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

Varun Sharma
Department of Electronics & Communication Engineering
Sri Sai College of Engineering and Technology, Pathankot, India

Randhir Singh
Department of Electronics & Communication Engineering
Sri Sai College of Engineering and Technology, Pathankot, India

Parveen Lehana
Department of Physics & Electronics
University of Jammu, Jammu

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. There has been a significant amount of research in the field of speech signal processing, and numerous models have been devised so far, but since harmonic plus noise model (HNM) proves to be a more promising of all the existing ones, this paper describes the application of HNM for analysis-synthesizes of children voice . For the analysis of speech signal we have carried out the recording of 7 children speakers (3 male and 4 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 the proportion of noise on the quality and intelligibility of speech signal of children has been investigated at different levels of noise keeping voice part constant. Results suggest that the quality of the children speech increases gradually until the value of noise part is 50% and remains constant till the noise part reaches a percentage of 70%. However as the noise percentage is increased the quality shows a slight decrease but remains constant afterwards. The optimum percentage of noise part for good speech quality has been found to be same for both male and female children speakers. 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 production, 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]. Speech signal is generated from human articulatory system and perceived through ears. It conveys different forms of information to the listener. Apart from the language being spoken and the emotion, gender etc, the identity of the speaker could also be the part of information [2, 3] Fig.1 shows the human speech production system. Lungs, larynx, vocal tract cavity, nasal cavity, teeth, lips, and connecting tubes are main part of this system. Variety of vibrations and spectral temporal compositions that form different speech sounds are produced by combined voice production mechanism [4]. Speech is produced by exhaling air from the lungs. 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 [5-7].

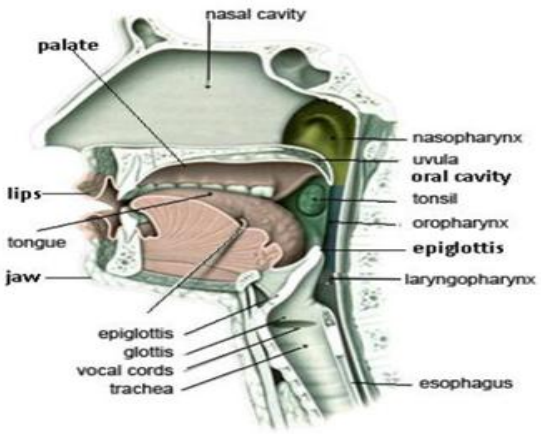

Fig. 1: Anatomy of the human speech production system

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. Hindi is one of the prominent languages of India 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 [8-10]. The objective of this paper is to compare the effect of varying the percentage of noise parts on the perception of synthesized speech of children with 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 in Dogri language. The analysis-synthesis model of harmonic plus noise model used for the assessment of the quality of the synthesized speech is discussed in section 2. Tools and techniques have been described in section 3, Results and discussions are presented in section 4 and the conclusions in section 5.

A. Harmonics plus Noise Model:

HNM which was developed by Stylianou et al. [11] 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. 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 [12].
Harmonics parts in lower band are modeled as sum of harmonics

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

Where $L(t)$ denotes the number of harmonics included in the harmonic part, $\omega_0(t)$ denotes the fundamental frequency while $A_k(t)$ can take on one of the following forms:

$$A_k(t) = a_k(t_i)$$
$$A_k(t) = a_k(t_i) + tb_k(t_i)$$
$$A_k(t) = a_k(t_i) + tc_k(t_i) + t^2 d_k(t_i)$$

Where $a_k(t_i), b_k(t_i), c_k(t_i), d_k(t_i)$ and are assumed to be complex numbers with $\arg\{a_k(t_i)\} = \arg\{c_k(t_i)\} = \arg\{d_k(t_i)\}$ where, arg, denotes the phase angle of a complex number [13-15]. Modulated noise dominates the voiced speech spectrum in the upper band of HNM. The noise part is given by expression

$s_n(t) = e(t)[h(\tau, t) * b(t)]$

where $*$ denotes convolution and $b(t)$ is white Gaussian noise. The synthetic signal is given by:

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

It is important that the noise part $s_n(t)$, be synchronized with the harmonic part $s_h(t)$ [13-14].

II. METHODOLOGY

The research work is divided into two major parts. Fig. 2 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 seven 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 [15]. Fig.2 shows the block diagram of the research methodology

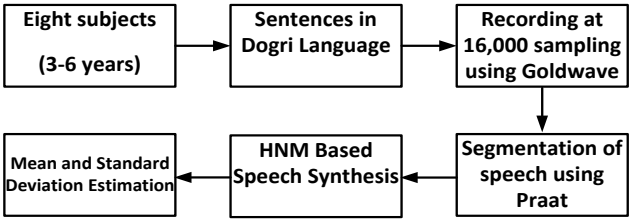

Fig. 2: Block diagram representation of proposed methodology

III. RESULTS AND DISCUSSION

Fig.3 shows the plot for mean and standard deviation of the HNM synthesized speech signal (at 100% voice and noise part) with respect to original speech for all the seven children speakers. The horizontal axis shows the percentage of voice and noise part and the vertical axis represents its corresponding PESQ score. sp1, sp3, and sp5 correspond to male speakers while female speakers are labelled as sp2, sp4, sp6, and sp7. From the histogram shown in fig.3

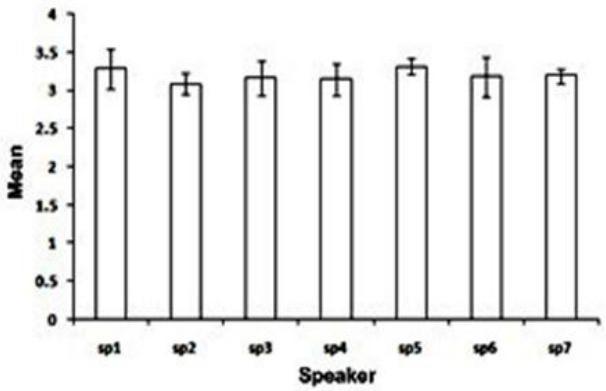

Fig. 3: Mean and standard deviation of all the HNM synthesized speech with respect to original speech

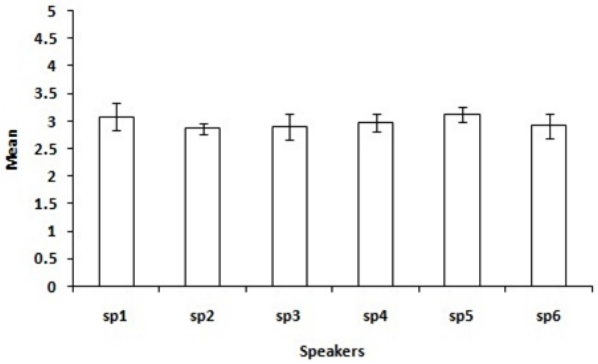

Fig. 4: Mean and standard deviation of all the HNM synthesized speech of all the six speakers at v50n100.

This can be seen from the histogram plotted for all the six speakers at v50% n100 in Fig.4

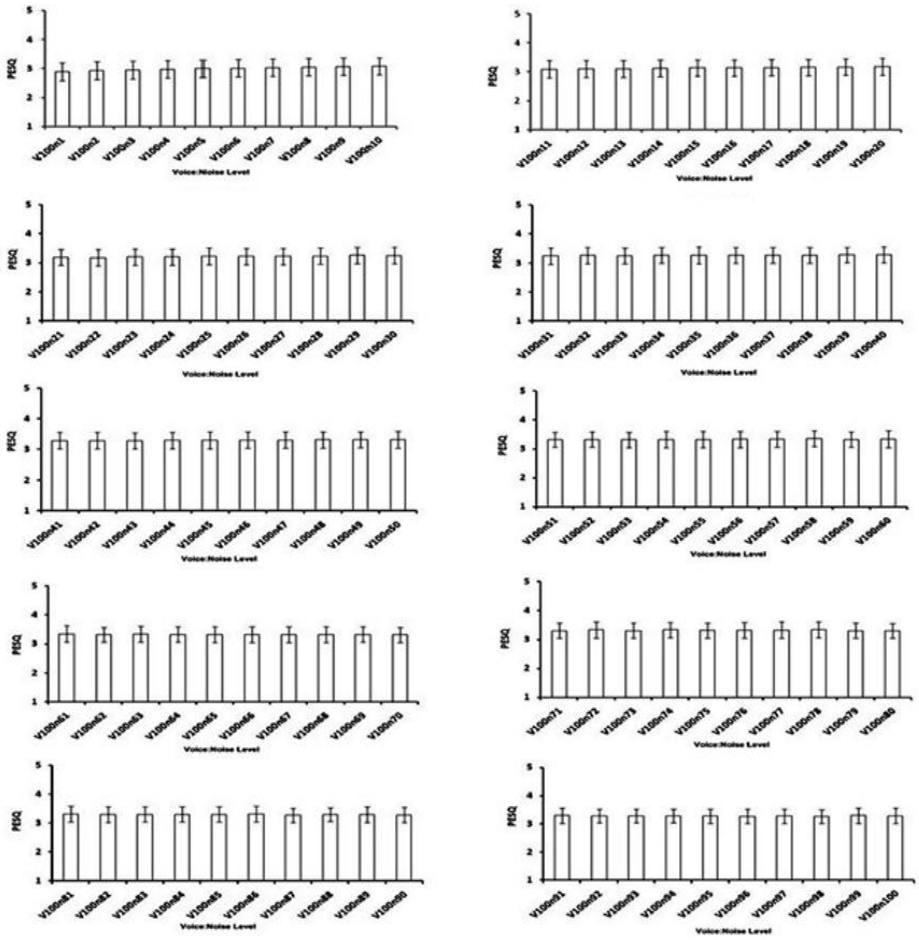

Fig. 5: PESQ score obtained at different level of noise ranging from n1 to n100 and constant voice level (v100) for sp1

From the histogram shown in Fig.5 it can be concluded that for all the speakers the average PESQ score comes out to be same i.e. the voice quality is quite comparable for all the children speakers (male and female). 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. In second experiment, effect of percentage of noise part at constant voice part (v100) on the synthesized speech signal is investigated for all the children speakers. Figure5 shows different plots for different percentage of noise part (n1-n100) e.g. n1 indicates .01% noise part) of the speech signal for speaker1 (sp1) synthesized by HNM at different level of noise part with fixed voice percentage added to it. From the histogram it can be analyzed that in the range from v100n1 to v100n50 the PESQ score obtained is quite acceptable (more than 3) has increased gradually. In the range from v100n50 to v100n70 the quality of speech remains the same. In between v100n70 to v100n100 quality has slightly decrease and becomes constant.

The worst quality of speech is obtained in between the range from v100n1 to v100n10 and the best quality of speech is seen in the range from v100n50 to v100n70.

Further from the results obtained for all the seven speakers (sp1–sp7) it has been seen that the quality of speech signal is not speaker dependent, i.e. the worst and best quality is same for both male and female speaker.

Fig.6 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 analysed 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 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 noise part on the synthesized speech quality and intelligibility has been discussed. HNM has been used as analysis and synthesis platform and the PESQ as the evaluation method for the quality. The quality of the children speech obtained shows a gradual increase until the value of noise part is 50% and later remains constant till 70 % noise is added to it, As the noise percentage is increased the quality decreases slightly but shows no degradation afterwards. From the results it is quite apparent that HNM model proves a robust model for children speech as it synthesizes all the voices quite.

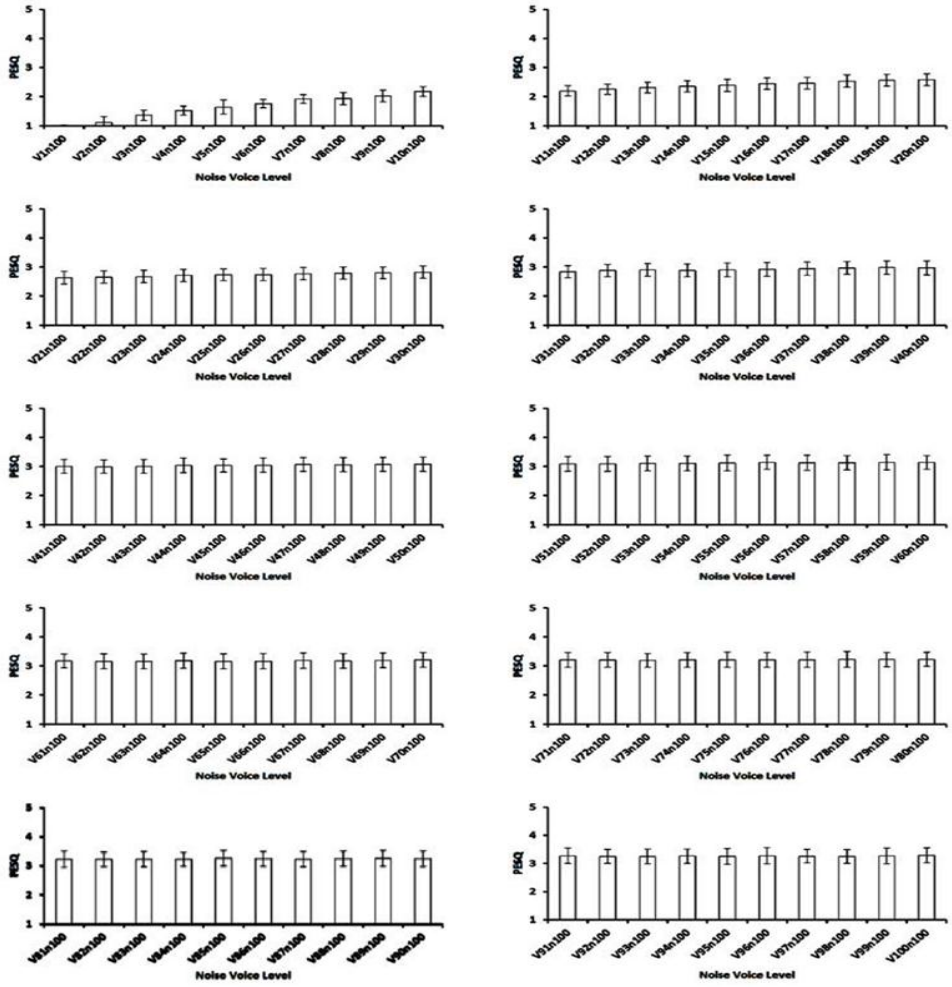

Fig. 6: PESQ score obtained at different level of voice ranging from v1 to v100 and constant noise level (n100) for sp1

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] Pawar RV. Jalnekar RM. Review on Speech Production Model International Journal of Engineering and Innovative Technology (IJEIT);2014,3(119).
[3] Reynolds DA. Automatic Speaker Recognition: Current Approaches and Future Trends.
[4] Mugitani R and Hiroya S. Development of vocal tract and acoustic features in children, Acoust. Sci. & Tech; 2012 , 33, (215)
[5] Moore BC. An Introduction to the psychology of hearing. Academic Press, London, second edition; 1982.
[6] Honda M. Human speech production mechanisms. NTT Technical Review;1(2).
[7] Qi Y and Hunt RB. Voiced- unvoiced-silence classifications of speech using hybrid features and a network classifier. IEEE Transactions on Speech And Audio Processing ;1993,1(2)
[8] Grierson, Sir George, ed. Linguistic Survey of India. Volume IX, Part 4. 1906. Reprinted by Motilal Banarsidas, 1967.
[9] Pushp, P. N., and K. Warikoo, eds. Jammu, Kashmir, and Ladakh: Linguistic Predicament. 2004.
[10] R. V. Pawar, Dr. R.M. Jalnekar, "Review on Speech Production Model,"International Journal of Engineering and Innovative Technology, vol. 3, 2014.
[11] Honda, M., "Human speech production mechanisms", NTT Technical Review, Vol. 2.
[12] Moulines E and Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones Speech Commun; 1990, 9(453).
[13] Hermes D. Synthesis of breathy vowels: Some research methods Speech Commun; 1991(38).
[14] Laroche J, Stylianou Y, and Moulines E. HNS: Speech modification based on a harmonic + noise model in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing'93, Minneapolis, MN, Apr. 1993.
[15] Scott. P., "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm", Proc. of MESAQIN, 2002