voice is what we have concentrated on in this work. While doing so the aspect of liftering and cepstrum analysis has been utilized to the fullest.

Index Terms – Cepstral Analysi, liftering, cepstrum

1.0 INTRODUCTION

Voice Processing or Speech Processing may be regarded as one of the most integral part of any module dedicated for either speech recognition, as an authentication for granting unique entity access or for speech to text or vice versa. Linguistically and phonologically we may not be able to decipher much from the available waveform due to enormous amount of data for each audio signal that is processed, where as if interpreted in a more sophisticated manner then we may be able to extract relevant unique information for every new incoming data. [1] While defining a cepstrum we can say that it is the output of the estimated spectrum's logarithm is computed and its Inverse Fourier transform is also calculated. The different categories under which the cepstrum may be divided is a complex cepstrum, a real cepstrum, a power cepstrum, and phase cepstrum. Power cepstrum finds its application in the analysis of human voice. As a baseline fact the term "cepstrum" basically originates from the word "spectrum" and is achieved by just reversing the first four letters. The fundamental procedures of cepstra may be listed as quefrency analysis, liftering, or cepstral analysis. The rate at which the various spectrum bands change may be called as cepstrum. Out of the many utilities of it the most primitive use of it was for characterizing the seismic echoes resulting from earthquakes and bomb explosions. But to the best of our concern it is also used to determine the fundamental frequency of human speech. Cepstrum pitch determination is particularly effective because the effects of the vocal excitation (pitch) and vocal tract (formants) are additive in the logarithm of the power spectrum and thus clearly separate. The best possible step will be to have smaller amount or relevant data which may act as a building block for the voice to be recreated. At the outset we can imagine that any signal emanating from a source is a cumulative result of the provided file and the basic response to the file in coordination to it. Both the input as well as the output component are to be treated separately for further processing and this process may be mathematically defined as the deconvolution process.[18]

There are algorithms like Linear predictive coding (LPC), Hybrid Harmonic/Stochastic (HYBRID H/S) and last but not the least the TD PSOLA which we can use to calculate the perturbation or the excitation.[2]

1.1 Organization of the paper

This whole work comprises of the the basic problems associated with the feature extraction of the voice samples. For feature extraction the various techniques available are discussed with their relative advantages and dis advantages. The problem definition has been clearly emphasized and the cepstral analysis part has been forwarded as the possible alternative.

2.0 PROBLEM DEFINITION

At the outset we can imagine that any signal emanating from a source is a cumulative result of the input excitation and the basic system response in coordination to it. Talking in the context of mathematics and digital signal processing the convolution of the input signal with the response of the system may be regarded as the actual output. Bothe the input as well as the output component are to be treated separately for further processing and this process may be mathematically defined as the deconvolution process.[38] Attempts have been made to extract the features of the audio signals more efficiently. If we have the extracted features then mathematically it should be possible to to uniquely identify a speech signal in its digitized form. But due to the massive diversity of the speech waveform and vast amount of data relative to a particular speech waveform we will always require a huge amount of information to be stored for a particular uttered phonetics. Due to this very hindrance it becomes difficult to recreate back the same

[1]*Department of Computer Science & Engineering Jaypee University of Engineering & Technology, Guna (MP) 473 226, India*
[2]*Department of Information Technology, NIT Durgapur West Bengal, India*
[3]*Aryabhatta Institute of Engineering and Management, Durgapur, India*
E-mail:[1]partha1010@gmail.com,
[2]baisakhi.chakraborty@it.nitdgp.ac.in and
[3]jaya2008.banerjee@gmail.com

waveform from the extracted features. [19] Our main aim would be concentrated on grouping the information available from each sample of audio signal into smaller parameters or features. There are many feature extraction techniques that are available, some of the important one are as listed as: FFT, LPC, PLP etc. [35] After the feature extraction has been done the reconstruction of the voice can be achieved by the use of the unique feature vectors.

3.0 REVIEW OF LITERATURE AND RELATED WORK DONE SO FAR

3.1 The voice reconstruction process may be divided into a sequence of steps:

3.1.1. The initial step may be regarded as the pre requisite step where the most unique features of both the source as well as the target dataset is traced and this phase is called the Analysis step.

3.1.2. In the second step we try to map the features computed in the previous step of that of the 3source voice to that of the "to be achieved voice" to as close a proximity as possible.

3.1.3. Synthesis: Last but not the least is the final phase where the modified parameters are used to synthesis or reconstruct the new speech which generally does have the target voice as well as the required prosody too if the module assists.

3.2 Voice processing

The researches on the crux topics of voice and speech is basically revolving around the paradoxical axis of speech or voice synthesis with application areas in the form of text to voice and vice versa with stress on characterization and bifurcation of two or more voice samples. Some of the broad application areas are as enumerated. [3-4]

Generally the combination issues of these are esoteric hence [5] tries to separate the basic glottal source spectra and vocal tract using glottal inverse filtering.

3.3. Review of Application Speech Synthesis Technology

The process of synthesis of speech may be classified as restricted and unrestricted when it is messaging and text-to-speech conversion respectively. These are useful for announcements and also for the people who are impaired visually.[2] In the sections to follow we are going to exemplify some of the topics.

3.3.1. Text-to-Phonetic Conversion

The basic problematic area encountered by any TTS system is linguistic to text and vice versa conversion. The process basically begins with the preprocessing of the text to be converted and then an in-depth analysis of the data for a unique and correct pronunciation. The last step involves the proper computation of the prosodic features. A few of the important steps have been discussed here.

3.3.2. Text preprocessing

The first step i.e. text preprocessing can be understood about its difficulty level from the following example where let's say any English numeral 121 may be at first read as one hundred and twenty one and 2014 as twenty fourteen if inferred as year or two thousand and fourteen for quantizing something for measurement. Some of the similar cases are the distinction between the any numeral and then stating pilot or people. The final area is the fractions and dates which are equally troublesome. 4/14 is four-fourteenths or April Fourteenth.

3.3.3. Pronunciation

The next important task is of correct pronunciation different areas in the text. There are certain types of words that bear a same spelling but have different meaning and sometimes different pronunciations also which are called homographs. Now these type of words are a big obstacle for the overall module.

3.3.4. Prosody

The rhetorical flow of any word or its segment will definitely consider the correct intonation, proper stress at the punctuations. The basic meaning of the uttered phrase and the emotional state of the speaker are some of the deciding factors for the prosodic characteristics. The dependencies of prosodic variations are shown in Figure 1. The ironic situation is that any information in written or textual format doesn't carry the traits of these qualities.[2]

A speech or voice processing system performs two types of functions which is modeling of the signal and the second one is the comparison of the features. [19]. The first stage corresponds to transformation of the signal to parameters and the second stage is the comparison of the similar features from the memory. In this regard hence forth we will be enlisting an analysis part of the related work study and then we will propose our work. Hence this part will be basically divided into

3.4. Feature Extraction:

Feature Extraction and Voice construction based on Pitch Synchronization Over-Lap Add (PSOLA) algorithm

As our study says that in this approach static voice conversion is basically taken care of. The parameters that cannot be changed by the speaker even if the person wants are called the Static parameters. They are vocal tract structure modification and these are basically inherent and natural.[8] The quantitative analysis of the algorithm is dependent on the Quality factor (Q) and The Resemblance factor (R). The parameters are applied for diversified sample of voice.

Flow of Implementation of Psola [9]

a) Application of silence removing algorithm for a silence free given input is the first step.

b) In the next step the voiced and unvoiced decision making algorithm takes care of the output of the previous step.

c) The voiced and unvoiced decision making algorithm is reutilized retrieving the qualitative data about the pitch.

d) The algorithm thus ends up right selection about the voiced and unvoiced aspect.

e) The qualitative data about the pitch is vital as markers for the pitch in the signal.

The PSOLA algorithm assesses these indexes on the pitch for scaling down of each unique signal.

f) The flow process at first checks the scale of the pitch and also converts the speech too. When we do have the data about the pitch indexes of both target as well as the source segments

then the mapping is performed for both and hence the conversion process continues. The conversion of the source to the target is achieved when the converted pitch indexes are processed through the PSOLA.
There are some other algorithms too like Linear predictive coding (LPC), Hybrid Harmonic/Stochastic (HYBRID H/S) and TD-PSOLA which we are not discussing since they are not that beneficial in the present norms.[10]

4.0 PLAN OF WORK, METHODOLOGY AND PROGRESS OF OUR OWN WORK

4.1 Our basic strategy for the module will follow the following steps: Generation of Speech Corpus or Sample (By recording the speech signal using microphone), Classification of Speech, Feature Generation, Feature Selection or Extraction, Recognition of a particular speech, Regeneration of Speech for various Prosodies.

4.2 The generation of the speech may be taken care by using condensational techniques.

4.3 Characterization or classification of speech as voiced and unvoiced and also as isolated words and connected words and sometimes as continuous speech and spontaneous speech.

4.3.1. Voiced and Unvoiced: Voiced or Unvoiced is a term used in phonetics segregate speech sounds, as either voiceless (unvoiced) or voiced. Voicing is defined as a voice when the vocal cords vibrate and is used to define phones [20]. In articulatory level, when the vocal cords vibrate a voiced sound is the output, where as a voiceless sound is one when the vocal cords do not vibrate. Its example is in the voicing of “s” and “z”.
4.3.2. Isolated word are the words which to be recognized requires each utterance to have noiseless which is disturbance less signal of the sample. Another name that might be applicable for this class is Isolated Utterance.[21]
4.3.3. The concept of Connected words can be understood as the utterances that are joined and may be regarded as separated words, but with separate utterances to be allowed with small amount of time difference between them.
4.3.4. In Continuous speech the speaker speaks in his or her natural speed and tone as a continuous voice sample. The continuous speech recognition is the most complex one because it is highly difficult to figure out the boundary of the speech.
4.3.5 The Spontaneous speech is the most natural one and is almost extempore. [22-23]

4.4 Feature Generation: For feature extraction of the speech we have employed OpenSMILE. The acoustic features in the form of low level descriptors are employed to have the values of intensity loudness and speech. [24-25].

4.5 Feature extraction may be considered as one of the most important steps in audio characterization and is similar to most pattern recognition problems. The basic audio features for a few sound samples stored in WAV files are computed and by using a measure of the reparability of the class by the use of histograms of the extracted features. [26-27]
The process of finding the features is as follows which extracts a particular feature and its statistics:
Features Statistics
Energy Entropy Standard Deviation (std)
Signal Energy Std by Mean (average) Ratio
Zero Crossing Rate Std
Spectral Rolloff Std
Spectral Centroid Std
Spectral Flux Std by Mean Ratio [34] [28-29]
4.5.1 Mel Frequency Cepstral Coefficient (MFCC) tutorial
The process of automatic speech recognition begins with computation of the unique features of the audio signal by which only the linguistic aspect of the sample can be identified. Hence by determining the shape of the vocal tract and by use of MFCCs we can correctly and accurately represent this envelope. [30-31]Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. We will deal with the main aspects of MFCCs, and how implement them. [32]
4.5.2. Feature extraction using multi signal wavelet Packet decomposition
It’s simply a feature extraction code using the wavelet packet Transform. [34] The code represents a generalization of the Multisignal 1-D wavelet decomposition.
4.5.3. Short time fourier transformation STFT and Inverse Short time fourier transformation ISTFT
The present code is a Matlab function that provides STFT of a given signal x(n) provides the following: stft - a matrix with complex stft coefficients with time across columns and frequency across rows, f - frequency vector, t - time vector.

5.0 PROPOSED RECONSTRUCTION METHODOLOGY

While implementing our system we may consider the following models which are already modified and are listed below:

5.1 Linear regression techniques on a z-transform implementation.

This idea consists of two voices, one is the source voice and other one is target voice. The first sample ie the target voice is the one in which form we try to observe the required input. The source voice is the sample which contains the information that we need to have reconstructed. [5] This method’s implementation consists of three major stages filter analysis, voice de filtering and voice conversion. The broad outline of each of these methods is as follows. In the first stage, we second using Machine Learning techniques such as minimizing the mean squared error, the components unique to any voice of human, subsequently this is what we refer to as the human voice filter. In the second stage, we use speech signal processing techniques like Z-transform to get the segment of the speech from the given speech signal,[33] by defiltering the unique voice content of the particular human voice. In the third and final stage, we now pass this de filtered voice into the

human voice filter of the target voice, and obtain the final speech in the target voice. [5]

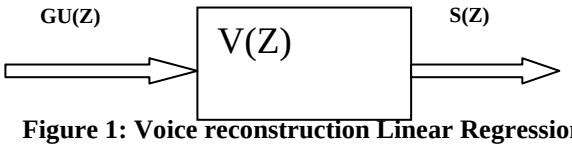

Figure 1: Voice reconstruction Linear Regression Techniques

5.2 The next basic idea is that of auto-regression on stationary time-frames where auto-regression is customized to the properties of the time-frame we consider. This is explained below. At stage 1 we implement the Dynamic Time Warping the second stage concentrates on K-Means clustering which emphasizes the fact that sound samples are stationary for relatively small time frames. This is justified by the fact that for small time frames, which are generally of 10ms, the sound varies very less. Each of the frames would have auto-regressive techniques performed on them.[6]

Stage three is Auto-Regression for time-frame where we use the auto-regression means on relatively stationary frames. While auto-regression process proposes that output samples are dependent on a few previous output and input samples. This uses a feedback from output to determine the future output samples. [7]

Stage four and five consists of Training phase and testing phase respectively where in the training phase we will perform clustering and will obtain coefficients where as in the testing phase we will start once the source speaker's voice sample is obtained. Then we will first split it into stationary time samples, as in the training phase. These stationary time samples, initially in our testing phase, are then detected to be part of a cluster, among the set of clusters obtained in the training phase. The output in the next stage is achieved by use of the cluster obtained. The second stage consists of predicting the output frame given the cluster the input frame. Once the cluster has been obtained, we pull out the coefficients corresponding to the cluster, and use it to linearly generate the samples which mimic the output.[6]

5.3 Cepstral Analysis

Before we start with the cepstral analysis part we can have a brief overview of the Speech Signal Analysis and this may be broadly classified into two basic steps. The first one is the fundamental frequency estimation in frequency domain and the second one is in the time domain.

5.4 The fundamental frequency estimation in frequency domain:

The basic problem associated with fundamental frequency determination is to consider a portion of the input signal and to trace the redundant dominant frequency. While doing so the problems that are encountered are that not all signals are repetitive and those of which are so may not be consistent in the time frame in which we are interested in. Next the signals may be associated with noise and a very interesting problem is that the repetitive signals with time interval of T are also periodic with interval 2T, 3T and hence forth hence we aim at finding the smallest repetitive sequence or the highest fundamental frequency and even signals of constant fundamental frequency may be changing in other ways over the interval of interest. The most trusted way of obtaining the unique fundamental frequency over the desired time frame for steady, noise free, and static speech signals is to use the cepstrum. [11]

5.5 Fundamental frequency estimation in time domain:

In this case the cepstrum is having a periodicity of log spectrum of the signal, but we are more interested in periodicity of the waveform itself. Here to get the fundamental frequency we take help of the autocorrelation. Expecting a proper correlation with itself for short delays is the crux. [11-12]

6.0 FORMANT FREQUENCY ESTIMATION

6.1 Prosody: The rhetorical flow of any word or its segment will definitely consider the correct intonation, proper stress at the punctuations and duration from written text is probably the most challenging problem for years to come. Estimation of formant frequencies is generally more difficult than estimation of fundamental frequency. The spectral shape of the vocal tract excitation strongly influences the observed spectral envelope, such that we cannot guarantee that all vocal tract resonances will cause peaks in the observed spectral envelope, nor that all peaks in the spectral envelope are caused by vocal tract resonances.[12]

The dominant method of formant frequency estimation is based on modelling the speech signal as if it were generated by a particular kind of source and filter:

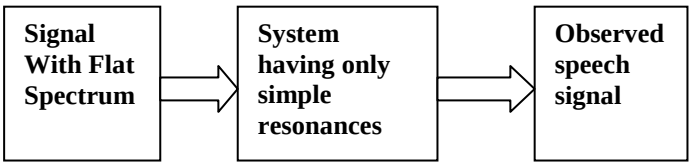

Figure 2: Estimation of formant frequencies

A generic human voice is basically a combination of excitation source and the vocal tract components or the system components. Now while going for the deconvolution process or the cepstral analysis our basic aim is to segregate the various speech components. [13] The convolution of the excitation and the unique traits of the vocal tract filter is the output speech signal.

If the excitation sequence $e(n)$ and the vocal tract filter sequence $h(n)$ then $s(n)$the speech sequence is:

$s(n) = e(n) * h(n)$

(1) The above mentioned can be mentioned in frequency domain as,

$s(\omega) = E\ (\omega) * H(\omega)$

6.2 The output is same in the time domain as expressed in Eqn 2 above. The system components and the multiplied source is transformed using cepstral analysis. [1]

6.2.1Cepstral analysis and its principle steps.
The provided speech spectrum's magnitude is as,

$|S(\omega)| = |E\ (\omega)| * |H(\omega)|$
6.2.3 The logarithmic representation of E(ω) & H(ω) will be,
$\log|S(\omega)| = \log|E\ (\omega)| * \log|H(\omega)|$
6.2.4 In Eqn. (4), for forther processing so as to take the summation the speech spectrum's log the excitation of the vocal tract component is taken into account. [16-17] While considering the vocal tract spectrum and the excitation spectra and its linear combination the bifurcation is performed by estimating the IDFT. Even though it changes but it remains same as that of the time domain.
$c\ (n) = \text{IDFT}(\log\ |S(\omega)|) = IDFT(\log\ |E\ (\omega)| + \log|H(\omega)|)$5. The basic cepstral analysis and the cepstral domain representation is shown below in Fig 3 and 4 respectively [14-15]. The results thus for the proper computation of the cepstrum is described in Fig 4, is given in Fig 5. In Fig 5, $s(n)$ is the voiced frame considered and $X(n)$ is the windowed frame. A multiplication by a hamming window to get $X(n) * |X(\omega)|$ in Fig 5 represent the spectrum of the windowed sequence $X(n)$. As the spectrum of the given frame is symmetric hence one half of the components is plotted. The $\log|X(\omega)|$ represents the log magnitude spectrum obtained by taking logarithm of the $|X(\omega)| * C\ (n)$ of Fig 5.
The various stages of cepstrum computation for an unvoiced frame is plotted in Fig 6.

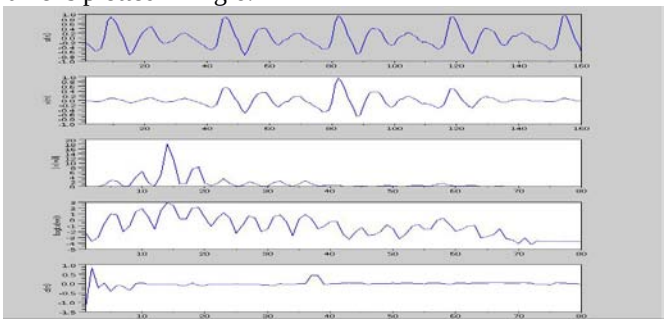

Figure 3: milli second cepstrum for voiced speech segment

6.2.5 Filtering operation or Liftering
In frequency domain when we perform filtering, we may call this overall process as Liftering. In Liftering we multiply the overall cepstrum with a rectangular window and this leads to our expected quefrency region of analysis at a desired time frame. High-time liftering and low time liftering are the two varients of liftering.

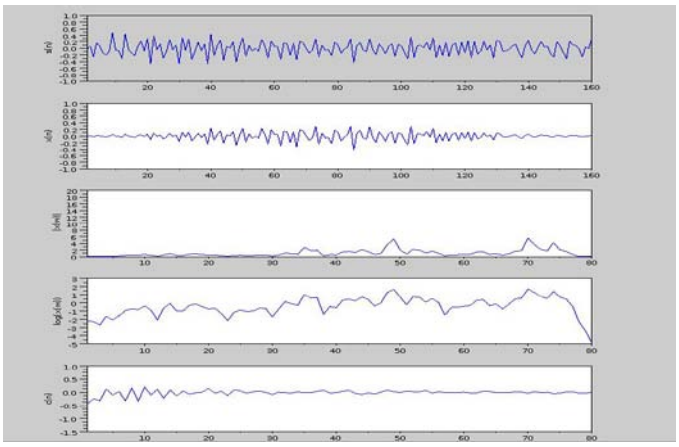

Figure 4: 20 milli second cepstrum for unvoiced speech segment

Low-time Liftering for Formant estimation
For quick varying vocal tract characteristics it's pretty easier to estimate the values but when the changes are relatively slower then low time liftering is used on the given speech sequence. This can be represented as follows,

$$W_e[n] = \begin{cases} 1,0 \le n \le L_c \\ 0, L_c \le n \le \frac{N}{2} \end{cases}$$

6.2.6. Where, L_c is the lower exclusion value of the liftering window,

$\frac{N}{2}$ is equal to the cepstrum length's half. In our case for the test case the value of L_c is 15 or sometimes 20. The parameters for the vocal tract characteristics are computed by the multiplication of cepstrum $C\ (n)$ and low time liftering window as mentioned in Eqn. (7) below.
$C_e(n) = W_e\ [n] * C(n)$
6.2.7. To get the log magnitude spectrum we have to have DFT applied on low time lifter. The output is basically the vocal tract spectrum of the provided short term speech and is denoted as below in Eqn. (8).
$\log[|H\ (W)|] = DFT\ [C_e(n)]$
The values of formant location and bandwidth can be calculated from vocal tract cepstrum. The heighest points of the smooth vocal tract spectrum are basically the values for the formant locations. The block diagram of formant estimation using low-time liftering is in Fig 7. Cepstrum of a voiced segment and low-time liftering window is shown in Fig 8, where as formant locations from vocal tract spectrum is depicted in Fig 9.

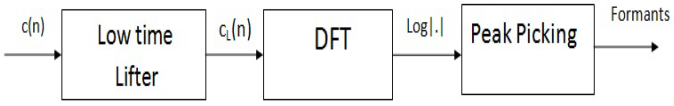

Figure 5: Block diagram representing low-time liftering

*6.2.*8. The values of formant location and bandwidth can be calculated from vocal tract cepstrum. The heighest points of the

smooth vocal tract spectrum are basically the values for the formant locations. The block diagram of formant estimation using low-time liftering is in Fig 4 and 5.

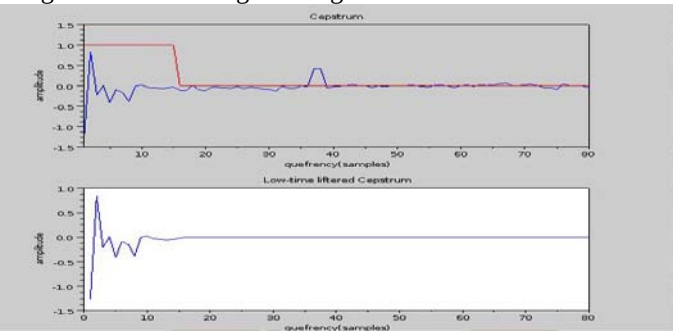

Figure 6: The voice segments Cepstrum

6.2.9. Liftering

The cepstrum liftering is the only way to find the vocal tract component and excitation component. The low-time liftering is applicable for the vocaltract component low-time liftering is done and high-time liftering is used for excitation component.

6.2.10. Computation of Formant frequency estimation:

This type of analysis is called source-filter separation, and in the case of formant frequency estimation, we are interested only in the modelled system and the frequencies of its resonances. To find the best matching system we use a method of analysis called Linear Prediction. Linear prediction models the signal as if it were generated by a signal of minimum energy being passed through a purely-recursive IIR filter. We will demonstrate the idea by using LPC to find the best IIR filter from a section of speech signal and then plotting the filter's frequency response.

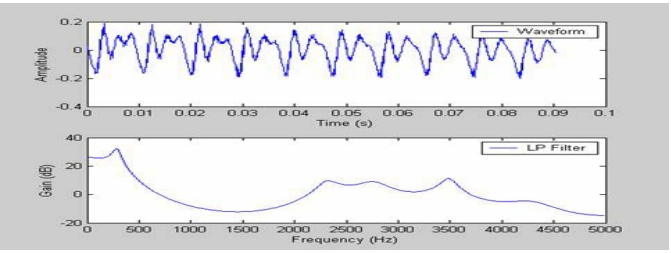

Figure 7: Formant frequency estimation

7.0 RESULTS

While plotting the fundamental frequency, the unique fundamental frequencies of speech are ovserved for their highest points in the corresponding quefrency region. In our case we searched for the the peak between 1 and 20ms in the whole cepstrum, which is clearly. The autocorrelation function for the speech signal's unique part can be seen in above incase of time domain estimation. The autocorrelation function is the peak when there is no latency and subsequently when the delays are ± 1 period, ± 2 periods, etc, we can estimate the it by finding out the highest point in the delay interval corresponding to the normal pitch range in speech, as in our case it is 2ms(=500Hz) and 20ms (=50Hz).

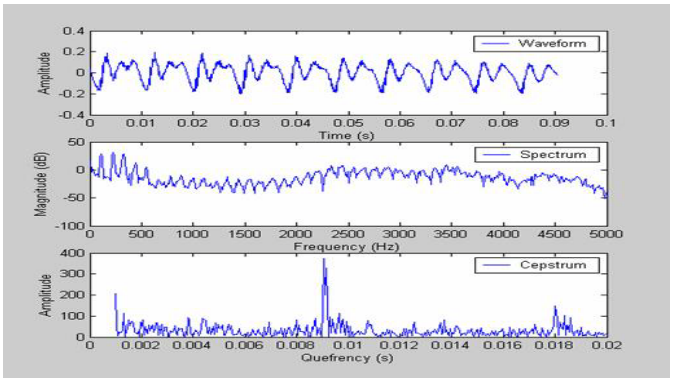

Figure 8: Fundamental Frequency Estimations

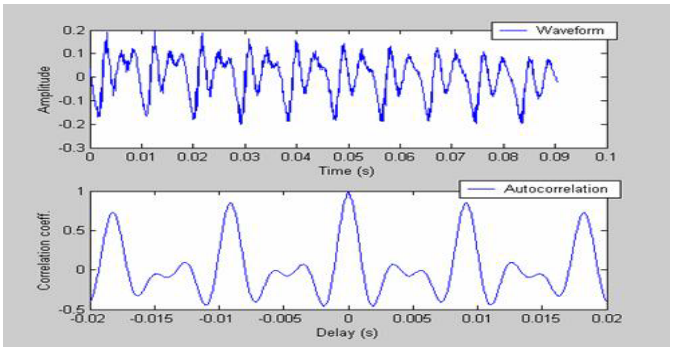

Figure 9: Fundamental Frequency Estimation in Time Domain

The cepstrum shows better results when the fundamental frequency is at the intermediate range and the fluctuations are not so high and also when the noise level is less. The cepstrum analysis is disadvantageous because of its high computation cost due to frequence domain processing. The autocorrelation function for the speech signal's unique part can be seen in Fig 8 & 9 incase of time domain estimation. The autocorrelation function is at the highest point when there is no delay and subsequently when the delays are ± 1 period, ± 2 periods, etc, we can estimate the it by finding out the highest point in the delay interval corresponding to the normal pitch range in speech, as in our case it is 2ms(=500Hz) and 20ms (=50Hz).

The autocorrelation approach works best when the signal is of low, regular pitch and when the spectral content of the signal is not changing too rapidly. The autocorrelation method is prone to pitch halving errors where a delay of two pitch periods is chosen by mistake. It can also be influenced by periodicity in the signal caused by formant resonances, particularly for female voices where F1 can be lower in frequency than Fx.

8.0 CONCLUSION AND FUTURE WORK

Even though in our present technological scenario we can observe a lot of applications of voice of speech processing in the form of recognizers, authentication systems, conversion

between speech to text and vice versa but the broad line benefits of voice processing may be a bit more exhaustive.
To have a better overview of these we may consider NLP as an area where we may use a syntactic parser on the input text and speech recognition's output may be used for information extraction techniques.

9.0 REFERENCES

[1]. Cepstral Analysis of Speech (Theory) : Speech Signal Processing Laboratory : Electronics & Communications : IIT GUWAHATI Virtual Lab.

[2]. P. S. Banerjee, Uttam Kumar Roy, "Modified PSOLA-Genetic Algorithm based approach for Voice Re-Construction", Journal on Information Technology, September – November 2013.

[3]. Voice Conversion for Unknown Speakers, Hui Ye and Steve Young Wiley India 2012.

[4]. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations, Hui Ye, Student Member, IEEE, and Steve Young, Member, IEEE.

[5]. Reconstruction of human voice for impersonation Final report Amritha Raghunath Gunaa Arumugam Veerapandian Vignesh Ganapathi Subramanian 18 November, 2013.

[6]. Ye. H. and S. Young (2003).” Perceptually Weighted Linear Transformations for Voice Conver- sion”. Eurospeech 2003, Geneva.

[7]. Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E. & Russell, S. (In press). The Caterpillar: A Novel Reading Passage for Assessment of Motor Speech Disorders. American Journal of Speech Language Pathology.

[8]. Ganvit,Y Lokhandwala, MA and Bhatt, NS (2012). ”Implementation and Overall Performance Evaluation of Voice Morphing based on PSOLA Algorithm”, International Journal of Advanced Engineering Technology.

[9]. “Pitch Conversion Based on Pitch Mark Mapping” Srikanth Mangayyagari and Ravi Sankar Department of Electrical Engineering, University of South Florida, Tampa, FL 33620, USA E-mail: {smangayy, sankar}@eng.usf.edu

[10]. Patel, R., Hustad, K. Connaghan, K.P. & Furr, W. Relationship Between Prosody and Intelligibility in Children with Dysarthria. Journal of Medical Speech Language Pathology.

[11]. http://svr-www.eng.cam.ac.uk/~ajr/SA95/node34.html

[12]. http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html

[13]. R. Deller Jr., J.H.L. Hansen, J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, 2000.

[14]. A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-Time Signal Processing, U. Saddle River, Ed. NJ:Prentice Hall, 1999.

[15]. L.R. Rabiner and R.W. Schafer, Theory and Application of Digital Speech Processing, First

[16]. Edition, Prentice Hall, New York, 2011.

[17]. D. O`Shaughnessy, Speech Communications: Human and Machine, Second Edition, University Press, India, 2004.

[18]. L.Rabiner, B.Juang, B.Yegnanarayana, Fundamentals of speech recognition, Pearson, India, 2010.

[19]. Urmila Shrawankar, Dr. Vilas Thakare “Techniques For Feature Extraction In Speech Recognition System : A Comparative Study” M.Tech dissertation Amravati University, 2011.

[20]. J. W. Picone, "Signal modelling technique in speech recognition," Proc. Of the IEEE, vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[21]. Shanthi Therese S.,Chelpa, “Lingam Review of Feature Extraction Techniques in Automatic Speech Recognition” International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.2, Issue No.6, pp : 479-484 1 June 2013.

[22]. Lawrence R. Rabiner, et. Al. Speech Recognition by Machine. 2000 CRC Press LLC.

[23]. Meysam Mohamad pour, Fardad Farokhi, “An Advanced Method for Speech Recognition”, World Academy of Science, Engineering and Technology 25, 2009.

[24]. Simon Kinga and Joe Frankel, Recognition, “Speech production knowledge in automatic speech recognition”, Journal of Acoustic Society of America, Oct 2006.

[25]. Tickle A, Raghu S and Elshaw M, “Emotional recognition from the speech signal for a virtual education agent” Sensors & their Applications XVII, IOP Publishing, Journal of Physics: Conference Series 450 (2013).

[26]. Eyden, F., Wollmer, M., and Schuller, B. 2010 Opensmile: the munich versatile and fast opensource audio feature extractor, MM '10 Proceedings of the international conference on Multimedia, pp. 1459-1462.

[27]. R. N. Khushaba, A. Al-Jumaily, and A. Al-Ani, “Novel Feature Extraction Method based on Fuzzy Entropy and Wavelet Packet Transform for Myoelectric Control”, 7th International Symposium on Communications and Information Technologies ISCIT2007, Sydney, Australia, pp. 352 – 357.

[28]. R. N. Khushaba, S. Kodagoa, S. Lal, and G. Dissanayake, “Driver Drowsiness Classification Using Fuzzy Wavelet Packet Based Feature Extraction Algorithm”, IEEE Transaction on Biomedical Engineering, vol. 58, no. 1, pp. 121-131, 2011.

[29]. J. Benesty, M. Sondhi, Y. Huang. Springer Handbook of Speech Processing. Berlin, Springer, 2008.

[30]. B. Boashash. Time Frequency Signal Analysis and Processing: A Comprehensive Reference. Oxford, Elsevier, 2003.

[31]. T. Dutoit, F. Marqures. Applied Signal Processing: A MATLAB-Based Proof of Concept. New York, Springer, 2009.
[32]. J. Allen. "Application of the short-time Fourier transform to speech processing and spectral Analysis". Proc. IEEE ICASSP-82, pp. 1012-1015, 1982.
[33]. J. Smith III, X. Serra. "PARSHL:An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation". Tokyo, Proceedings of the International Computer Music Conference (ICMC-87), pp. 290 – 297, 1987.
[34]. Vogt, T., André, E., and Bee., N. 2008 EmoVoice - A framework for online recognition of emotions from voice. In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee.
[35]. www.mathwork.com
[36]. arxiv.org
[37]. dspace.thapar.edu:8080
[38]. Cepstral Analysis of Speech (Theory): Speech Signal Processing Laboratory: Electronics & Communications: IIT GUWAHATI Virtual Lab.