IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 11, Issue 6 (Nov. - Dec. 2021), pp.07-15
e-ISSN: 2319-4200, p-ISSN: 2319-4197
www.iosrjournals.org

Evaluating the Performance of the ARMA Model Based Speech Synthesis for Male and Female voices

H.M.L.N.K Herath[1]
[1]*The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka*

Abstract:
Today the synthetic speech of highest quality, generated by existing speech synthesis systems are not close to actual human speech. As the number of applications for synthetic speech increases, the naturalness and the intangibility of synthetic speech will become an important factor in determining its use. This paper presents a comparative study on investigating the naturalness and the intelligibility of the newly proposed Auto Regressive Moving Average (ARMA) based speech synthesis method for female voice. The synthesized speech was analyzed based on subjective and objective measure. The subjective measures, signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR) and mean square error (MSE) were considered, whereas as an objective measure, the Diagnostic Rhyme test was used. The results show that more than 75% of the synthesized words were identified by the subjects in the Rhyme test. The results of the objective measure demonstrate the synthesized speech signals have lower MSE and higher PSNR, SNR values which indicate the signal quality is high. The Pearson's correlation Coefficient values confirmed the synthesized signal is 95% closer to the recoded speech signal. Once the overall results for male voice was compared with female voice, it can be concluded that the proposed method yields better results for synthesizing male voice than female voice.Also the experiment concludes that proposed ARMA model based speech synthesis system produces more natural and intelligible speech for any gender

Key Word*: Auto Regressive Moving Average (ARMA), Speech intelligibility; Naturalness; signal-to-noise ratio (SNR); peak signal-to-noise ratio (PSNR) and mean square error (MSE); Diagnostic Rhyme Test*

I. Introduction

During the last few years, a considerable number of speech synthesis methods has beenproposed and developed in order to boost speech intelligibility while maintaining the naturalness of the synthesized speech. Improving the naturalness and intelligibility of artificially generated speech is a hot topic in speech synthesis field. But still the intelligibility and naturalness of artificially generated speech is not closer to humans' speech. The challenge is still exist.

Vast variety of speech synthesis models were developed in the past few decays by combining with various research fields like Artificial Intelligence, Neural Networks and Deep Learning etc. But the specific, traditional strategies for speech synthesis models, Articulatory Synthesis [1], Concatenation Synthesis [2][3], Parametric Speech Synthesis are still the basic structure for novel techniques. The concatenation synthesis is the most widely used speech synthesis method. The speech is generated by concatenating different speech units (phonemes, diaphones and syllables) in to a single word. The naturalness and the intangibility of the synthesized speech is highlydepend on the recorded speech units in the database. The high-quality natural sounding synthetic speech is produced from larger databases of well-ordered and well –labeled speech. However, it cannot avoid the transition between the speech units which often produce auditory discontinuity and leads to unnaturalness. In addition to that, designed voices for particular application may often sound inappropriate for another application. Recording different voices in both male and female categories needs larger databases and high bit rates involved in transmission of the speech.

Speech in parametric speech synthesis approach is generated by the parameters that are extracted from the human speech samples like fundamental frequency (vocal source), duration (prosody), noise level, etc… Formant Synthesis, Hidden Markov Model (HMM) [4] speech synthesis, Auto Regressive (AR) [5][6] model based Linear Predictive Coding (LPC), Auto Regressive Moving Average (ARMA)[6] are some of the techniques that are used in parametric speech synthesis. Majority of the systems generate artificial, robotic sounds with less naturalness and intelligibility. The resent speech synthesis models produce speech by combining the traditional speech synthesis models with Artificial Intelligence, Neural Network, Deep Learning etc. WaveNet is a speech synthesis model based on neural network for generating raw audio waves. It is fully probabilistic and autoregressive, and it reduces the gap between the state of the art and human-level performance by over 50% for both US English and Mandarin Chinese. As a result, it is capable of producing

audio that are very similar to a human voice [7]. Trcotron [8], Deep voice I [9]are some of the other systems that use Artificial Intelligence, Deep Neural Network for synthesis more natural speech.

Nowadays text-to-speech (TTS) systems have become far more common and an ordinary feature of everyday life. The increasing number and uses of TTS in various day-to-day life applications like reading and communication aids for the blind, deafened and vocally handicapped, educational systems, telephone enquiry systems, e-mail readers, human-machine interactions etc. In near future it may also be used in language interpreters or several other communication systems, such as videophones, video conferencing, smart environment, virtual assistant, intelligent robots or talking mobile phones. As the number of applications for synthetic speech increases, the performance of the synthetic speech with different voices, styles and emotions will become an important factor in determining its use.

Quality of a speech assesses "how" a speaker produces an utterance and includes attributes such as "natural", "raspy", "hoarse", "scratchy" and so on. Quality is known to possess many dimensions, encompassing many attributes of the processed signal such as "naturalness", "clarity", "pleasantness", "brightness", etc. Typically for practical purposes the speech quality is restricted to few dimensions depending on the application[10]. Naturalness is the most important quality and it is described as how much the synthetic voice is similar to the human voice. Speech intelligibility is a different attribute that measures "what" the speaker said, i.e., the meaning or the content of the spoken words. Hence, different methods need to be used to assess the quality and intelligibility of artificially generated speech.

In some applications the naturalness of the synthetic speech is less important than the intelligibility of the synthetic speech.If more natural artificially generated voice message cannot understand or hard to understand by the listener then it is not useful. On the other hand, in some applications like reading and communication aids for the disable people, use synthetic speech as a vocal prosthetic. Those people need more intelligibility speech than others because they understand the context by listing to the speech and it should be clear and understandable. Most of the TTS systems were developed for blind people, produce unnatural speech signals, which affect the emotion component of speech communication. Thus, it may lose several information to be communicated and the listener may find it hard to develop a trust on the speaker. Communication will be more comfortable, attractive and attentive when the speech sounds more humanly and hence the synthesized speech will be more useful This will help to increase the interaction of the user in many applications such as telephone answering machines, e-mail readers. etc

The applications that use speech synthesis systems have a great need in producing both male and female voices, especially in voice prosthesis and translating telephony. Early attempts of producing various voices like male, female and child with high quality output voices have not been very successful. More recent attempts have synthesized female voice mainly by transforming a male synthetic voice or copying an utterance of a female speaker [11]. Speaker independent speech generating systems producequality speech with minor change of values for parameters. A similarity of the synthesized speech is still less than the natural speech. Among number of approaches speaker depend DNN-based TTS and HMM model speaker depended models are working model that used now. But still the artificiality of the synthesized speech is still identified. This study investigates a ARMA modeled based speech analyze algorithm for male and female voices analysis. The synthesized speech evaluated for the naturalness and the intangibility of male, female voice in the context of subjective and objective measurements.

II. Speech Quality Assessments

The usefulness of speech synthesis systems is highly depending on the performance, the naturalness and the intangibility of artificially generated speech of the system. Evaluation of the performance of synthetic speech provides important information about the speech synthesizers in comparison to competing products. Diagnostic evaluation is important for researchers to understand, where the relative strengths and weaknesses of a particular synthesizer and can assist in the development effort by pinpointing specific problems in synthesis. The quality and intelligibility of speech synthesized systems can be quantified using subjective and objective measures.

2.1 Subjective Quality Measurements

Subjective quality measures of speech are obtained by conducting a listening test to the hearing-impaired subjects in their language. There were several subjective tests were available for measure the naturalness and intangibility of synthesized speech. The Diagnostic Rhyme Test (DRT) is the common test for evaluating the speech intangibility.

The Diagnostic Rhyme Test, (DRT) is a testconsists of 96 monosyllabic (single-syllable) word pairs which are distinct from each other only by one acoustic feature in the initial consonant [12]. These fall into one of the six categories of Voicing, Nasality, Sustenation, Sibilation, Graveness, and Compactness [12]. Some illustrative word pairs from the DRT are shown in Table 1; note the similarities and differences between the initial consonants of each word pair.

Table no 1: The Diagnostic Rhyme Test, (DRT) words

Voicing		**Nasality**		**Sustenation**		**Sibilation**		**Graveness**		**Compactness**	
veal	feel	meat	beat	vee	bee	zee	thee	weed	reed	yield	wield
bean	peen	need	deed	sheet	cheat	cheep	keep	peak	teak	key	tea
gin	chin	mitt	bit	vill	bill	jilt	gilt	bid	did	hit	fit
dint	tint	nip	dip	thick	tick	sing	thing	fin	thin	gill	dill
zoo	sue	moot	boot	foo	pooh	juice	goose	moon	noon	coop	poop

2.2 Objective Quality Measurements

Objective measures of speech quality are computed from properties of an original and synthesized speech wave form. Mean Square error (MSE), Peak Signal to Noise Ratio (PSNR) and Signal to Noise Ratio (SNR) measures the naturalness and intangibility of synthesized speech as Objective quality measures.

2.2.1 Mean Square Error

Mean Square error measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimation of speech synthesis.

$$\text{MSE} = \frac{1}{N}\left[\left(r(n) - x(n)\right)^2\right] \quad \text{.....(1)}$$

Where, N is length of input speech signal, x(n) is input speech signal and r(n) is reconstructed speech signal.

2.2.2. Signal to Noise Ratio

Signal to Noise Ration measure used to quantify how much a signal has been corrupted by noise. It is defined as the ratio of signal power to the noise power corrupting the signal in decibars. A ratio higher than 1:1 indicates more signal than noise.

$$\text{SNR (dB)} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_d^2}\right) \quad \text{.....(2)}$$

Where, σ_x^2 is the mean square of speech signal and σ_d^2 is the mean square difference between the original and reconstructed speech.

2.2.3 Peak Signal to Noise Ratio

This ratio is often used as a quality measurement between the original and a reconstructed one. The higher the PSNR, the better the quality of the signal. The PSNR represents a measure of the peak error between the reconstructed and the original signal in decibels.

$$\text{PSNR (dB)} = 10\log_{10}\left(\frac{Nx^2}{\|x-r\|^2}\right) \quad \text{.....(3)}$$

III Experiment

3.1Speech material

Speech samples (the words in DRT) were recorded by a female speaker with British accent. The recording was at laboratory conditions. The voice was recorded through the MKH800 microphone, with the volume set at 60 dB. The recording wav files were all in single channel, with frequency at 16 kHz.

3.2 Speech analysis and Synthesis

Vocal tract of humans molded in terms of both poles and zeros by the Autoregressive Moving Average filter models. It described the unknown model with a pole-zero filter as following form

$$y(n) = \sum_{k=0}^{q} b_k\, x(n-k) + \sum_{k=1}^{p} a_k\, y(n-k) \quad \text{.....(4)}$$

x(n) and y(n) are the input output signal of the unknown system { a_k,k=1, ..,p}and { b_k, k=0, ..,q} are the filter coefficients corresponding to poles and zeros. Performing z transforms on both sides of equation (4), the equation becomes

$$Y(z) = \frac{B(z)}{A(z)}\,.X(z) \quad \text{.....(5)}$$

Where, $A(z) = 1 - \sum_{k=1}^{p} a_k\ z^{-k}$ $\qquad$ $B(z) = \sum_{k=0}^{q} b_k\ z^{-k}$

Y(z) and X(z) are the z transform of y(z) and x(z) respectively. The system transfer function is given by B(z) / A(z), where B(z) and A(z) are called the Moving Average part and Auto Regressive part of the model[5];

$$H(z) = \frac{Y(z)}{x(z)} = \frac{\sum_{k=0}^{q} b_k\ z^{-k}}{1 - \sum_{k=1}^{p} a_k\ z^{-k}} \quad \text{.....(6)}$$

Equation (13) can be expressed as follows,

$$H(z) = \frac{y(z)}{x(z)} = \frac{b_0 + b_1 z^1 + \cdots + b_q z^{-q}}{1 - a_1 z^1 - \cdots - a_p z^p} \quad \text{.....(7)}$$

Speech parameters frequency, phase, amplitude and attenuation coefficient derived according to the equation (7) given in ARMA model, the partial fraction representation H(z) express as,

$$H(z) = \frac{B(z)}{A(z)} = \frac{r_m}{s-p_m} + \frac{r_{m-1}}{s-p_{m-1}} + \cdots + \frac{r_0}{s-p_0} + k(z) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

Where, the values $r_m \ldots r_0$ represents the residues, the values $p_m \ldots p_0$ are poles and k(z) is a polynomial in z, which is usually 0 or constant. The real and imaginary parts of the complex transform of residues r_m are used to estimate the amplitude A_n and the phase $\emptyset_n$

$A_n = |r_m| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$ (9)

$$\emptyset_n = \tan^{-1}\left(\frac{r_{im_n}}{r_{Re_n}}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

Pole locations pm used to calculate the frequency and attenuation coefficient r_n

$$f_n = \tan^{-1}\left(\frac{p_{im_n}}{p_{Re_n}}\right) \times \left((Fs/2)/\pi\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(11)$$

$$r_n = |p_m| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(12)$$

Where, fs sampling frequency, n designate the frequency increment (n= 0, 1,…,N) and R_e an I_m are the real and the imaginary parts of the $r_m \ldots r_0$and $p_m \ldots p_0$ transform.Most dominant poles gain from ARMA model were converted to frequency, phase, amplitude and exponential decay values. Then filter the values by given different conditions to obtained the most important details.

Parametric speech synthesis model calledSinusoidal Model Noise model were used to resynthesized the speech signals. It models the speech or music signals as sum of sinusoids each with time-varying amplitude, frequency and phase. Since the sinusoidal noise model has the ability to remove irrelevant data and encode signals with lower bit rate, it has also been successfully used in audio and speech coding. Equation (13) represents a decaying sinusoidal wave.

$$x(t) = \sum_{i=0}^{N} A_i(t)\, e^{-\alpha t} \cos\left(2\pi f_i t + \phi_i\right) + r(t) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(13)$$

Where, $A_i(t)$, radian frequency $2\pi f_i$ and ϕ_i phase in radians of sinusoidal *i*at time *t*, and *r*(*t*) is a noise residual, α is the exponential Decay and $e{-}\alpha t$ is the decay rate.

3.3 Objective and Subjective Test

The Diagnostic Rhyme Test was carried out in order to evaluate the ARMA based speech synthesis model. A total of 10 non-native English speakers participated in the experiment, with 2 male and 8 females. Allparticipants were graduates between 25 to 30 years of age. They have not participated in any subjective test whatever for at least the previous six months and not in any listening-opinion test for at least one year. They have never heard the same word lists before. The words were played from the PC to the test participant via headphones in laboratoryconditions. The participants choose the correct word between two presented word pairs defined by DRT test.

Mean Square Error, Peak signal to Noise Ratio, Signal to Noise Ratio and Pearson's Correlation coefficient between the original and a reconstructed speech signal were calculated.For comparison purpose these tests were repeated for Male voice.

IV Results and Discussion

4.1The Diagnostic Rhyme Testy

The subjects were able to identify synthesized words correctly between two words that are given to them in the Diagnostic Rhyme Test, in all six categories. Each category was consisted of 16-word pairs.In the diagnostic rhyme test, the difference between the word pair is in one phoneme and most of the time they are pronounced to be small difference. The percentage of number of words identified by each subject was as depicted in the Table 2.

Table 2: The percentage of no of words identified by each subject

Subject	Voicing	Nasility	Sustenation	Sibilation	Graveness	Compactness
S1	93.75	93.75	43.75	75	62.5	81.25
S2	93.75	87.5	56.25	87.5	75	81.25
S3	87.5	93.75	75	87.5	87.5	81.25
S4	87.5	93.75	62.5	87.5	62.5	68.75
S5	87.5	87.5	68.75	81.25	75	81.25
S6	81.25	93.75	87.5	75	87.5	75
S7	68.75	87.5	75	68.75	81.25	62.5

S8	81.25	93.75	87.5	75	68.75	81.25
S9	75	87.5	81.25	81.25	81.25	81.25
S10	81.25	87.5	87.5	81.25	87.5	81.25

The 75% of the word in the rhyme test were successfully identified by each subject in each category. The average percentage of each category is shown in the Figure 1 for male and female voices.

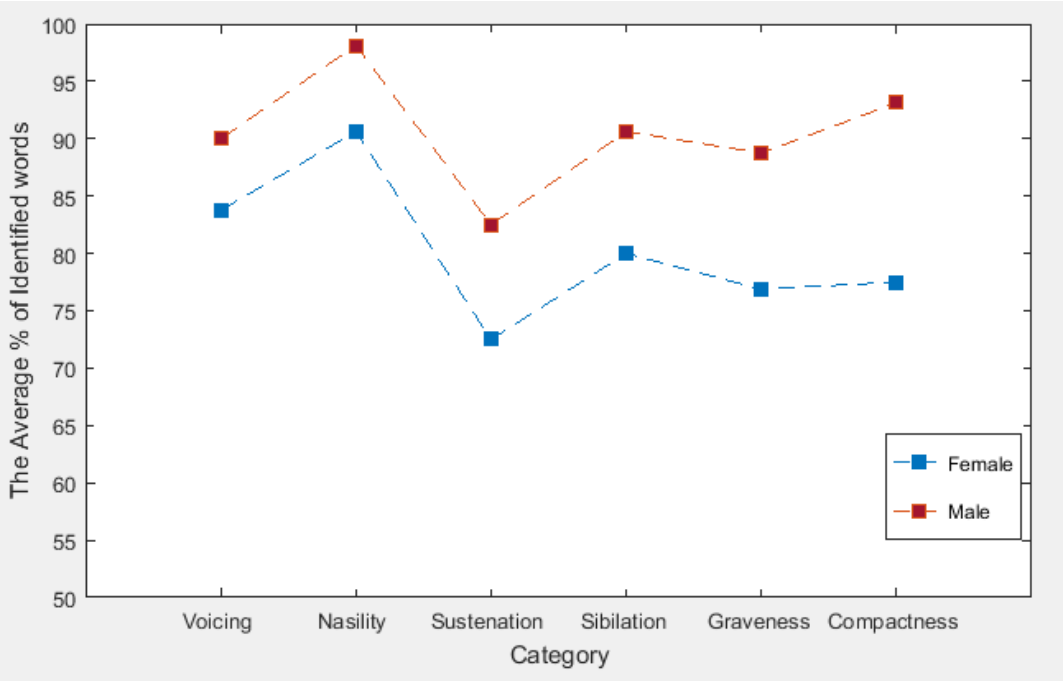

Figure 1: The average percentage of identifiedwords in each category, Male and Female speech samples

The average percentage of identification of words(Figure 1) synthesized using Female voice is more than 70% for allcategories. The Nasality category has the highestpercentage value (90%) among all other categories.The Sustenation category has 73% of wordidentification percentage and it is lower than allother categories.

Figure 1 illustrates 80% and 75% of the words that belong to any category in male or female respectively can be identified clearly.The Nasality category has the highest average identification percentage value while the Sustenation category has the lowest average identification percentage value in both male and female voice. The identification percentage of synthesized words belong to male voice, higher than the female voice identification percentage in all categories. The pattern of identification percentage of each category of male and female voice was same.According to the above result any of the words that are belong to any category can be identified clearly.

4.2 Peak Signal to Noise Ratio

The average peak signal-to-noise ratio value of synthesized female and male voices were shown in the Table 3.

Table 3: Average PSNR of each category

	Voicing	Nasility	Sustenation	Sibilation	Graveness	Compactness
Female	23.33	25.28	23.59	23.03	24.16	24.28
Male	35.53	36.31	34.26	34.50	33.15	38.86

According to the Table 3, average PSNR values of all categories are higher in male voice than female voice. In both female and male voices, maximum average value was obtained in Nasility and compactness category.

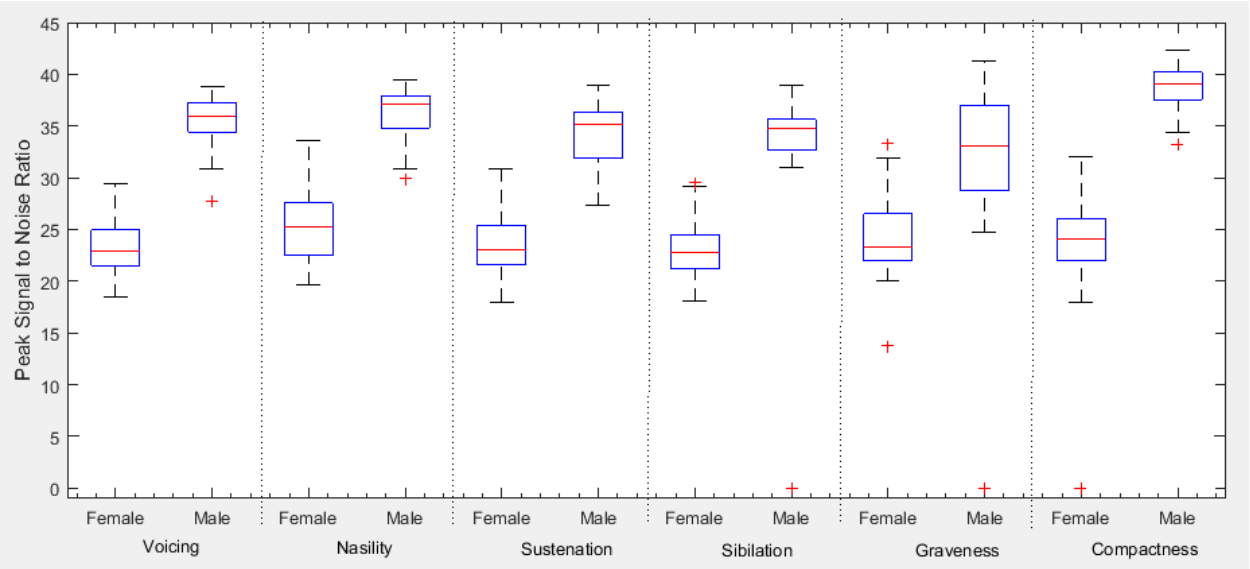

Figure2: PSNR of each category, male and female speech samples

Boxplot of the PSNR values of Male and Female synthesized voices were shown in the Figure 2. It clearly illustrations that the synthesized male voices have higher PSNR median value than female voices in each category. The variation of the PSNR values around the median is very little in most categories except Graveness ofmale voice.The experiment clearly shows that male voice has higher PSNR than the female voice. When PSNR is higher, the quality of the signal is better. That means quality of the synthesized male voice is higher than the female voice.

4.3 Mean Square Error

The average of the Mean Square Error (MSE) for male voice is lesser than the female voice (Figure 3). For male voices it is less than 0.001 but female voice it is between 0.004 and 0.006.That implies the male voices have lower error than female voices.But the error of both male and female voices are considerably lower and thus the quality of male as well as female voices are high

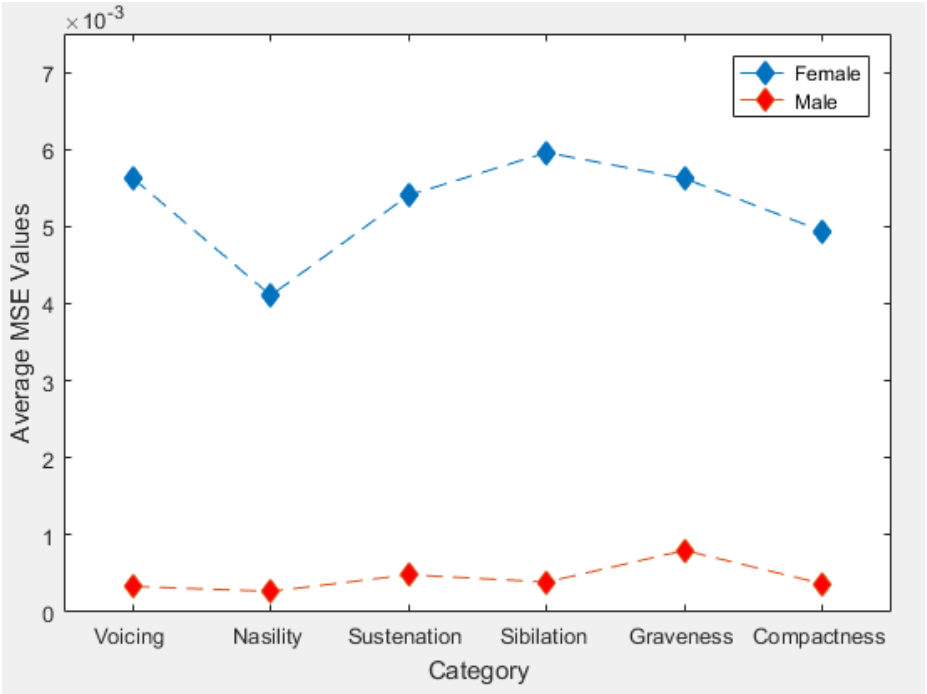

Figure3: The average of MSE in each category, Male and Female speech samples

4.4 Signal to Noise Ratio

The SNR of the recorded signal and the synthesized signal of each word of voicing category is shown in the Figure 4.SNR values of synthesized words of female voice have higher value than the recorded female words. The difference of the SNR value between the synthesized word and the recorded word is smaller. This pattern is similar in all categories of male and female voice samples. SNRmeasure how much a signal has been corrupted by noise. The results indicate that the re-synthesized signals have no additional noise is generated from the synthesized algorithm. If that happens the results may have huge variation of the SNR values of the recorded signal and the synthesized signal. The synthesized signal has the noise component as the recorded signal. It also shows that the both synthesized and the recorded signals have same proportional of signal and noise component.

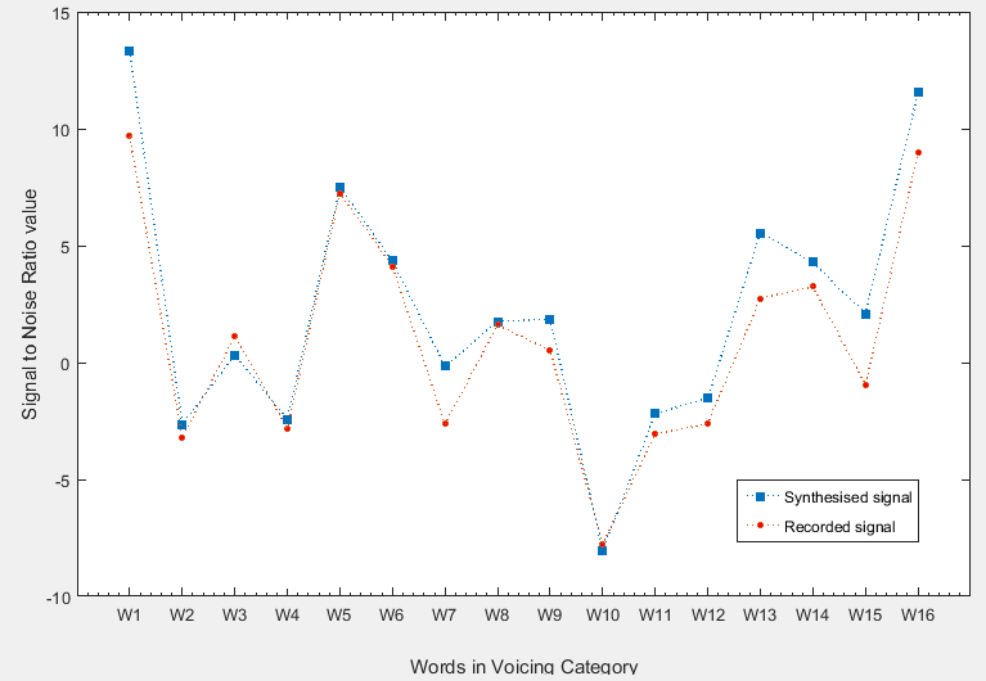

Figure4: The SNR value of Synthesized and Recoded Female sample words in Voicing category

4.5 Correlation Coefficient Value

The Pearson's Correlation Coefficient between the synthesized speech signal and the recorded signals in each category is shown in the figure 5.

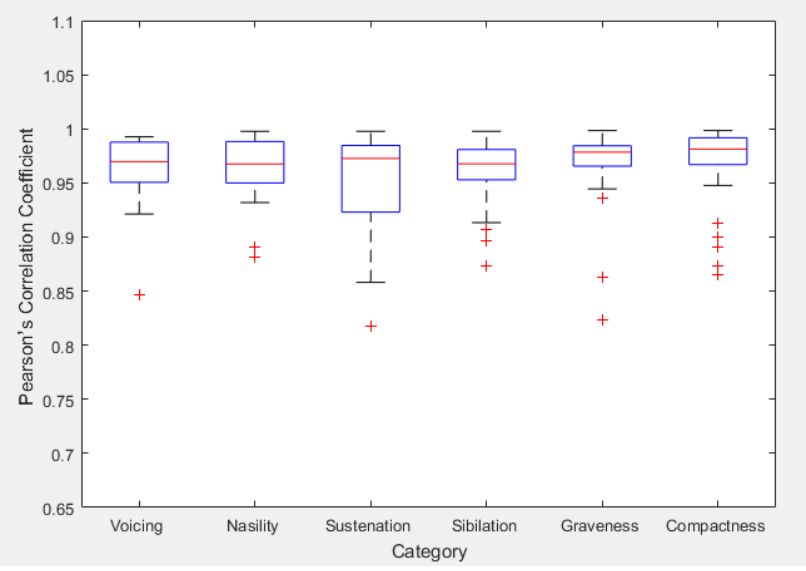

Figure5: Pearson's Correlation Coefficient between the synthesized speech signal and the recorded signals in each category.

Median Pearson's correlation coefficient value of each category is greater than 0.95. The variation of the Pearson' correlation coefficient does not vary in large rage in all categories. All correlation values that have

obtained are more than 0.8. The average Pearson's correlation coefficient values of male and female voice were shown in figure 6.

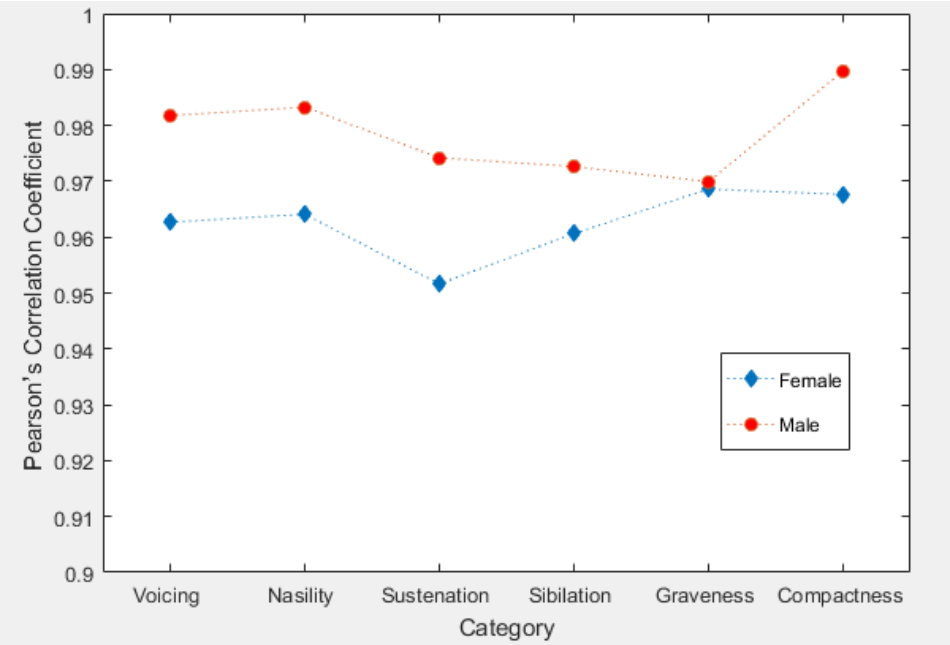

Figure6: The Average Pearson's Correlation Coefficient between the synthesized speech signal and the recorded signals in each category.

The Average Pearson's Correlation Coefficient values were greater than 0.95 for all categories in both male and female voices.Lowest average Pearson's Correlation Coefficient value was observed in Sustenation category for both male and female voices, whereas the largest value was observed in Compactness category.The average Pearson's Correlation Coefficient values for male voice were greater than female voice in all categories. These resultsindicate the speech signals that were reconstructed are closer to the recorded speech signals forbothgender.

Overall results implyfemale voice has higher Identification percentage of the DRT, higher PSNR, average Pearson's Correlation Coefficient and lower MSR value. That meansthe synthesized words of female voice can be identified clearly in all categories and the similarity between the synthesized speech signal and recorded speech signals were closer. The output also points out the percentage identification of the DRT, PSNR and average Pearson's Correlation Coefficient values were higher in male voices than female voice while the MSR is lower in male voice than female voice. Among all the categories the words belong to Nasality category has the highest percentage of identification value and the higher values of PSNR and lower MSR value than other categories. The Sustenation category has the opposite reaction for that.

VI Conclusion

The study investigates the naturalness and the intangibility of synthesized male and female voices using an unique process. One algorithm was used to analyze the recoded samples in both male and female voices. The quality of the proposed speech analysis and synthesized system is evaluated using both subjective and objective measurements for naturalness and the intangibility. The subjective quality measurement, the diagnostic rhyme test proves that synthesized words in any category can be identified clearly for both male and female voices in higher percentage. The objective quality measurements of PSNR, MSE and SNR verifies that the synthesized signals have lower error and lesser noise with higher quality signal for both male and female voice. The Pearson's correlation coefficient values show that generated signals were similar to the original recorded signals. For all the subjective and objective quality measurements conclude that the proposed speech analysis and synthesized method generated more natural and more intangible speech for both male and female voices. Furthermore, the naturalness and the intangibility of synthesized male voice is greater than the synthesized female voice. The speech system extracts the speech information of Nasality category than all other categories. The experiment concludes that the ARMA based speech analysis algorithm extract the most dominant speech informationusing unique filter conditions from both male and female voices. The proposed model can be used to resynthesized speech signals more naturally and more intangible for both male and female voice.

References

[1]. Vaseghi, S., V., *Multimedia Signal Processing: Theory and Applications in Speech, Music and communications.* John Wiley and Sons Ltd, 2007.

[2]. Keller. E., Baily, G., Monaghan, A., Huckvale, M., *Improvements in Speech synthesis, COST 258: The Naturalness of Synthetic Speech,* John Wiley & Sons, LTD .

[3]. Taylor, P., *Text to Speech Synthesis*, Cambridge University Press, 2009.

[4]. C.V. Botinh˜ao, *Intelligibility enhancement of synthetic speech in noise* Doctor of Philosophy thesis, Institute for Language, Cognition and Computation School of Informatics University. of Edinburgh. 2013

[5]. Wang, M., *Speech Analysis And Synthesis Based On ARMA Lattice Model,* Master's Thesis, University of Windor, 2003.

[6]. Rabiner, L., Juang. B., *Fundamentals of speech Recognition*, Prentice Hall International, 1993.

[7]. Avdnoord. A, Sedielem, D, Heigazen, Simonyan, K, Vinyals. O, Gravesa. A, KalchbrennerN. ,Senior A, Korayk. K, *Wavenet: A Generative Model For Raw Audio*,arXiv.org (2016)

[8]. Wang. Y., Skerry-Ryan. R., Stanton. D., Wu. Y., Weissy. R., Jaitly. N, Yang. Z, Xiao. Y, Chen. Z, Bengioy. S,Le. Q, Agiomyrgiannakis. Y.,Clark. R., Saurous. R.,*_Tacotron: Towards End-To-End Speech Synthesis*, arXiv.org (2017)

[9]. Arik. S., Chrzanowski. M., Coates. A, Diamos. G., Gibiansky. A, Kang. Y, Li. X., Miller. J., Ng. A., Raiman. J, Sengupta. S., Shoeybi. M, *Deep Voice: Real-time Neural Text-to-Speech*, arXiv.org (2017)

[10]. Lin. W., *Multimedia Analysis, Processing and Communications*, Springer-Verlag Berlin Heidelberg 2011, pp 623-654

[11]. Karlsson. I., *Female voices in speech synthesis*, Journal of Phonics, (1991)19, 111-120

[12]. D. Ferris, *Techniques and Challenges in Speech Synthesis*, A thesis of Bachelor of Engineering in Electrical Engineering at The University of Newcastle, Australia. 2016

[13]. Herath L., *Evaluation Of Intelligibility And Naturalness Of Synthesized Speech Based On Arma Model*, Proceedings of the 2nd South Asia Conference on Multidisciplinary Research 2019 Vol. 2

[14]. Kayte. S., Mundada. M., Kayte. C.,*Performance Evaluation of Speech Synthesis Techniques for Marathi Language,International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, 2015*

[15]. Kayte. S., Mundada. M., Kayte. C., *Performance Calculation of Speech Synthesis Methods for Hindi language*, IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 13-19

[16]. Nusbaum. H., Francis. A., Henly. A., *Measuring the Naturalness of Synthetic Speech*, International Journal of Speech Technology, 1, 7-19 (1995)

[17]. Herath, H.M.L.N.K., Wijayakulasooriya,J.V.*Modeling of Phoneme Transitions for Natural Synthesis of Speech* (2018), International Journal of Computer Applications –Volume 181-No 23-