# Formant Frequency Estimation for Speaker

**Aabha Sahu[1] Ms. Lakhwinder Kaur[2]**
[1]Research Scholar [2]Professor
[1,2]Department of Electronics & Telecommunication Engineering
[1,2]Rungta College of Engineering & Technology, Bhilai, India

*Abstract*— The normal resonant modes of vibration of the vocal tract are called formants. The formant is one of the most important features in speech signals, and is used for many applications, such as speech recognition, speech characterization, and synthesis. Formants frequencies represent the most immediate source of articulatory information and are critical in speech perception. Previous formant extraction methods can largely be classified into spectral peak picking, root extraction, and analysis by synthesis. The spectral peak picking methods and their variants have been widely used for a long time because of low computational complexity, but they often seriously suffer from the peak merger problems, where two adjoining formants are identified into a single one. In this paper, we propose a new formant extraction algorithm that the short term spectral. In the proposed algorithm, the formant candidates are found by using the short term spectral method. It requires a data segment with suitable length such that the harmonics can be resolved and the data within that length should be approximately stationary. Windowing can be done to extract the required length of the data segment.

***Keywords:*** Formant, LPC, Short-Term Spectral Analysis

## I. INTRODUCTION

In our everyday lives there are many forms of communication, for instance: body language, textual language, pictorial language and speech, etc. However amongst those forms speech is always regarded as the most powerful form because of its rich dimensions character. Except for the speech text (words), the rich dimensions also refer as the gender, attitude, emotion, health situation and identity of a speaker. Such information is very important for an effective communication. From the signal processing point of view, speech can be characterized in terms of the signal carrying message information. The waveform could be one of the representations of speech, and this kind of signal has been most useful in practical applications. Extracting from speech signal, we could get three main kinds of information: Speech Text, Language and Speaker Identification.

### *A. Formant*

A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency, roughly one in each 1000Hz band. Or, to put it differently, formants occur at roughly 1000Hz intervals. Each formant corresponds to a resonance in the vocal tract. Formants can be seen very clearly in a wideband spectrogram, where they are displayed as dark bands. The darker a formant is reproduced in the spectrogram, the stronger it is (the more energy there is there, or the more audible it is). A formant is a peak in an acoustic frequency spectrum which results from the resonant frequencies of any acoustic system. It is most commonly used in phonetics or acoustics involving the resonant frequencies of vocal tracts or musical instruments. However, it is equally valid to talk about the formant frequencies of the air in a room.

Formants are the distinguishing or meaningful frequency components of human speech and of singing. By definition, the information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. Formants are the characteristic partials that identify vowels to the listener. The formant with the lowest frequency is called f1, the second f2, the third f3, and the fourth f4. Most often the two first formants, f1 and f2, are enough to disambiguate the vowel. These two formants are primarily determined by the position of the tongue. f1 has a higher frequency when the tongue is lowered, and f2 has a higher frequency when the tongue is forward. Generally, formants move about in a range of approximately 1000 Hz for a male adult, with 1000 Hz per formant. Vowels will almost always have four or more distinguishable formants; sometimes there are more than six.

### *B. Speech Recognition*

During the past four decades, a large number of speech processing techniques have been proposed and implemented, and a number of significant advances have been witted in this field during the last one to two decades, which are spurred by the high speed developing algorithms, computational architectures and hardware. Speech recognition refers to the ability of a machine or program to recognize or identify spoken words and carry out voice. The spoken words are digitized into sequence of numbers, and matched against coded dictionaries so as to identify the words. Speech recognition systems are normally classified as to following aspects:

1) Whether system requires users to train it so as to recognize users'
2) speech patterns; Whether system is able to recognize continuous
3) Speech or discrete words; whether system is able to recognize small vocabulary or large one.

A number of speech recognition systems are already available on the market now. The best can recognize thousands of words. Some are speaker-dependent, others are discrete speech systems. With the development of this field speech recognition systems are entering the mainstream, and are being used as an alternative to keyboards.

## II. PROPOSED METHODOLOGY

Types of formant frequency estimation -

1) Short term spectral analysis.
2) Linear prediction coefficients (LPC).

### *A. Short-term Spectral Analysis*

Speech signal changes continuously due to the movements of vocal system, and it is intrinsically non-stationary.

Nonetheless, in short segments, typically 20 to 40ms, speech could be regarded as pseudo-stationary signal. Speech analysis is generally carried out in frequency domain with short segments across which the speech signal is assumed to be stationary, and this kind of analysis is often called short-term spectral analysis.

Short-term speech analysis could be summarized as following sequences:

1) Block the speech signal into frames with the length of 20 to 40ms, and overlap of 50% to 75% (the overlap is to prevent lacking of information);
2) Windowing each frame with some window functions (windowing is to avoid problem brought by truncation of the signal);
3) Spectral analyzing frame by frame to transfer speech signal into short-term spectrum.
4) Features extraction to convert speech into parameter representation.

*1) Window Functions*

Windowing is to reduce the effect of the spectral artifacts from framing process. In time domain, windowing is a pointwise multiplication between the framed signal and the window function. Whereas in frequency domain, the combination becomes the convolution between the short-term spectrum and the transfer function of the window. A good window function has a narrow mainlobe and low sidelobe levels in their transfer function. The windows commonly used during the frequency analysis of speech sounds are Hamming and Hanning window. They both belong to raised cosine windows family. These windows are formed by inverting and shifting a single cycle of a cosine so that to constrain the values in a specific range: [0, 1] for Hanning window; [0.054, 1] for Hamming window. Based on the same function, shown as follow:

$$W(n)=\varphi-(1-\varphi).\cos(2\pi n/(N-1))$$

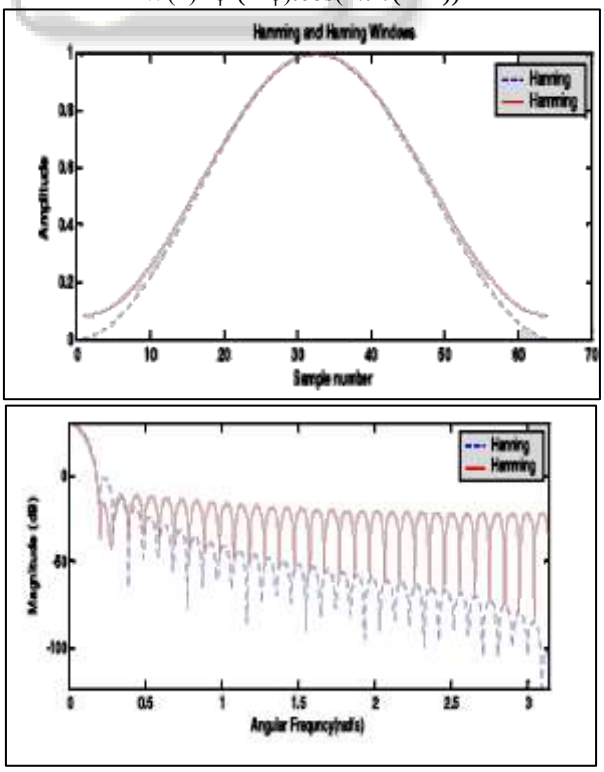
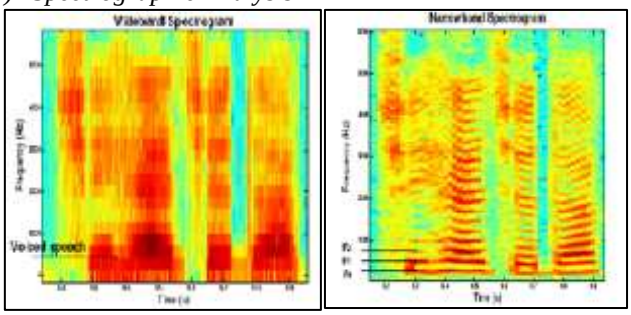



Fig. 3.1: Hamming and hanning windows

Figure 3.1 shows the waveforms and magnitute resposes of hamming window (red and solid line) and Hanning Window (Blue anddash line) with 64 samples. In time domain, Hamming window dows not get as close to zero near the edges as the Hanning window does not get as close to zero near the edges as the Hanning window does. In frequency domain, the main lobes of both Hamming and hanning have the same width which is $8\pi/N$ whereas the Hamming window has lower side lobes adjacent to the main lobe than the Hanning window has, and side lobes farther from the main lobe are lower for the hanning window.

*2) Spectrographic Analysis*



(a)(b) show the spectrogram of the same utterance with different size of windows. (a) is the wide band spectrogram with 56 samples at 16 kHz, corresponding the time spacing is 3.5 ms. (b) is the narrowband spectrograms with 512 samples at 16 kHz, corresponding the time spacing is 32 ms. The pointed part in (a) shows the voice speech. Therefore wideband spectrograms can be used to track the voiced speech. Whereas in (b) harmonics, the red horizontal bars, can be clearly identified. The three arrows from bottom up in (b) point out the fundamental frequency F0, the first formant F1, and the second formant F2.Thus narrowband spectrograms can be used to reveal individual harmonics, and to estimate F0.

### *B. Linear Prediction Coefficients*

Linear prediction is one of the promising technique for speech coding. Linear prediction can be used for recognition as well as synthesis. The most common approach for formant frequency estimation is the Linear Predictive Coding (LPC) method, which can extract the formant frequencies effectively by finding the roots of the linear predictor polynomial, or by picking the peaks of the LPC spectrum. However, in noisy environments, the LPC method is affected by noise and less effective. Linear prediction coefficients are used as coefficients of digital filter. Depending upon the coefficients, the signal is passed/ filtered by the digital filters.

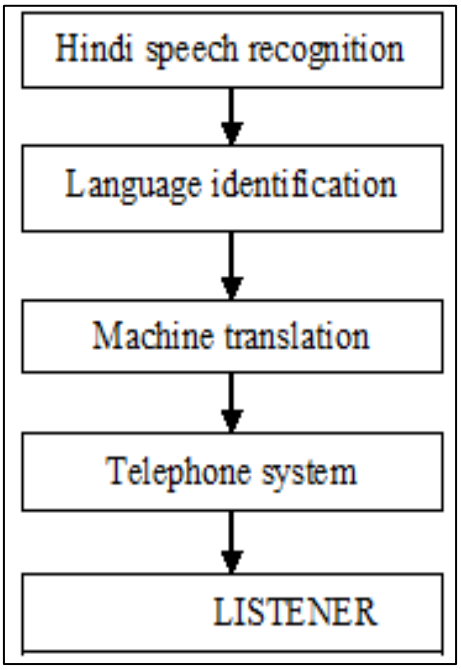
## III. System Design

### *A. Speech Signal Processing*

*1) Speech fundamentals*

1) Speech production
2) Speech processing
3) Speech perception/compression/enhancement

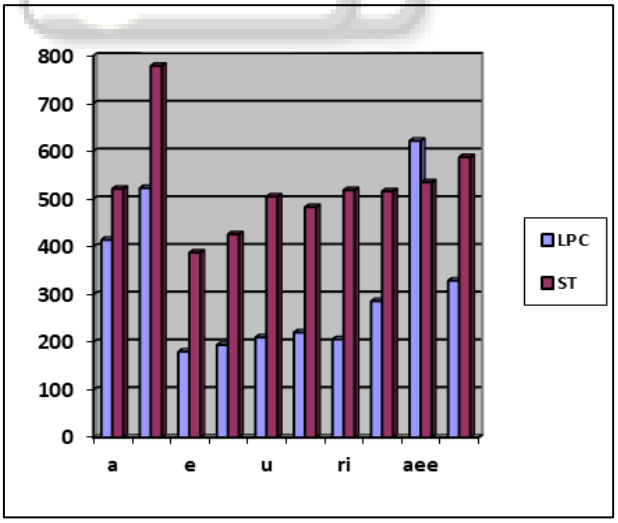
– Speech input
– Speech output




## IV. Results and Discussion

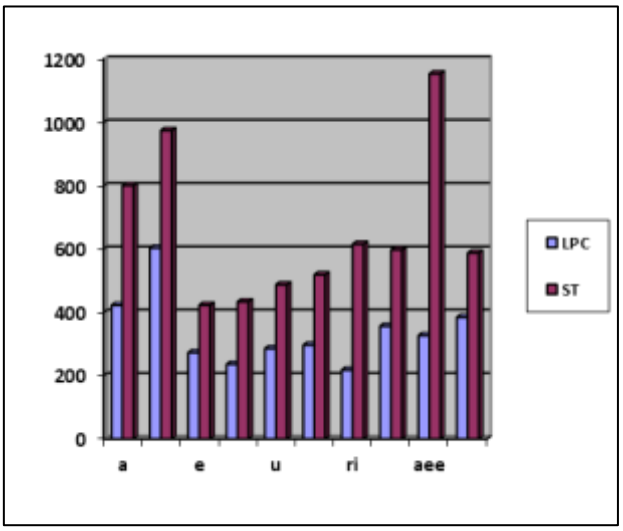
Listeners outperform Automatic speech recognition systems in each and every speech recognition task. Modern high-tech automatic speech recognition systems perform very well in environments, where the speech signals are reasonably clean. In most of the cases recognition by machines degrades dramatically with slight adjustment in speech signals or speaking environment, thus this complex algorithms are used to represent this unpredictability. So, the speech can be easily recognized through the spectrogram.

### *A. Average frequencies of F1 for Males*




Graph 1: F1 (Males) between LPC & ST

### *B. Average frequencies of F1 for Females*




## V. Conclusion

In this paper, the weakness of the LPC stems from two major sources: it partially eliminates some of the critical features of the original speech in the process of removing speech redundancies, and it notoriously lacks robustness. Therefore the improvement of LPC speech is directly related to the appropriate use of speech preprocessing, proper selection of speech samples for analysis, and suppression of various forms of interference. Hence we use the short-term spectral analysis method to get the original speech frequencies. The formant frequencies of each alphabet of speech are determined very accurately using short-term spectral analysis.

## Acknowledgment

Expression of giving thanks is just a part of those feelings which are too large for words but shall remain as memories of beautiful people with whom I have got the pleasure of working during the completion of this work. I am grateful to RCET BHILAI, Chhattisgarh, which helped me to complete my work by giving an encouraging environment. I want to express my deep and sincere gratitude to the HOD of Digital Electronics department of RCET, Bhilai. His comprehensive knowledge and his logical way of thinking have been of great value to me. His understanding, encouraging, and personal guidance has provided a sound basis for the present work.

## References

[1] S.Furui, "Digital Speech Processing, Synthesis, and Recognition" Marcel Dekker, 1989.
[2] L. Rabiner and B-H. Juang, "Fundamentals of Speech Recognition", Englewood Cliffs: Prentice Hall, 1993.
[3] Joseph W. Picone, "Signal Modeling Techniques in speech recognition," Proceedings of the IEEE, vol.81, no.9, pp.1215-1247, 1993.
[4] M. Bellanger: "Digital Processing of Signals: Theory and Practice", John Wiley & Sons, 1989.
[5] D.A. Reynolds, L.P. Heck, "Automatic Speaker Recognition", AAAS 2000 Meeting, Humans, Computers and Speech Symposium, 19 Feb 2000.

[6] J. P. Campbell, JR., "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol. 85, no.9, pp. 1437-1462, Sep 1997.
[7] D. Schwarz, "Spectral Envelopes in Sound Analysis and Synthesis", IRCAM Institut de la Recherche et Coordination Acoustique/Musique, Sep 1998.
[8] J. R. Deller, J. H.L. Hansen, J. G. Proakis, "Discrete-Time Processing of Speech Signals", IEEE Press, New York, NY, 2000.
[9] J. G. Proakis, D. G. Manolakis, "Digital signal processing. Principles, Algorithms and Applications", Third ed. Macmillan, New York, 1996.
[10] Samuel Kim, Thomas Eriksson, Hong-Goo Kang, Dae Hee Yuon, "A Pitch Synchronus Feature Extraction Method for Speaker Recognition", 2004
[11] P. Alexandre, P. Lockwood, "Root Cepstral Analysis: A Unified View", 1993
[12] Umit Yapanel, John H.L. Hansen, Ruhi Sarikaya, Bryan Pellom, "Robust Digit Recognition in Noise: An Evaluation Using the AURORA Corpus", 2001
[13] A M Ariyaeeinia, P Sivakumaran, "Comparison of VQ and DTW Classifiers for Speaker Verification", 1997
[14] Eamonn J. Keogh, Michael J. Pazzani, "Derivative Dynamic Time Warping", 2001
[15] Elena Rodriguez, Belén Rtdz, Juagel Garcia-Crespo, Femando Garcia, "Speech/Speaker Recognition Using a HMM/GMMHybrid Model", 1997
[16] Johan Olsson, "Text Dependent SpeakerVerification with a HybridHMM/ANN System", 2002
[17] Shai Fine, Jiři Navrátil, Ramesh A. Gopinath, "A Hybrid GMM/SVM Approach to Speaker Identification", 2001
[18] Stefan Schacht, Jacques Koreman, Christoph Lauer, Andrew Morris, Dalei Wu, Deitrich Klakow, "Fram Based Features", 2007
[19] Abdulnasir Hossen, Said Al-Rawahi, "A Text-Independent Speaker Identification System Based on the Zak Transform", 2010.
[20] Mostafa Hydari, Mohammad Reza Karami, Ehsan Nadernejad, "Speech Signals Enhancement Using LPC Analysis based on Inverse Fourier Methods", Contemporary Engineering Sciences, Vol. 2, 2009, no. 1, 1 - 15
[21] Hyunsin Park, Tetsuya Takiguchi, and Yasuo Ariki, Research Article Integrated Phoneme SubspaceMethod for Speech Feature Extraction, Hindawi Publishing Corporation EURASIP Journal on Audio, Speech, and Music Processing Volume 2009
[22] Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson, IEEE Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model IEEE Transations on audio, speech and language processing, VOL. 16, 2008.
[23] Ishizuka K.& Nakatani T.: A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. Speech Communication, Vol. 48, Issue 11, pp. 1447-1457, 2006
[24] K.R. Aida–Zade, C. Ardil and S.S. Rustamov, Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems, Proceedings of world academy of science , engineering and technology, Volume13, ISSN 1307-6884, 2006
[25] Burak Tombaloğlu, Hamit Erdem,"Development of a MFCC-SVM Based Turkish Speech Recognition system", IEEE 24th Conference on Signal Processing and Communication Application (SIU),ISBN: 978-1-5090-1679-2, 2016.
[26] Usha Sharma,Sushila Maheshkar, A. N. Mishra"Study of robust feature extraction techniques for speech recognition system", IEEE International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), ISBN: 978-1-4799-8433-6, 2015
[27] Jun Tao, Xiaoxiao Jiang, "A Domestic Speech Recognition Based on Hidden Markov Model", IEEE CCIS, pp.606-609, 2011.
[28] Mansour Alsulaiman, Ghulam Muhammad, Zulfiqar Ali, "Comparison of Voice Features for Arabic Speech Recognition", IEEE, pp.90-95, 2011.
[29] Eleonora D'Arca, Neil M. Robertson and James Hopgood, "Using the Voice Spectrum for Improved Tracking of People in a Joint Audio-Video Scheme",IEEE, 2013.
[30] Joanna Grzybowska, Maciej Klaczynski,"Computer-assisted HFCC-based learning system for people with speech sound disorders", IEEE, pp.1-5, 2014.
[31] Peng Dai, Ing Yann Soon, "An Adaptive Soft Voice Activity Detector for Automatic Speech Recognition System", IEEE, pp.1-5, 2011.