ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

Effect of Voice Part on the Quality of Children Speech in Dogri Language

Varun Sharma[1], Padmini Rajput[2], Randhir Singh[3], Parveen Lehana[4]

M.Tech Student, Department of Electronics and Communication, Sri Sai College of Engineering, Punjab, India[1]

Associate Professor, Department of Electronics and Communication, Sri Sai College of Engineering, Punjab, India[3]

PhD. Scholar, Department of Physics & Electronics, University of Jammu, Jammu, India[2]

Associate Professor, Department of Physics & Electronics, University of Jammu, Jammu, India[4]

ABSTRACT: Speech is the most innate and fastest means of communication between humans. Computers with the ability to understand speech and speak with a human like voice are expected to contribute to the development of more natural man-machine interface. For the analysis of speech signal we have carried out the recording of six children speakers (3 male and 3 female) in Dogri language between the age group of 3-6 years. Harmonic plus noise model HNM has been employed as the analysis-synthesis platform as it outperforms almost all models of speech production in terms of important characteristics like naturalness, intelligibility, and pleasantness. PESQ method is used for evaluation of the quality of the speech synthesized from HNM. Mean and standard deviation (SD) is estimated for original and synthesized speech. Effect of different proportion of voice part on the quality and intelligibility of speech signal of children has been investigated at different levels of noise keeping noise part constant. Results suggest that the quality is quite poor at lower levels of voice part but increases gradually until the value of voice part is 50%. However as the voice percentage is increased the quality remains constant afterwards (till v100%). Results suggest that the percentage of voice part plays an important part for the quality of speech. With no voice part the quality is quite poor. Further the results prove that HNM is an excellent model for children speech. Also the worst and best speech quality is not same for male and female children speakers.

KEYWORDS: Speech processing, HNM, PESQ.

I. INTRODUCTION

Speech generation is one of the most important areas of research in speech signal processing which is now gaining a serious attention. The attributes of speech signal are so fascinating that we rarely pause to define it. Speech is the most natural kind of communication different forms of information to the listener. Among them, the content of the message is most important nevertheless, other information like the emotion, gender and identity of speaker is also an essential part in the oral swap over of communication [1]. Figure 1 shows the different parts of human articulatory system. The speech signal is generated from human articulatory system and perceived through ears. Speech is a natural form of communication in human beings and seems as natural to humans as walking, and only less so than breathing [2]. Brain arranges thoughts into sequence of words for articulation. The indistinguishable units that constitute words are called phonemes. The pronunciation of phonemes depends upon contextual effects, speaker characteristics and emotions [3]. Human speech is dynamic rather than static, as the articulators keep moving during articulation this fact leads to an assumption that we begin to articulate the next segment before completing the previous one [4]. Speech signal processing has many efficient and intelligent applications, like speech recognition, speaker transformation and text-to-speech (TTS) systems [6].

DOI: 10.15680/IJIRSET.2015.0402067

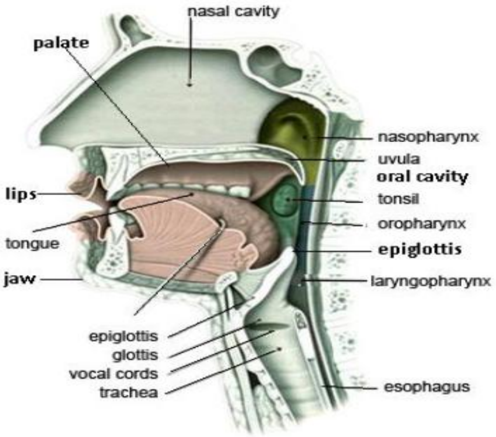

Fig. 1 Anatomy of the human speech production system.

II. RELATED WORK

Research on Indian languages has been used for developing Text-to-Speech synthesis systems for few Indian languages like Hindi, Tamil, Kannad, Marathi, and Bangla. Speech is an information rich signal exploiting frequency modulated, amplitude modulated and time modulated carriers such as resonance, harmonics and noise, pitch intonation, power, and duration to convey information [7]. Information in speech signals is primarily occupied in 4 KHz telephone bandwidth and above that speech energy conveys quality and sensation [6]. Figure 1 shows the outline anatomy of the human speech production system. It consists of the lungs, larynx, vocal tract cavity, nasal cavity, teeth, lips, and the connecting tubes. The process of speech production begins with exhaling air from the lungs. This air sounds like a random noise with no information, if exhaled without subsequent modulation. The air is frequency modulated by closing and opening of the glottal folds. This signal is applied to the vocal tract which is further shaped by the resonance of the effects of the nasal cavities and the teeth and lips. Speech will sound like a random noise with no information if the air is exhaled without modulation. The frequency of closing and opening of glottal fold determines the type of information in speech signal. These signals are the passed to vocal tract as excitation signal which shapes the resonance of the vocal tract and the effects of the nasal cavities, teeth, and lips [8-10]. A lot of studies have been carried out on adult vocal tract, but only few on the children vocal tract. Since the infants vocal tract is a miniature version of adults. Therefore vocal tract is expected to grow in different manner and different timing**.** Effect of the proportion of harmonic and noise part on the quality of synthesized speech using HNM in hindi language reveals that with 50% voice part the optimum noise percentage for acceptable speech quality is found to be 40%. As the percentage of voice part is increased beyond 50%, the speech quality doesn’t degrade even if the noise part is increased. Further, the optimum percentage of the noise part for good speech quality has been found speaker and phoneme dependent as well [11]. Hindi is one of the prominent languages in India and belongs to Indo- European language of the indo Aryan subfamily while dogri is spoken in the regions like Jammu, parts of Kashmir, Himachal, and northern Punjab. Dogri was given the honor of the national language on 22nd December, 2003. Dogri was given the honor of national language on 22nd December, 2003. The objective of this paper is to determine the effect of varying the percentage of voice parts on the perception of speech of children in Dogri language synthesized by harmonic plus noise (HNM) model. The brief detail on the synthesis model of HNM has been discussed in section 3. Tools and techniques have been described in section43, Results and discussions are presented in section 5 and the conclusions in section 6.

III. HARMONIC PLUS NOISE MODEL

HNM which was developed by Stylianou et al. [9] and it is used for high quality time/pitch scale modification of speech and voice transformation. Techniques developed for the synthesis of speech like Hidden Markov Models (HMM), Mel frequency Cepstral Coefficients (MFCCs), Line Spectral Pairs (LSPs), and Harmonics plus Noise Model (HNM) are widely used to model spectra in many synthesis and conversion systems.

For HNM synthesis as shown in fig. 2, the center-of-gravity is used for eliminating the inter-unit phase mismatches. Prosody (fundamental frequency F0, duration, and amplitude) may be altered as desired. Around unit concatenation points, we smooth the HNM parameters in order to minimize residua discontinuities (after unit selection) by employing a simple linear interpolation over a small number of frames. The actual synthesis is done following the overlap-and add paradigm. For each frame, the noise part is high-pass filtered according to the maximum voiced frequency found during analysis (that is zero for unvoiced speech). Also, the noise part is modulated by a parametric triangular envelope synchronized in time with the pitch period [12]. The harmonic part and noise part constitute the quasi-periodic components and non-periodic part respectively [12].HNM decomposes speech into two components: a harmonic component and a noise component. HNM expresses the sampled speech signal s[n] as the sum of two components: sh(t)and, sn(t) which correspond to the harmonic and noise, or stochastic, components of the signal, respectively.

$$\hat{s} = s_h(t) + s_n(t) \tag{1}$$

where

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{jk\omega_0(t)} \tag{2}$$

and

$$s_n(t) = e(t)[h(\tau, t) * b(t)] \tag{3}$$

The frequency that separates the two bands is called maximum voiced frequency Fm.. The lower band represents the signal by harmonic sine waves, slowly varying in amplitudes and frequencies:

$$s'(t) = \mathrm{Re} \sum_{l=0}^{L(t)} a_l(t) \exp\{ j[\int_0^t l\omega_o(\sigma) d\sigma + \theta_l]\} \tag{4}$$

where a_l *(t)* and Θ_l *(t)* represent the amplitude and phase at time t of the l^{th} harmonic, while w_o*l(t)* are the fundamental and time-varying number of harmonics included in the harmonic part. AR model represents the upper band constituting the noise part modulated by the time domain amplitude envelope. The noise part *n '(t)* is obtained by filtering a white Gaussian noise *b(t)* by a time varying, normalized all-pole filter *h (τ: t)* . The result obtained is multiplied by an energy envelope function w (t):

n '(t)=w(t) [h(τ; t)*b(t)] (5)

In addition to obtaining the maximum voiced frequency F_m, other parameters like voiced/unvoiced, amplitudes and phase of harmonics of fundamental frequency (pitch), glottal closure instants, parameters of noise part, and pitch are calculated for each frame. Figure 2 depicts the analysis using HNM. Speech signal is fed by the voicing detector which states the frame either voiced or unvoiced. HNM analysis is pitch synchronous so their lies the exact inference of the glottal closure instances (GCIs) [12]. GCIs can be calculated either by means of the speech signal or electroglottogram (EGG). Speech signal or EGG is given at the input side of GCI. Maximum voiced frequency F m is calculated for each voiced frame. The analysis frame is taken twice the local pitch period. Form each GCI the voiced part is analyzed for calculating amplitudes and phase of all the pitch harmonics up to F m. The synthesized portion of the voice part is calculated from equation (4) for obtaining noise parameters while the remaining fraction obtained as result of the subtraction of the noise from the speech signal is the voiced part. Noise part is later analyzed for the LPC coefficients and energy envelope. For both voiced and unvoiced frames the length of the analysis window for noise part is taken as two local pitch periods. However for unvoiced frames the local pitch is the pitch of the last frame and for voiced frames the local pitch is the pitch of the frame itself and [12]. The addition of the synthesized speech.HNM based synthesis can be used for good quality output with relatively small number of parameters. Using HNM, pitch and time scaling are also possible without explicit estimation of vocal tract parameters [12]. Speaker transformation and voice conversion method has been a hot area of research in speech processing research for the last two decades [13-16]. These techniques are also implemented in the framework of the HNM system, which allows the high-quality modifications of speech signals. In comparison to earlier methods based on the vector quantization, HNM based conversion scheme results in high quality modification of speech signal [15].

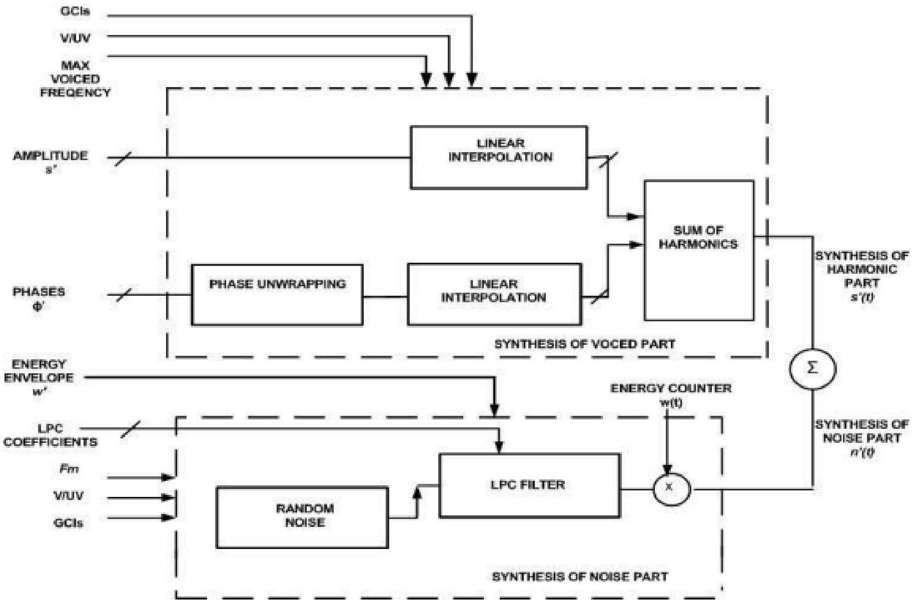

Fig.2 Synthesis of speech using HNM [12]

Harmonics are represented by the lower band and modulated noise is represented by upper band. These validations are useful from perception point of view which leads to simple speech model, providing high quality synthesis and modification of the speech signals [16-20]. Research has shown that all vowels and syllables can be produced with a better quality syllables by the implementation of HNM [21]. Results obtained from many speech signals including both male and female voices are quite satisfactory with respect to the background noise and inaccuracies in the pitch [22].

IV. METHODOLOGY

The research work is divided into two major parts. Figure 3 shows the block diagram of the research methodology.
In first part speaker selection, speech recording and segmentation is done, while in the second part analysis-synthesis of speech has been performed by using HNM model and objective evaluation of speech quality has been estimated by perceptual evaluation of speech quality (PESQ). Eight different phrases in Dogri language are recorded using Goldwave software at the sampling rate of 16,000 KHz. The material was recorded in an acoustically treated environment and segmented and labeled manually using Praat software. HNM based speech synthesis of Dogri language is carried out in this research work taking six speakers in the age group of 3-6 years using HNM algorithm. The deviation between the original and HNM synthesized speech were analyzed. Second aspect of the investigation is to estimate the effect of voice percentage on children speech signal in Dogri language. Perceptual Evaluation of Speech Quality (PESQ) one of the methods for objective has been used for the comparison of the original and synthesized speech quality. It predicts subjective MOS score by comparing the synthesized speech with original version of the speech signal [23].

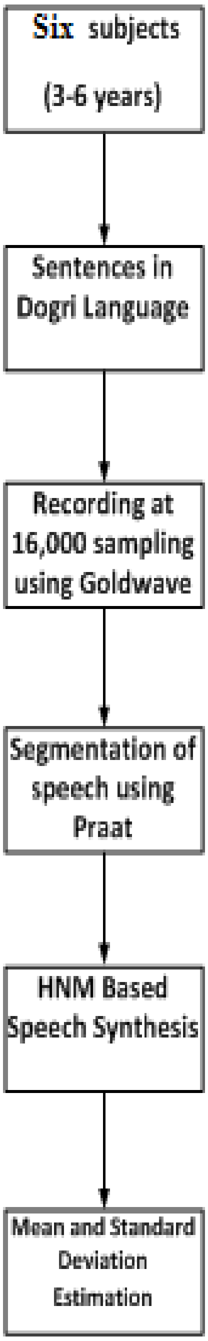
Six subjects
(3-6 years)
Sentences in Dogri Language
Recording at 16,000 sampling using Goldwave
Segmentation of speech using Praat
HNM Based Speech Synthesis
Mean and Standard Deviation Estimation

Fig. 3. Block diagram representation of proposed methodology

V. RESULTS AND DISCUSSIONS

The histograms in figure 4 show the quality of speech signal at different percentage of voice and noise level (v1-10% at n100%) for all the six speakers. The horizontal axis shows the percentage of voice and noise part e.g. v1 indicates .01% voice part) and the vertical axis represents its corresponding PESQ score (speech quality). sp1, sp3, and sp5 correspond to male speakers while female speakers are labelled as sp2, sp4, and sp6. From the histograms as shown in figure 4 the

DOI: 10.15680/IJIRSET.2015.0402067

range of voice part from v1-v10%, it can be analyzed that the quality of the speech signal increases substantially for all speakers (sp1-sp6).

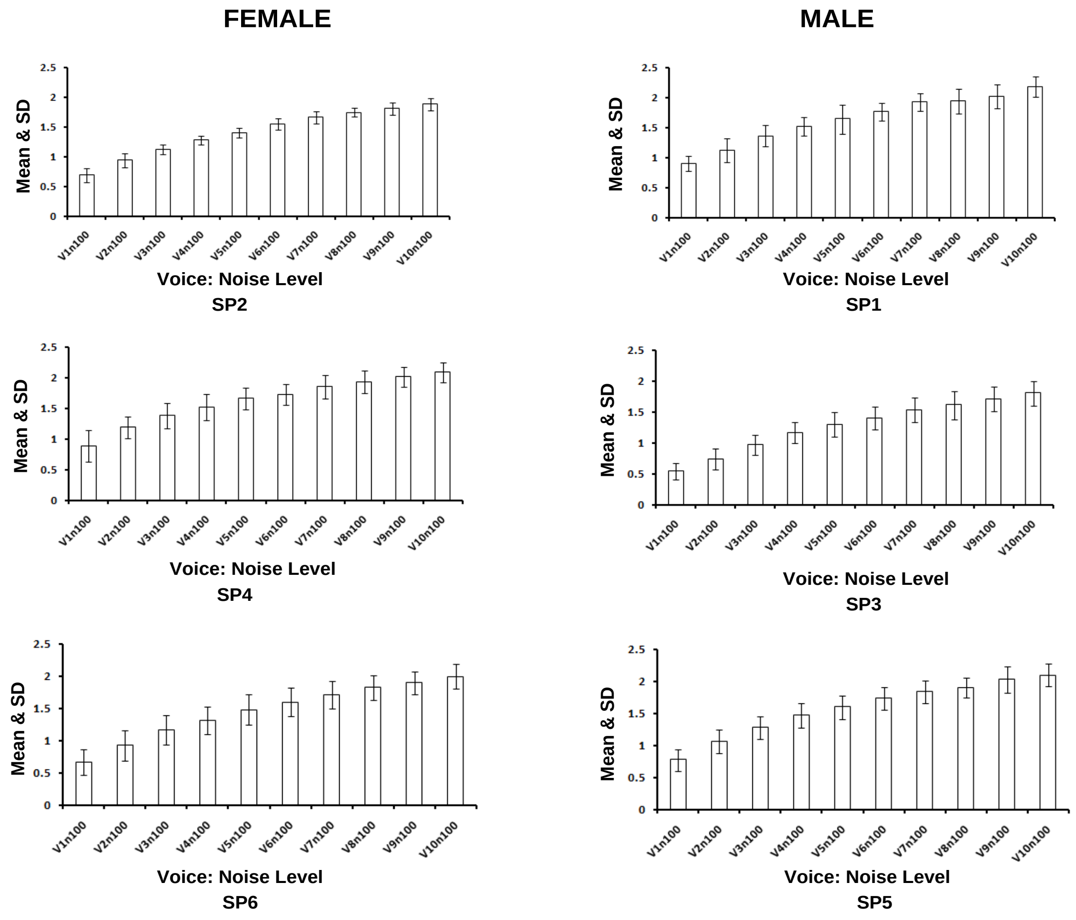

Fig.4. PESQ score of six speakers for v1-v10% at n100%

Figure 5 shows different plots for different percentage of voice part (v1-v100) of the speech signal for speaker1 (sp1) synthesized by HNM at different level of voice part with fixed noise percentage added to it. From the histograms it can be analyzed that voice part serves as an important constituent in speech signal. With no voice added there is no sound heard even at 100 percent noise part. For constant noise part (100% noise) there is a gradual but considerable increase in speech quality as the percentage of voice part increases from 2% till 10 %. However beyond 10% voice level quality increases slightly till a proportion of 50% voice part has reached. The quality becomes quite appreciable (more than a PESQ score of 3) after 50% voice proportion has reached and remains almost same till a percentage of 100 % voice part. This shows that at least 50 % voice part is needed for appreciable speech quality. This can be seen from the histogram plotted for all the six speakers at v50% n100 in figure 6. From the plot it is clear that for all the speakers the quality of signal is appreciable thus HNM proves to be a robust model for children voice as well.

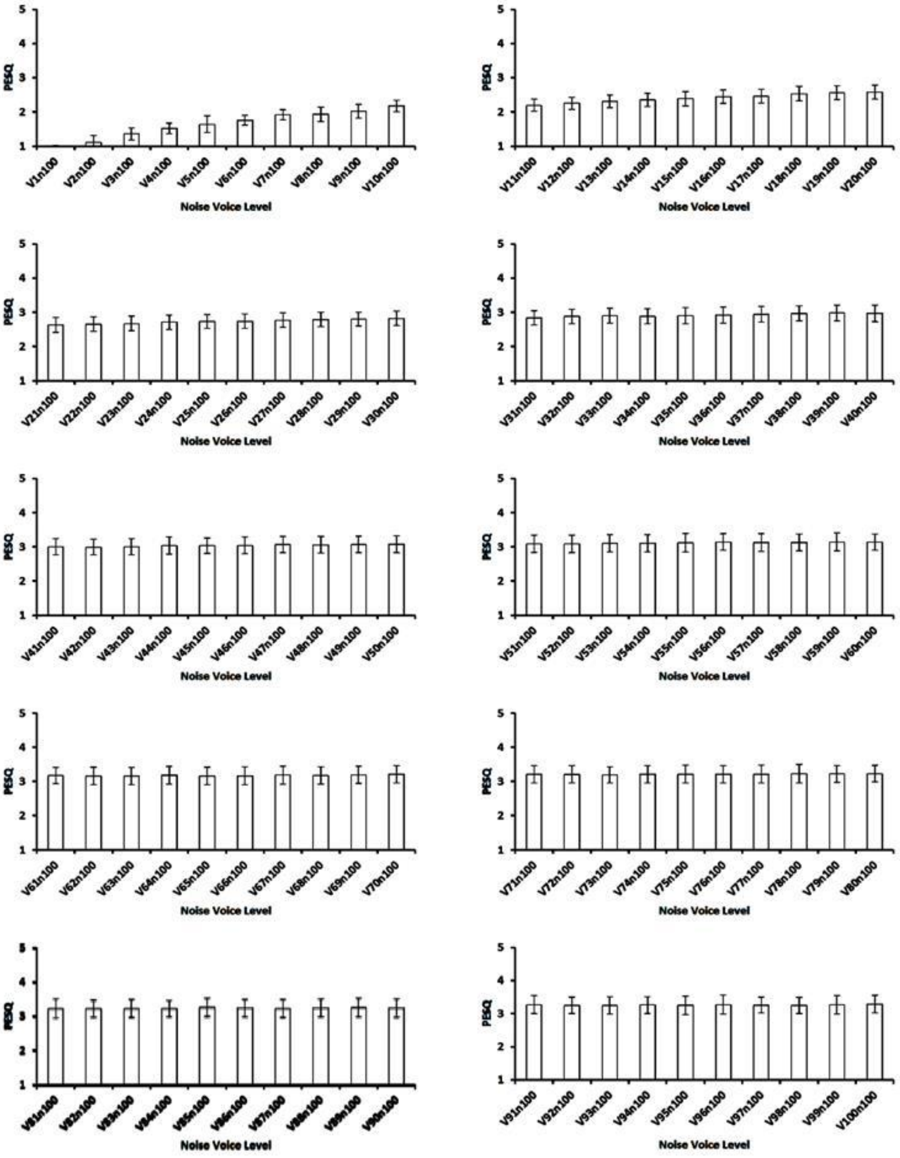

Fig. 5. PESQ score obtained at different level of voice ranging from v1 to v100 and constant noise level (n100) for sp1

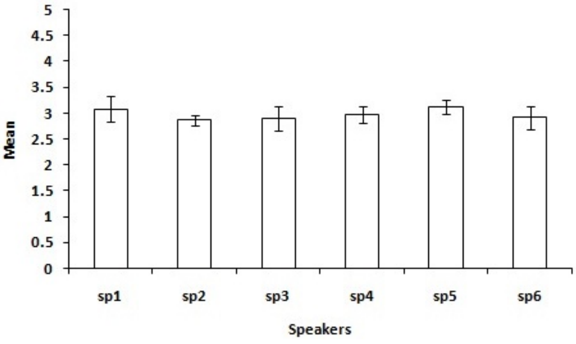

Fig. 6. Mean and standard deviation of all the HNM synthesized speech of all the six speakers at v50n10.

VI. CONCLUSION

Research work is carried out to evaluate and compare the quality of synthesized speech of children in Dogri language. The effect of the proportion of voice part on the synthesized speech quality and intelligibility has been discussed. HNM has been used as analysis and synthesis platform and PESQ as the evaluation method for the speech quality. The results show that voice part serves as an important component in the speech signal. As the voice percentage is increased the quality increases, however the quality of the children speech obtained shows a substantial increase until the value of voice part reaches 50% and later shows a slight increase till v100%. From the results it is quite apparent that HNM model proves a robust model for children speech as it synthesizes all the voices quite clearly. This sheds light on a significant result that HNM model works well with children voice as well as it synthesizes all the voices quite clearly. At 1% voice level shows 0% PESQ score. The worst quality of speech is obtained in between the range from v1n100 to v10-n100 and the best quality of speech is seen in the range from v50n100 to v100n100. Further from the results obtained for all the six speakers (sp1–sp6) it has been seen that the quality of speech signal is speaker dependent, i.e. the worst and best quality is not same for both male and female children speaker.

REFERENCES

1. Prasanna, S. R. M., and Zachariah, J. M., "Detection of vowel onset point in speech", in Proc. IEEE Int Conf. Acoust Speech Signal Processing Orlando, Vol. 4, pp. 4159.2002.
2. Mahendru, H. C. "Quick review of human speech production mechanism", International Journal of Engineering Research and Development, Vol.9, 2014.
3. Chaudhari, U. V, Navaratil J, and Maes, S. H., "Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition", IEEE Trans. On Speech and Audio Processing, Vol. 11, No. 1, pp. 61-69, 2003.
4. Ahmed, A., Mohamed D., "A novel approach to speech segmentation using the wavelet transform", Fifth International Symposiumon Signal Processing and its Applications (ISPRA), pp. 127-130. 1999.
5. Moataz, Ayadi E. l., Mohamed S. K., and Fakhri K., "Survey on speech emotion recognition: Features, classification schemes, and database", ELSEVIER Pattern Recognition, pp. 572-587. 2011.
6. Mugitani , R and Hiroya, S, "Development of vocal tract and acoustic features in children", Acoust. Sci. & Tech, 2012.
7. Rabiner, L. R., and Juang, B. H., "Fundamentals of speech recognition", Prentice Hall Englewood Cliffs NJ, 1993.
8. Moore, B. C, "An Introduction to the psychology of hearing," Academic Press, London, second edition; 1982.
9. Honda, M., "Human speech production mechanisms", NTT Technical Review, Vol. 2.
10. Qi, Y., and Hunt, R. B., "Voiced- unvoiced-silence classifications of speech using hybrid features and a network classifier", IEEE Transactions on Speech and Audio Processing, 1993.
11. Rajput, P., and Lehana, P., " Effect of the Proportion of Harmonic and Noise part on the Quality of Synthesized Speech using HNM in Hindi Language, MAGNT Research Report", Vol.2 (7). pp.116-133, 2014.
12. Lehana, P. K., and Pandey, P. C. "Effect of GCI perturbation on speech qualityin Indian languages", in Proc. Convergent Technologies for the Asia Pacific (IEEE TENCON-2003), Vol. 3, pp. 959-963, 2003.
13. Marwan, A., A., "Fractal Speech Processing", The Press Syndicate of University of Cambridge, pp. 3-4,2003.
14. Denes, P, and Pinson. E., "The Speech Chain," Bell Telephone labs, Murray Hill, New Jersey, 1963.
15. Rabiner. L. R., and Schafer, R. W., "Digital processing of speech signals," Prentice-Hall Inc., Englewood Cliffs. New Jersey, 1978.
16. Furui, S., and Sondhi, M.M, "Advance in speech Signal Processig", Marcel Dekker, New York, 1992.
17. Erro, D, Sainz, I, Navas, E., and Hernaez, I., "HNM-based MFCC+F0 extractor applied to statistical speech syn-thesis", Proc. ICASSP; 2011.
18. Yannis. S., "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE transactions on speech and audio processing; 2001.
19. Moulines, E., and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Commun, 1990.
20. Stylianou Y, Laroche, J., and Moulines. E., "High-quality speech modification based on a harmonic + noise model", in Proc. Eurospeech, 1995.
21. Dudley, H., "The carrier nature of speech", The Bell Syst. Tech. Journal. Vol. 9, No.4, pp. 495- 515, 1940.
22. Dudley, H., and Tarnozy, T. H., "The speaking machine of Wolfgang von kempelen", Acoustic Society of America, Vol. 22, No. 2, pp. 151-166. 1950.
23. Scott. P., "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm", Proc. of MESAQIN, 2002.