

Speech Training System based on Resonant Frequencies of Vocal Tract

V. Salai Selvam¹, V. Thulasibai² and R. Rohini³

Department of Electronics and Communication Engineering, Sriram Engineering College, Perumalpattu, Tamil Nadu, India
vsalaiselvam@yahoo.com, thulasi_bai@yahoo.com, rohini2002@gmail.com

Abstract—Speech sounds are air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the vocal tract as it is pushed out through the lips and nose. Speech signals, in addition to communicating the linguistic information, convey a multitude of other information including gender, age, accent, intent, emotion, humor and the state of health of the speaker. There are several neurological (e.g., aphasia, dysarthria, and apraxia) or anatomical (e.g., cleft lip and palate) factors that could affect the intelligibility and audibility of the human speech. In this paper we present our work in developing a training system based on the resonant frequencies (widely known as formants) of the vocal tract to help a human subject with speech impairment train himself or herself to improve the intelligibility and audibility of his or her speech.

Keywords—formants, impaired speech, speech training system, vocal tract resonance, cepstrum

I. INTRODUCTION

A. Speech Production

Speech sounds are air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the vocal tract as it is pushed out through the lips and nose. Figure 1 shows the anatomical model of the speech mechanism.

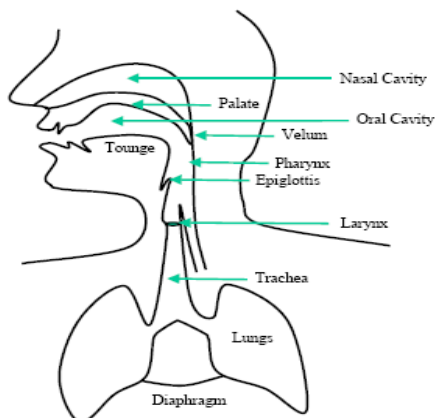


Figure 1. Anatomy of human speech mechanism

The movable parts present inside the oral cavity are called the articulators and their movement is known as the articulation. The articulators in the oral cavity include velum, palate, tongue, teeth, and lips and, in general, the lower jaw [5] [7]. The relative positions of these articulators decide the speech sound produced and its intelligibility. Such ‘intellectual’ speech sounds produced are broadly classified into two: *voiced* sounds and *unvoiced* sounds [5] [7]. Speech signals convey not only the linguistic information but also the gender, age, accent, intent, emotion, humor and the state of health of the speaker [3]-[5] [7].

B. Speech Signal Properties

The Spectral properties of the Speech signal vary with time as the vocal tract shape varies. However, speech can be considered as quasi-stationary over short segments, typically 10-30 milliseconds. The statistical and spectral properties of speech are thus defined over short segments. The voiced signals are quasi-periodic and exhibit spectra with frequency spacing equal to the pitch frequency [3]-[5] [25]. The unvoiced signals exhibit rather constant spectra. The effective bandwidth of speech is approximately 7 kHz [4] [5] [25].

C. Speech Units

The elementary linguistic unit of spoken speech is called a *phoneme* and its acoustic realization is called a *phone* [3]-[5] [7] [17]-[21] [23]. The British English phonemes are grouped into ten categories [5]. The US English can be similarly classified [5]. Table 1 and 2 shows the British English phonemes and the US English phonemes respectively [5].

TABLE 1. A LIST OF AMERICAN ENGLISH PHONEMES

Phoneme	Example	Phoneme	Example	Phoneme	Example	Phoneme	Example
IV	best	IX	roses	NX	sing	V	val
IH	bit	ER	bird	P	pet	DH	that
EY	bait	AXR	butter	T	ten	Z	zoo
EH	bet	AW	down	K	kit	ZH	azure
AE	bat	AY	buy	B	bet	CH	church
AA	bat	OY	boy	D	debt	JH	judge
AH	but	Y	you	H	get	WH	which
AO	bought	W	wit	HH	hat	EL	battle
OW	boat	R	rent	F	fat	EM	bottom
UH	book	L	let	TH	thing	EN	bitten
UW	boot	M	met	S	sat	DX	batter
AX	about	N	net	SH	shut	O	(glottal stop)

TABLE 2. A LIST OF BRITISH ENGLISH PHONEMES

Vowels	Semivowel	Nasal	Vfricative	Whisper
aa	l	m	v	hh
ae	r	n	dh	sil
ah	w	ng	z	
ao	y	Affricative	zh	
ax	Diphthong	ch	UnVstop	
eh	aw	jh	p	
er	ay	UnVfricative	t	
ih	ea	f	k	
iy	ey	s	Vstop	
oh	ia	sh	b	
ow	oy	th	d	
uh	ua		g	
uw				

D. Impaired Speech

Speech or language impairment involves a communication disorder with the rhythm articulation, vocal qualities or the components of language. Generally the speech or language impairment is classified into [1] [2] [25]: *Articulation Impairment*, which includes Omissions, substitutions or, distortions of sounds, *Voice Impairment*, which includes inappropriate pitch, loudness, or voice quality, *Fluency Impairment*, which includes abnormal rate of speaking, Speech Interruptions and repetition of words, sounds, phrases or sentences which interferes with effective communication, and *Language Impairments* which is Phonological, Morphological, Syntactic, Semantic, or Pragmatic use of oral language. Articulation impairment is the cause of concern for us in this research.

E. Speech Parameters [3]-[5] [7]

Speech parameters can be listed as follows: (i) spectral envelope (ii) the speech formants and (iii) the pitch. Formants refer to pre-dominant peaks in the harmonic spectrum of a single sound or phoneme, produced by the resonance of the vocal tract of shape peculiar to that sound [3]-[5] [7]. Because of their resonant origin, they tend to stay essentially the same when the frequency of the fundamental or pitch is changed. The ranges of frequencies is shown in Table IV for the first three formants of a male voice [5] [22] [24]. It has been recognized that F_1 is influenced by the height of the tongue and F_2 is influenced by the frontness or backness of the tongue body [5] [7] [24].

II. MATERIALS AND METHODS

A. Materials

The impaired speech from 52 adults with the articulation problem was recorded using a P-IV computer with 2 GB RAM, 160 GB HDD, PC-based microphone, 16-bit sound card, and free audio recording and editing software at a sampling rate of 8 KHz. The subjects were requested to pronounce the words listed in Table 3 through the microphone. As seen from Table 3, there are sixteen words among which are distributed 43 phonemes.

B. Method

As the set of first three (or five) formants decide the sound unit being produced, one with articulation problem can trained to improve the speech if the formants for the phonemes of the impaired speech is compared with the 'templates' of the formants for the same set of phonemes

derived from the normal speech. This can help 'shy' subjects to train their articulations 'off' the line i.e., on their own without any trainer using a simple desktop computer and this Formant Based Speech Training System (FBSTS), which is a computer based program. The FBSTS is shown in Figure 2.

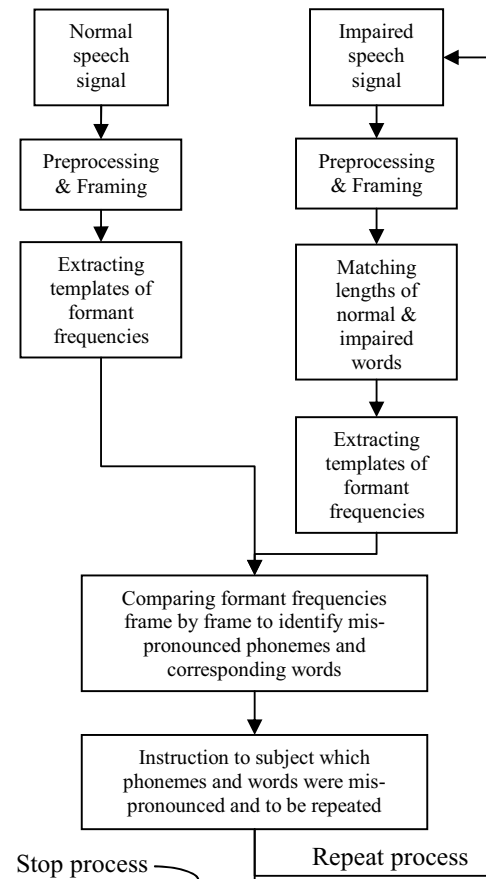


Figure 2. FBSTS

TABLE 3. LIST OF WORDS USED AND PHONEMES COVERED

S. No.	Words	Phonemes
1	ZERO	Z-I-R-OW
2	USHER	AH-SH-AX
3	ANY	E-N-I
4	SEE	S-I-
5	FUR	F-AA
6	HANG	H-AE-NX
7	ARM	AA-M
8	THAW	TH-AO
9	GOT	G-AA-T
10	YOU	Y-U-
11	WOOD	W-u-D
12	BOUQUET	B-U-K-AY
13	JOIN	JH-OY-N
14	THERE	DH-EH-AX
15	LIP	L-i-P
16	I	-AY-

C. Extraction of Formants

The formants or resonant frequencies of the vocal tract can be extracted from the speech signal either using the LP

technique [11] [12] [13] [14] or the cepstral technique [8] [9] [10]. In this research work, the cepstral technique was used to extract the formant frequencies.

The cepstral technique is a well-known process of separating convolved signals such as speech signal which is a convolution of the voicing signal (also known as the excitation signal, source signal speech residue) and the vocal

tract response [3]-[6] [7] [9] [10] [12] [25]. The (complex) cepstrum, $s_p'(n)$ of a signal, $s(n)$ is defined as the inverse Fourier transform of the logarithm of the Fourier transform of $x(n)$ as shown in Figure 3 (a) and (b).

Computation of complex cepstrum requires phase unwrapping, which is difficult due to some theoretical and practical reasons [8].

TABLE 4. RANGE OF FORMANT FREQUENCIES FOR MALE VOICE

First Formant Frequency (F_1)	150-850 Hz
Second Formant Frequency (F_2)	500-2500 Hz
Third Formant Frequency (F_3)	1500-3500 Hz

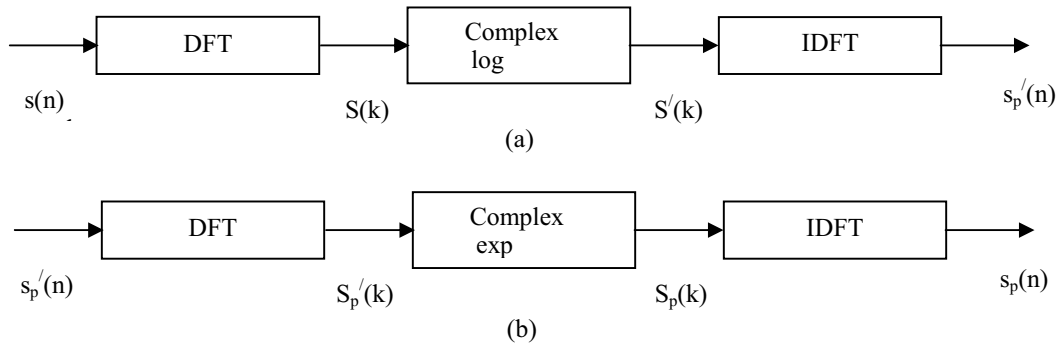


Figure 3. (a) Complex cepstrum and (b) its inverse (after Oppenheim & Schafer).

The real cepstrum $s_r'(n)$ of a signal $s(n)$ is defined as the inverse Fourier transform of the logarithm of the signal magnitude spectrum [8]. Real cepstrum is not invertible but provides a minimum phase reconstruction of the signal [8].

This is depicted in Figure 4 (a) and (b). Cepstral domain is known as **quefrequency** (coined from 'frequency') domain [8] [25].

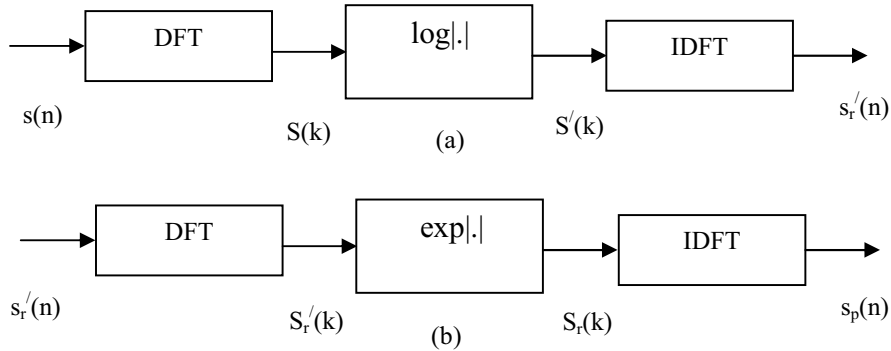


Figure 4. (a) Real cepstrum and (b) its inverse (after Oppenheim & Schafer).

Since speech is convolution of a fast varying glottal excitation signal, $e(n)$ and a slowly varying vocal tract response, $h(n)$, the cepstrum of a speech segment consists of the glottal excitation signal that occupies the low quefrequency region and the vocal tract response that occupies the high quefrequency region [8] [25]. Since the phase information is not as much important as the magnitude information in a speech spectrum, the real cepstrum is used due to its computational easiness [8] [25]. The formant frequencies are extracted as peaks of the smoothed log spectrum obtained from the cepstral coefficients of the vocal tract.

The steps involved in creating the formant templates for the words listed in Table 3 are as follows:

- 1) The words listed in Table 3 are recorded from 100 normal subjects.
- 2) The recorded signals are then lowpass-filtered to 4 kHz and the (arithmetic) mean is removed from each signal.
- 3) Then from the mean removed signals, the (normal) speech segments (utterances) alone are extracted from the background noise i.e., the leading and trailing non-speech segments (non-utterances) are removed using a simple end-point detection algorithm based on the magnitude and the zero-crossing rate [13].
- 4) The extracted similar (normal) utterances are then time-scaled to be length-matched with each other using the phase-vocoder technique [15].

5) The extracted (normal) utterance is then split into frames of 10 ms duration with successive frames overlapping by 5 ms. (Shorter the frame, then longer the process time but more phonemes are covered.)

6) Then the frame is windowed using the Hamming window to avoid the spectral leakage [11] [16].

7) The windowed frame is then pre-emphasized to cancel the contributions of the larynx and lips [25] using the filter $H(z)=1-z^{-1}$.

8) Then the real cepstrum of the pre-emphasized frame is computed using the Fast Fourier Transform (FFT) as the inverse Discrete Fourier Transform (DFT) of the logarithm of the magnitude of the DFT of the frame [8] [9].

9) The first 40 samples of the cepstrum of the frame, which represents the (quefrequency-domain) vocal tract response, is obtained by the simple liftering process [8] [9].

10) The (time-domain) vocal tract response is then obtained by taking the inverse DFT of the exponentiation of the magnitude of the DFT of the cepstrum of the (quefrequency-domain) vocal tract response.

11) The smoothed log spectrum is then obtained as the log of the absolute value (magnitude) of the DFT of the (time-domain) vocal tract response.

12) The first three dominant peaks from this smoothed log spectrum is then measured using a simple peak-picking algorithm. These are stored.

13) Steps from 6) to 12) are repeated for all frames of the same (normal) utterance.

14) Steps 5) to 13) are repeated for the set of the same 100 (normal) utterances. The set of formant templates for the set of words listed in Table 3 is formed by averaging the set of formants of similar frames of the same (normal) utterances.

The steps involved in the Formant Based Speech Training System for the subjects with the articulation impairment to improve their articulation are as follows:

15) The words listed in Table 3 are recorded from the subject with the articulation impairment.

16) Then the steps 2) to 12) are followed to extract the set of formants for the utterance.

17) Then the Euclidean distance (a similarity measure) between the set of formants for the impaired utterance with the set of formants for the same but normal utterance is measured. The frames of the impaired utterance for which this similarity measure is above a (fixed) threshold are considered to be the 'dissimilar' frames and the corresponding phonemes are considered to have been mispronounced by the subject.

18) Then the subject is instructed to repeat those words with mis-pronounced phonemes giving more emphasis to the mis-pronounced phonemes.

III. RESULT AND DISCUSSION

The output of each step of the entire speech training system just discussed has been presented from Figure 5 (a) to (j) for a clear understanding of the steps involved in the process. To create a set of formant templates, a unique list of words with maximum possible total number of phonemes with very little or no repetition of phonemes was finalized. This enabled us to train several subjects with different articulation problems without changing the system. As the computation involved is very simple requiring only few resources of the desktop PC used, it can be very much implemented in real-time if a dedicated DSP system is used. One major drawback is that the system fails to help people with articulation problem resulting in mispronunciation of unvoiced sounds like "sh" due to the very fact that the smoothed log spectra of unvoiced sounds do not show prominent peaks.

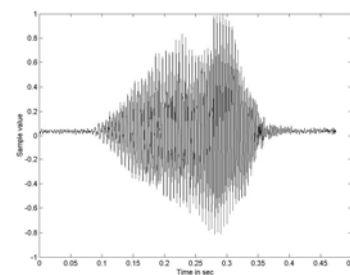


Figure 5. (a) Speech signal (utterance plus noise) "a"

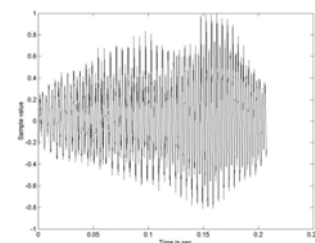


Figure 5. (b) Utterance "a" extracted from background noise

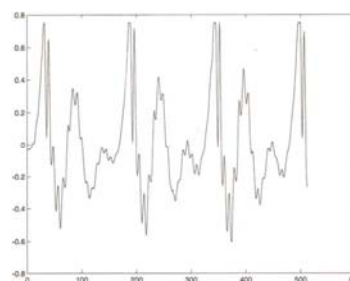


Figure 5. (c) One mean-removed frame of utterance "a"

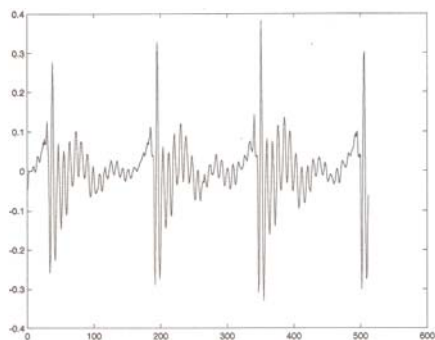


Figure 5. (d) Pre-emphasized frame of utterance "a"

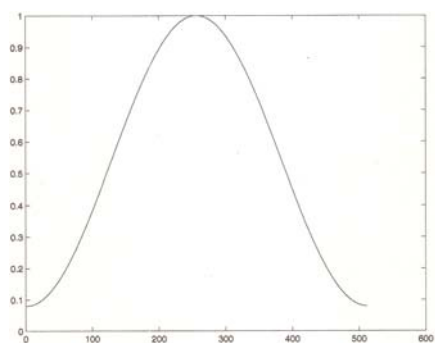


Figure 5. (e) Hamming window

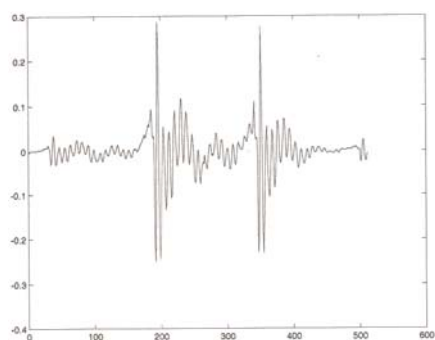


Figure 5. (f) Hamming-windowed frame of utterance "a"

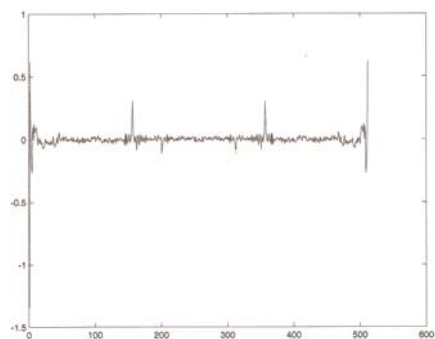


Figure 5. (g) Real cepstrum of frame of utterance "a"

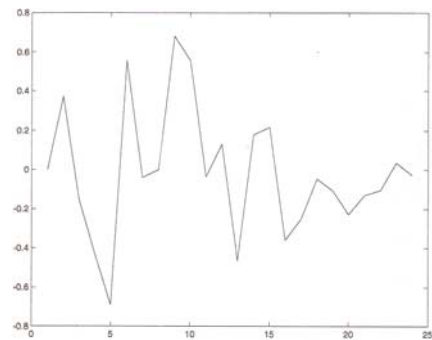


Figure 5. (h) Cepstrum of vocal tract response obtained by liftering

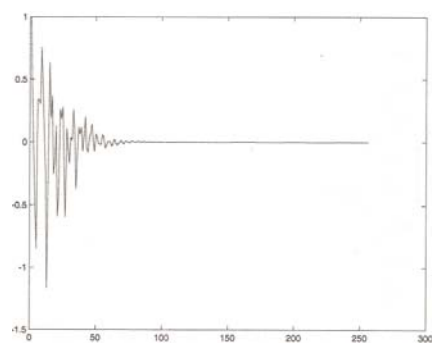


Figure 5. (i) Vocal tract response

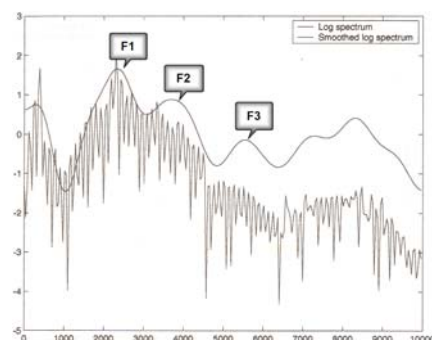


Figure 5. (j) Smoothed log spectrum and formant 'peaks' F_1 , F_2 & F_3

Another problem which we found was the difficulty in training even normal subjects in standard foreign (British) accent. We had to standardize the accent with the normal Indian accent.

However, manually training the subjects to pronounce more similarly and then re-recording the words then, reduced these differences. Once the differences became negligible and regular, a threshold was set for the deviations in the formants to judge the mispronunciation of words.

This threshold was then adjusted to make a fairly accurate judgment, which was not too sensitive, at the same time not less than required sensitivity to the changes in the vocal tract. This was done on a trial and error basis. An average was found from the formants of the different

speakers. This was then to be used as a reference for comparison with people problem in their articulation.

IV. FUTURE DEVELOPMENT

The reliability of the FBSTS can be improved by using a more reliable identity for the unvoiced sounds such as “hiss” and “hush”, which lack proper resonant peaks.

The real time processing can be improved to an extent that a dictionary can be included and the words can be recognized automatically with the aid of a set of data of all phonemes for identifying the phonemes. The phonemes not having a match with the reference set can be identified as an error and suitable correction may be recommended.

But implementing all this will load conventional computer systems to an extent that on line processing becomes impractical. This situation can be overcome by implementing the project in DSP Processors.

Matching of similar words spoken by different speakers can be improved by adopting a more accurate technique of Time Warping.

ACKNOWLEDGMENT

The authors would like to thank Late Shri. M. E. Deverajan, Former Chairman, Sriram Educational Trust, Chennai, Tamil Nadu, India, for his consistent encouragement during the project period. They thank the present members of the management for their extended support in all our developmental activities. They also thank the Principal, Sriram Engineering College, Chennai, Tamil Nadu, India, for his support and guidance in writing this paper

REFERENCES

- [1] NICHCY Disability Fact Sheet No. 11 (2004), ‘Speech & Language Impairments’, NICHCY.
- [2] Department of Education (2002), ‘Special Education Programs and Services Guide’, State of Michigan State.
- [3] Shuzo Saito and Kazuo Nakata (1985), ‘Fundamental of Speech Signal Processing’, Academic Press, London.
- [4] Rabiner L. R. and Schafer R. W. (1978), ‘Digital Processing of Speech Signal’, Prentice-Hall, Engliwood Cliffs, NJ.
- [5] Rabiner L. R. and Juang B. H. (1993), ‘Fundamentals of speech recognition’, Prentice-Hall, Engliwood Cliffs, NJ.
- [6] Rabiner L. R. and Bernard Gold (1992) ‘Theory and Application of Digital Signal Processing’, Prentice-Hall of India, New Delhi, Chapter 12.
- [7] Thomas F. Quatieri (2004), ‘Discrete-Time Speech Signal Processing’, Pearson Education, Singapore.
- [8] Oppenheim A. V. and Schafer R. W. (1992), ‘Discrete-Time Signal Processing’, Prentice-Hall of India, New Delhi.
- [9] Oppenheim A. V. (1969), ‘Speech Analysis-Synthesis based on Homomorphic Filtering’, Journal of Acoustic Society of America, Vol.45, pp. 458-465.
- [10] Oppenheim A. V. (1976), ‘Signal analysis by homomorphic prediction’, Proc. IEEE, ASSP, Vol. 24, No. 4, pp. 327.
- [11] Proakis J. G. and Manolakis D. G. (2000), ‘Digital Signal Processing’, Prentice-Hall of India, New Delhi.
- [12] Tony Robinson (1998), ‘Speech Analysis’ Lent Term.
- [13] Nipul Bharani, Sara MacAlpine, JP Slavinsky and Aamir Virani (2000), ‘A Project on Speaker Recognition’ Rice University.
- [14] Makhoul J. (1975), ‘Linear Prediction: A Tutorial Review’, Proc. IEEE, Vol. 63, No. 4, pp. 561-580.
- [15] Jean Laroche and Mark Dolson (1999), ‘New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects’, Proc. IEEE WASPAA, Oct. 17-20.
- [16] Murat Kunt (1986), ‘Digital Signal Processing’, ARTECH HOUSE, Norwood.
- [17] Lynne E. Bernstein, Moise H. Goldstein, Jr., and James J. Mahshie (1988), ‘Speech training aids for hearing-impaired individuals: I. Overview and aims’, Journal of Rehabilitation Research and Development Vol. 25 No. 4 Pages 53—62.
- [18] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, Robert C. Armstrong and David Pisoni (1987), ‘From Text To Speech: The MITalk System (Cambridge Studies in Speech Science and Communication)’, Cambridge University Press, Cambridge, England.
- [19] Akira Watanabe, Shingo Tomishige, and Masahiro Nakatake (2000), ‘Speech Visualization by Integrating Features for the Hearing Impaired’, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol.8, No.4, July 2000, p.454-466.
- [20] Gerasimos Potamianos and Chalapathy Neti (2001), ‘Automatic Speech Reading of Impaired Speech’, Proceedings of the Audio-Visual Speech Processing Workshop, Scheelsminde, Denmark, September 7-9, 2001.
- [21] Klatt and Dennis (1987), ‘Review of text-to-speech conversion for English’, The Journal of the Acoustical Society of America 82 (3): 737-793.
- [22] Mark Hasegawa-Johnson (2000), ‘Lecture Notes in Speech Production, Speech Coding, and Speech Recognition’, University of Illinois, Urbana-Champaign.
- [23] Furui S. (1989), ‘Digital Speech Processing, Synthesis and Recognition’, Marcel Dekker.
- [24] W. Tecumseh Fitch (1997), ‘Vocal Tract Length and Formant Frequency Dispersion Correlate with Body Size in Rhesus Macaques’, Journal of Acoustic Society of America, p.1213-1222.
- [25] G. Ravindran, S. Shenbagadevi and V. Salai Selvam (2010), ‘Cepstral and Linear Prediction Techniques for Improving Intelligibility and Audibility of Impaired Speech’, J. Biomedical Science and Engineering, 2010, 3, 85-94.