Nigerian Journal of Engineering
Faculty of Engineering
Ahmadu Bello University
Samaru - Zaria, Nigeria

Vol. 25, No. 1, Sept. 2018
ISSN: 0794 - 4756

THE EFFECT OF THE USE OF WHITE NOISE FOR MASKING STUTTERED SPEECH RECONSTRUCTION

S. A. Alim*[1], N. K. Alang Rashid[2] and I. A. Abdullateef[3]

[1]. Mechanical Engineering Department, Ahmadu Bello University, Zaria, Nigeria.
[2]. Nuclear Engineering Department, Universiti Teknologi Malaysia, Johor, Malaysia.
[3]. Electrical Engineering, University of Ilorin, Ilorin, Nigeria.
***Corresponding Author email: moaj1st@yahoo.com**

ABSTRACT

Stuttering can be defined as the unintentional disruption in the normal flow of speech by dysfluencies, which include repetitive pronunciation, prolonged pronunciation, blocked or stalled pronunciation at the phoneme or the syllable level. The effect of noise masking on the reconstructed stuttered speech is the focus of this study. This study aimed at finding out the effects of white noise masking on the reconstruction of stuttered speech. Three stuttered words; anniversary, department and sales were masked with 5dB white noise. LPC analysis – synthesis was used for the speech reconstruction, while Welch power spectral density (PSD) estimates was used in evaluating the speech signals in frequency domain. The algorithm effectively recreated the speech samples via reconstruction. The dominant peaks from about 2 kHz were modulated by the masking noise. As such, all the repetition in the noise masked region have reduced power, while the lowest frequency points also had its power increase for the three stuttered words considered. The added white noise as a masking noise thus effectively reduced the repetitions and by extension the stuttering in the speech.

Keywords: *Noise masking, Speech Reconstruction, LPC analysis, LPC synthesis.*

INTRODUCTION

Stuttering can be defined as the unintentional disruption in the normal flow of speech by dysfluencies, which include repetitive pronunciation, prolonged pronunciation, blocked or stalled pronunciation at the phoneme or the syllable level (Chee, Ai, Hariharan, and Yaacob, 2009a; Hariharan, Chee, and Yaacob, 2012; Zhang, Dong, and Yan, 2013). Some of the unusual behaviors of stuttering is that it is variable. It can be manipulated and altered by a wide variety of strategies (Voigt, Hewage, and Alm, 2014). Stuttering cannot be completely treated, however, it may disappear after some time, or stutterers can be trained to adjust their speech to speak fluently with the aid of suitable speech pathology treatment. This shaping has its effects on the effort, tempo, duration, or loudness of their utterances (Awad, 1997; Hariharan *et al*., 2012).

Dysfluencies associated with stuttering can be classed into four main categories. Bursts stuttering occurs when a syllable is repeated when speaking for example ‘He wa-wa-was a great man’ or ‘caaaaaaaaaake’. Reciprocating stuttering occurs when some syllables are repeated when speaking, for example ‘He wwwas a great man’ or ‘u-um-um-um’ or elongated for instance ’uuuum’ or recurring syllable before speaking, for instance ’wa wa wa wa water’. Blocking stuttering occurs when a word is difficult to pronounce in a sentence for a few seconds unsuccessfully, such as ’He w——as a great man’. Interjections are added to the sentence for example ‘I have *um, um,* a test today’ or ‘School is, *you know,* fine’ or ‘The test was, *well,* hard’ (Awad, 1997; Hollingshead and Heeman, 2004; Hariharan *et al*., 2012; Zhang *et al*., 2013; Manjula and Kumar, 2014).

Burst stuttering and reciprocating stuttering are the most frequent forms of stuttering and are part of the main issues that affect speech fluency (Zhang *et al*., 2013). There is a larger quantity of repetition in general, as compared with other types of dysfluencies that stutterers experience (Chee, Ai, Hariharan, *et al*., 2009b). Repetitive pronunciation is a common characteristic of the two categories of stuttering, therefore, they are together named repetitive stuttering (Zhang *et al*., 2013). Many stutterers, find it challenging to terminate sentences. The more severe the stuttering, the more difficulty they experience in starting and ending sentences (Acton, 2004).

According to psychoacoustics theory, masking is an essential component in human hearing (You, Rahardja, and Koh, 2007). It is usually challenging to hear one sound when a much louder sound is present, this task is called masking. The masking effect is a property of the human auditory system that efficiently sets a sound level or threshold for auditory perception. Therefore, any speech or noise components below the masking threshold will not be heard by the listener (Djebbar, Abed-Meraim, Guerchi, and Hamam, 2010). Noise masking improves the speech recognizer performance by decreasing the signal-to-noise ratio to a static value. Noise masking eradicates low-energy spectral details that are only evident in (very) clean speech situations but which are not relevant in more realistic situations (in the presence of noise) (Zhang, Demuynck, and Van hamme, 2010).

Only about 5 to 10% of the human population has a completely normal form of oral communication in relation to numerous speech features and healthy voice. The rest of the population (about 90 to 95%) exhibit some forms of speech disorder such as stuttering, apraxia of speech, dysarthria and cluttering (Manjula and Kumar, 2014). Nearly 2% of adults exhibit stuttering, while about 5% of children stutter (Conture and Yaruss, 2002; Oliveira, Cunha, and Santos, 2013). This study is part of the attempts to proffer some solutions to stuttering as a type of speech disorder.

Some audio parts cannot be heard when they are masked by other audio parts. This implies that human listeners cannot differentiate between the original speech and the speech distorted by a processing step if the distortions in the processed speech are masked by some components of the original speech retained in the processed speech. Masking effects occur not only when sounds are presented concurrently but also when they are not (You *et al.*, 2007). However, the choice of the masking signals for active protection of speech information against the leakage on acoustic channels is an open issue. The masking signals can be pink or white noise, as well as music, speech-like signals or speech cocktail signals (a mixture of speech signals of many speakers) - are often used for the shielding of speech information (Seitkulov, Boranbayev, Yergaliyeva, Davydov, and Patapoviche, 2014). The study aimed at finding out the effects of the use of white noise masking on the reconstruction of stuttered.

METHODS

For the purposes of evaluating the effects of white noise on stuttered speech reconstruction, three stuttered words from the same speaker were used, anniversary, department and sales. The online database was the easiest method to have access to stuttered speech samples. The stuttered speech were gotten from UCLASS (University College London Archive of Stuttered Speech) database. UCLASS had only English speakers. The three stuttered words used were extracted from the speech samples obtained from the UCLASS website. A 5dB white noise was added to the words before the speech reconstruction using Linear Prediction Coefficient (LPC) was carried out. As a result of an experiment conducted during the PhD research by Alim, 2017, masking speech with 5dB white noise gives a compromise between speech quality and intelligibility. In order to obtain the frequency domain representation, several methods are available. In this study, a Welch Power Spectral Density (PSD) estimate was used.

Noise Masking

White noise was used for the masking of the stuttered speech. It is a randomly generated Gaussian noise that has a constant Power Spectral Density (PSD). The important criteria for masking signals are that they are made in an indiscriminate way. White noise can be made from thermal noise of semiconductor or other natural types of noise from normal physical activities. Moreover, white noise has to be restricted in frequency range and extend only for the range of speech signals, (from 125 to 5600 Hz), with the collapse of characteristics out of the array of transmission of 12 dB per octave (Seitkulov *et al*., 2014).

LPC Speech Reconstruction

Linear predictive coding (LPC) is most commonly used for low or medium bit-rate speech coders (Mansour and Al-Abed, 2010). The reflection coefficients are calculated from each frame of speech samples. Because significant details about the vocal tract model is extracted as reflection coefficients which have fewer redundancy than the original speech. Thus, fewer number of bits are needed to quantize the residual. This quantized residual along with the quantized reflection coefficients are transmitted or stored. The output of the filter, termed the residual signal, has fewer redundancy than original speech signal. Speech is reconstructed by taking the residual signal through the synthesis filter. If both the linear prediction coefficients and the residual sequence are existing, the speech signal can be recreated by applying the synthesis filter. The diagram of the LPC reconstruction algorithm is shown in Figure 1.

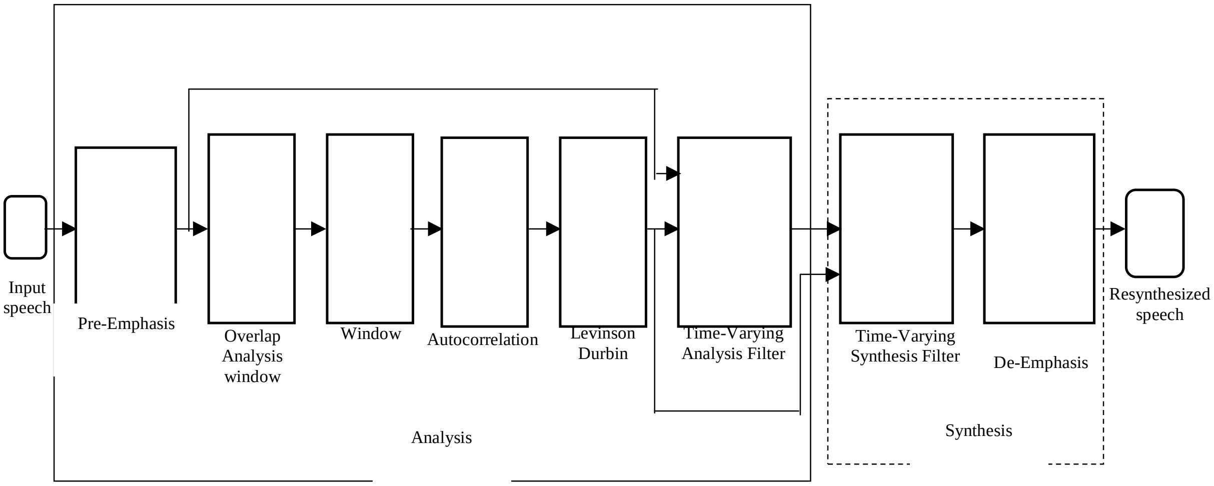


Figure 1: Flow diagram of the LPC analysis-synthesis algorithm

Figure 1 shows the flow of the LPC analysis-synthesis algorithm. It clearly lists out all the processes that the speech goes through in order to create a resynthesized speech. The stages such as pre-emphasis, de-emphasis and window are just filter. While the first two are to remove background noise, the last one is to remove the discontinuities at the edges of the frames after the overlap analysis windowing. The equations for the analysis and synthesis filters are discussed in the subsequent sub-section.

LPC Analysis Filter

Linear Predictive Coding is the most efficient form of coding technique (Jones *et al*., 2009; Suman, 2014) and it has been used in various speech processing applications for depicting the envelope of the short-term power spectrum of speech. In LPC analysis of a speech sample is predicted by a linear combination of past samples, and given by Equation 1 (Rabiner and Schafer, 1978):

$$\hat{s}(n) = \sum_{k=1}^{n} a_k . s(n-k) \tag{1}$$

where $\hat{s}(n)$ is the predictor signal, a_k are the LPC coefficients and p is the LPC order. The residual signal $e(n)$ is derived by subtracting $\hat{s}(n)$ from $s(n)$:

$$e(n) = s(n) - \hat{s}(n) \tag{2}$$

$$= s(n) - \sum_{k=1}^{p} a_k . s(n-k) \tag{3}$$

Applying Z-transform to the equation (3),

$$E(z) = S(z) - \sum_{k=1}^{p} a_k . z^k S(z) \tag{4}$$

$$= S(z)\left[1 - \sum_{k=1}^{p} a_k . z^k\right] \tag{5}$$

But $A(z) - 1 - \sum_{k=1}^{p} a_k z^{-k}$

$$E(z) = S(z)A(z) \tag{6}$$

where $E(z)$ and $S(z)$ are the z-transforms of the residual and the speech signals respectively, and $A(z)$ is the LPC analysis filter.

The short-term correlation of the input speech signal is removed by assigning an output $E(z)$ with a flat spectrum. After implementing the analysis filter, the speech signal is quantized. The quantized signal is then synthesized to get the speech signal.

LPC Synthesis Filter

The short-term power spectral envelope of the input speech signal can be depicted by the all-pole synthesis filter which is expressed as (Rabiner and Schafer, 1978):

$$H(z) = \frac{1}{A(z)} \tag{7}$$

where $A(z)$ is the LPC analysis filter and $H(z)$ is the LPC synthesis filter.

Equation 7 is the basis for the LPC analysis model. The LPC synthesis model on the other hand consists of an excitation source $E(z)$, which gives input to the spectral shaping filter $H(z)$, which provides the synthesized output speech $S(z)$ (Suman, 2014):

From equation (6)

$$\frac{E(z)}{s(z)} = A(z) \tag{8}$$

Putting eqn. (6) in eqn. (7)

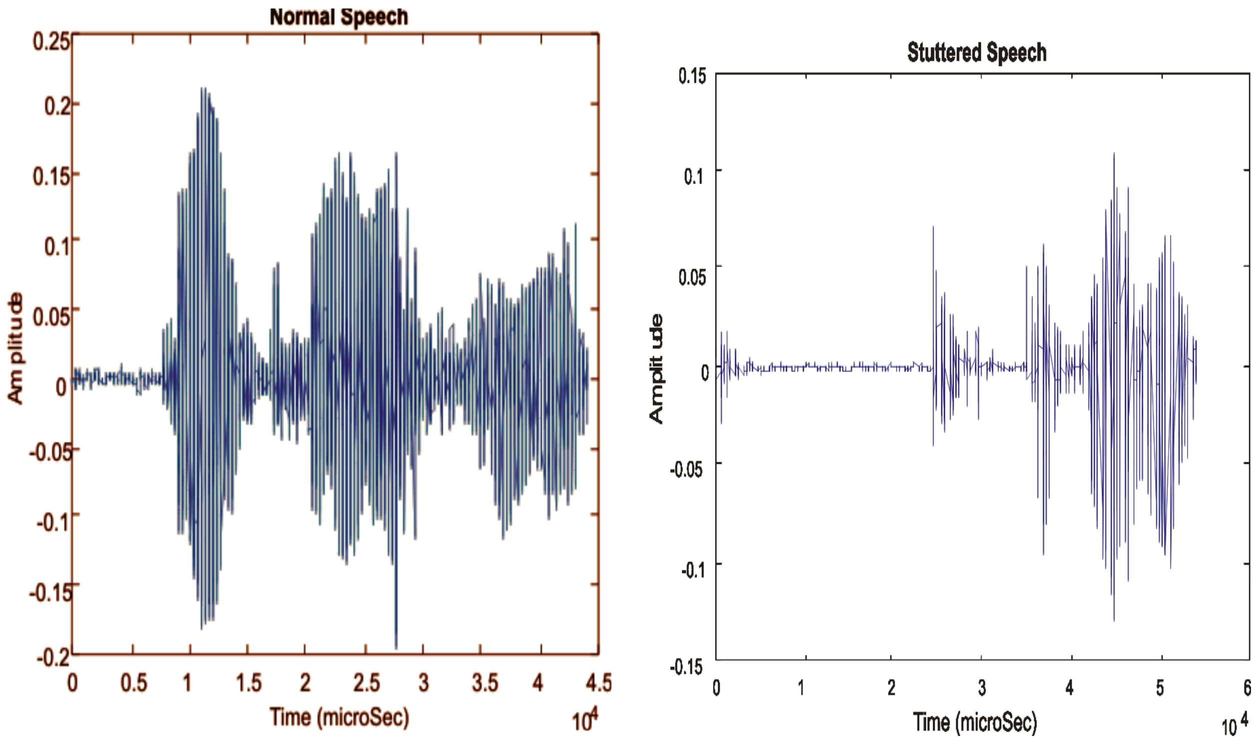
$$H(z) = \frac{1}{\frac{E(x)}{S(x)}} \tag{9}$$

$$S(z) = H(z).E(z) \tag{10}$$

In order to identify voiced or unvoiced sound, the LPC analysis of each frame acts as a decision-making process. The impulse train is used to signify voiced signal, while white noise is used to represent unvoiced frame. Consequently, either impulse train or white noise becomes the excitation of the LPC synthesis filter. Hence, it is essential to highlight the gain, pitch and coefficient parameters that will be fluctuating with time and from one frame to the other. The above model in equation 10 is called the LPC model (Jones *et al*., 2009; Suman, 2014).

RESULTS AND DISCUSSION

Figure 2a and 2b show the speech waveform of the word ‘*department*’, for both the normal pronunciation and the stuttered pronunciation. The speech samples are from two different speakers pronouncing the word department. Some of the speakers in the database read the same passages, making it easy to get the waveform for the normal and stuttered speech. Blocking stuttering and reciprocating sttttereing are the type of stuttering present in Figure 2b and pronounced as ‘*d-----d-department*’. There is about 2×10^4 microseconds block in the pronunciation. Subsequently, there was another short block of about 0.5×10^4 microseconds. The syllable ‘*d*’ was repeated two times before the word was eventually pronunced. The silence observed at the beginning of Figure 2a is the normal inter-word silence which is expected to be a maximum of one second for normal speech.



(a) (b)

Figure 2: Pronunciation of the word ‘department’ (a) normal speech (b) stuttered speech

Figures 3 (a-c), 4 (a-c) and 5(a-c) show the PSD estimates of the stuttered words anniversary, department and sales. The cases of each word were considered, namely; before reconstruction, after reconstruction without noise masking and after reconstruction with noise masking. These diagrams give a clearer picture of what has happened in the time domain. The power estimates show that the speech before and after reconstruction without noise masking are very similar.

Distinct peaks in Welch PSD estimates indicate points of periodicity which in the stuttered speech can be assumed to be some of the points where speech sounds are repeated. Five of these distinct peaks were randomly selected for each of the stuttered word. Out of these five peaks, the most distinct peak is the first peak selected. In addition to these five peaks, the lowest point on each of the plots was also identified and indicated in the plot. From Figures 3 (a and b), Figure 4 (a and b) and Figure 5 (a and b), it would be observed that without the addition of a masking signal (white noise), the reconstructed speech is almost the same with the original speech before reconstruction. The slight differences occur as a result of approximation of values during the reconstruction process from the residual (S(z) = H(z).E(z)).

Considering Figures 3 (b and c), Figure 4 (b and c) and Figure 5 (b and c), the power of the first peak remain relatively the same and no visible effect of the white noise is seen. But from the second peak, some slight changes set in. There are no significant changes in the second peak for stuttered words anniversary and department because the peaks are situated below 5 kHz where the effect of the white noise is just beginning. However, there is significant reduction in the power of the second peak for the word sales as this peak is located beyond 5 kHz. Furthermore, there is a significant increase in the power of the lowest frequency point for each of the words after reconstruction with noise masking. This is because white noise being a random noise tends to reduce the power of distinct peaks and increase the power of the lowest point in its effective area. Therefore, all the repeated speech in the masked region has reduced power and are not likely to be heard by the speaker during playback of the speech. The implication is that not all the repetition in the stuttered speech would be heard by the speaker during speech playback.

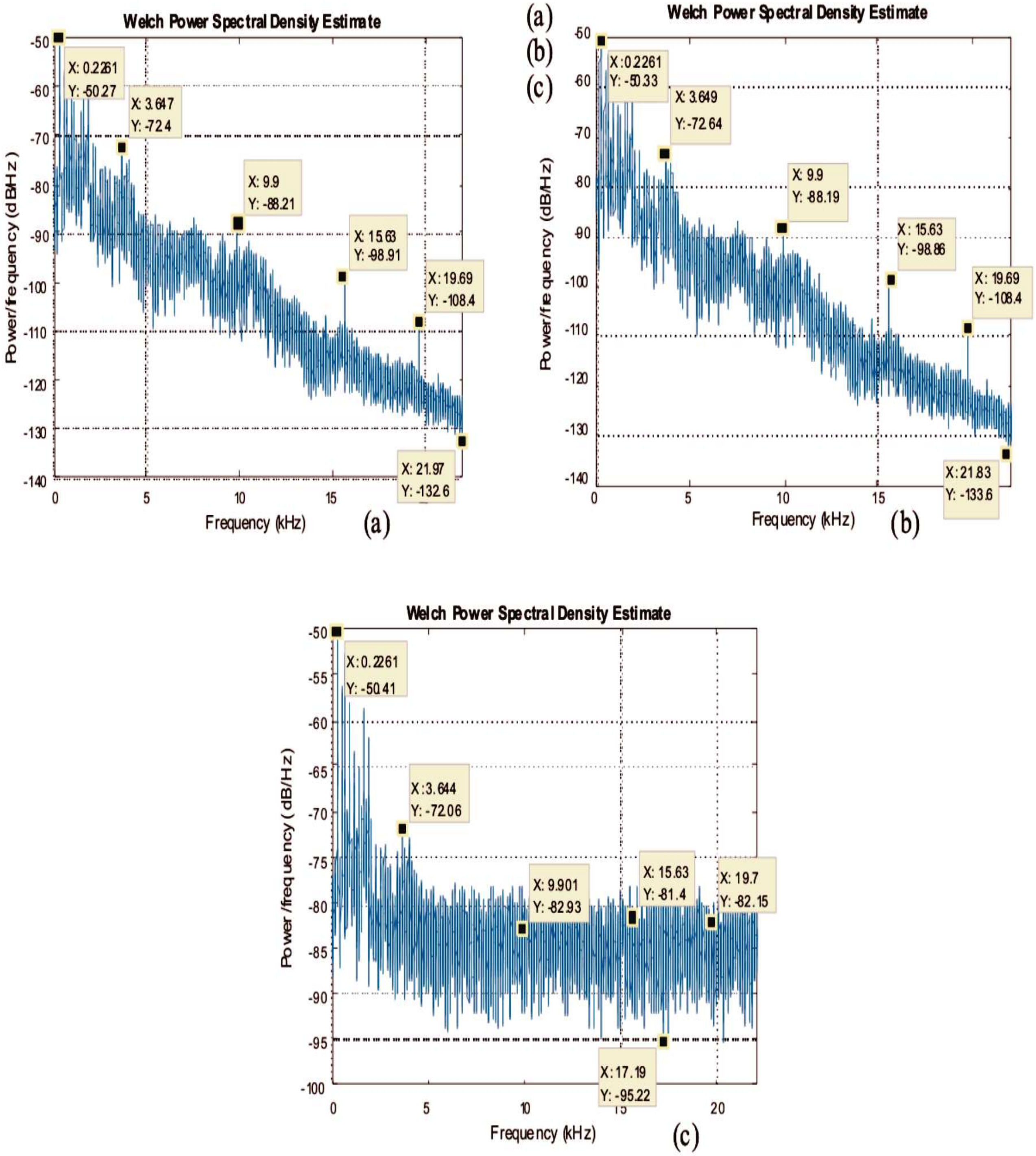


Figure 3: PSD of anniversary (a) before reconstruction (b) after reconstruction without noise masking (c) after reconstruction with noise masking

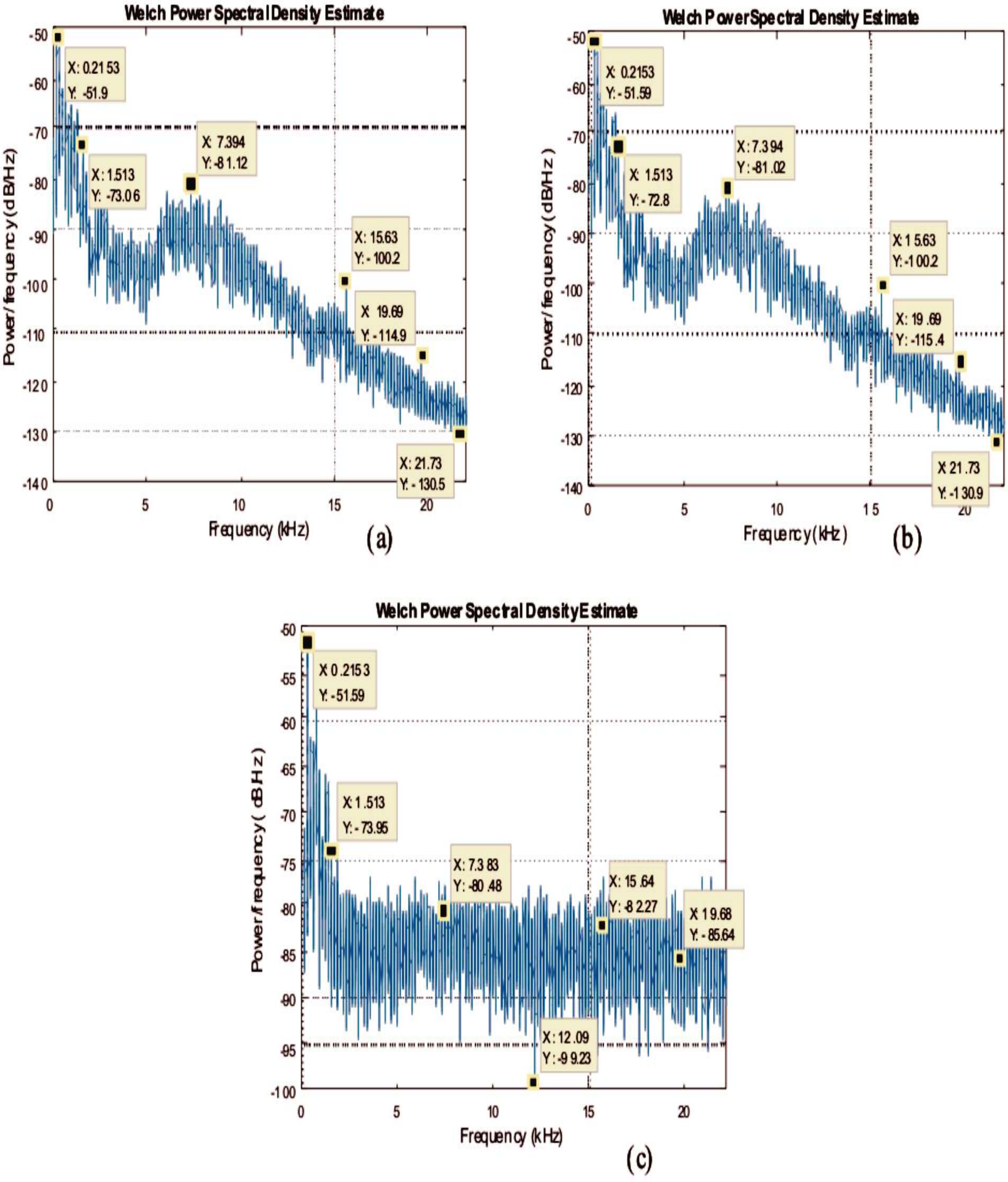


Figure 4: PSD of department (a) before reconstruction (b) after reconstruction without noise masking (c) after reconstruction with noise masking

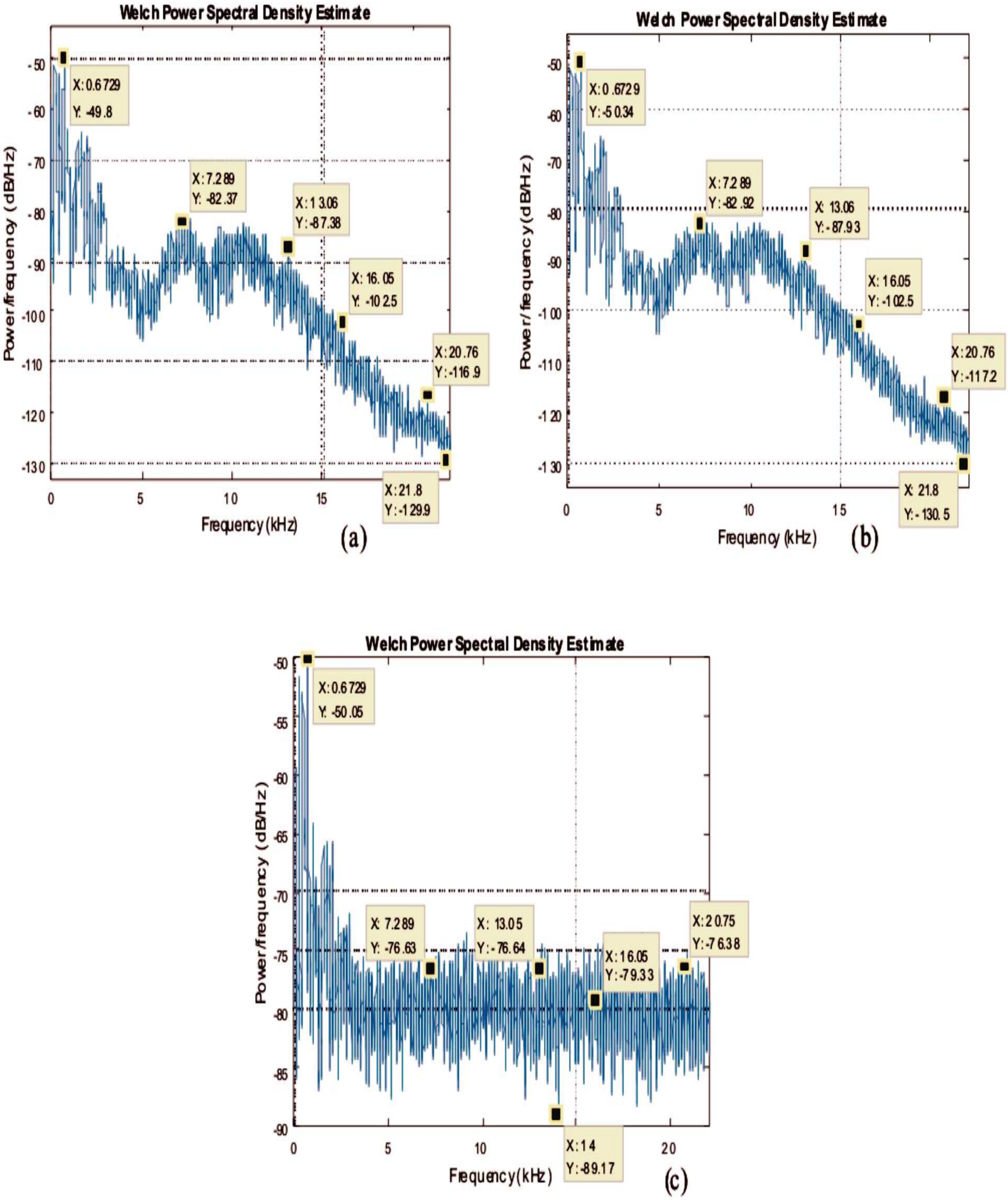


Figure 5: PSD of sales (a) before reconstruction (b) after reconstruction without noise masking (c) after reconstruction with noise masking

CONCLUSIONS

LPC analysis-synthesis algorithm effectively and efficiently reproduced speech by the process of reconstruction. The added white noise effectively reduced the power (dB) of most of the periodicity observed in the speech signal. The effects of the white noise can be visibly seen from the Welch PSD estimates from about 2 kHz forward for the three stuttered words considered. It could therefore be concluded that the added white noise as a mask effectively reduces the repetitions and by extension the stuttering in the speech.

REFERENCES

Acton, C. (2004). A conversation analytic perspective on stammering: Some reflections and observations. *Stammering Research*, *1*(3), 249–270.

Alim, S. A. (2017). Development of Stuttered Speech Reconstruction System, an unpublished thesis at the Department of Mechatronics Engineering, International Islamic University Malaysia.

Awad, S. (1997). The application of digital speech processing to stuttering therapy. In *IEEE Sensing, Processing, Networking, Instrumentation and Measurement Technology Conference, IMTC 97* (pp. 1361–1367).

Chee, L. S., Ai, O. C., Hariharan, M. and Yaacob, S. (2009a). MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA. In *2009 IEEE Student Conference on Research and Development (SCOReD)* (pp. 146–149).

Chee, L. S., Ai, O. C., Hariharan, M. and Yaacob, S. (2009b). Automatic detection of prolongations and repetitions using LPCC. In *International Conference for Technical Postgraduates 2009, TECHPOS 2009* (pp. 1–4).

Conture, E. G. and Yaruss, J. S. (2002). Treatment Efficacy Summary. *American Speech-Language Hearing Association*, (1993), 20850.

Djebbar, F., Abed-Meraim, K., Guerchi, D. and Hamam, H. (2010). Dynamic energy based text-in-speech spectrum hiding using speech masking properties. In *2010 2nd International Conference on Industrial Mechatronics and Automation (ICIMA)* (pp. 422–426).

Hariharan, M., Chee, L. S. and Yaacob, S. (2012). Analysis of infant cry through weighted linear prediction cepstral coefficients and Probabilistic Neural Network. *Journal of Medical Systems*, *36*(3), 1309–15. http://doi.org/10.1007/s10916-010-9591-z.

Hollingshead, K. and Heeman, P. (2004). *Using a uniform-weight grammar to model disfluencies in stuttered read speech: a pilot study*. *Center for Spoken Language Understanding*. Oregon.

Jones, D., Appadwedula, S., Berry, M., Haun, M., Janovetz, J., Kramer, M. and Wade, B. (2009). Speech Processing: Theory of LPC Analysis and Synthesis. *Connexions. June*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4525504.

Rabiner, L. and Schafer, R. (1978). *Digital processing of speech signals*. (A. V. Oppenheim, Ed.). Prentice-Hall.

Seitkulov, Y., Boranbayev, S., Yergaliyeva, B., Davydov, G., and Patapoviche, A. (2014). Rationale for the method of formation of the combined speech masking signals. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). IEEE.

Suman, M. (2014). *Enhancement of compressed noisy speech signal*. Koneru Lakshmaiah Education Foundation. Retrieved from http://shodhganga.inflibnet.ac.in/handle/10603/25341.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.9427and rep=rep1andtype=pdf

Manjula, G. and Kumar, M. (2014). Stuttered Speech Recognition For Robotic Control. *International Journal of Engineering and Innovative Technology (IJEIT)*, *3*(12), 174–177.

Mansour, I. and Al-Abed, S. (2010). A New Architecture Model for Multi Pulse Linear Predictive Coder for Low-Bit-Rate Speech Coding. *Dirasat: Engineering Sciences*, *33*(2).

Oliveira, C., Cunha, D. and Santos, A. (2013). Risk factors for stuttering in disfluent children with familial recurrence. *Audiology-Communication Research*, 18(1), 43–49.

Qi, Y., Wang, H. and Yuan, J. (2008). Speech Information Hiding Method Based on Itakura-Saito Measure and Psychoacoustic Model. In *IEEE International Conference on Networking, Sensing and Control, (ICNSC)* (pp. 1739–1742).

Voigt, T., Hewage, K. and Alm, P. (2014). Smartphone support for persons who stutter. In *13th international symposium on Information processing in sensor networks* (pp. 293–294).

You, C., Rahardja, S. and Koh, S. (2007). Audible noise reduction in eigendomain for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *15* (6), 1753–1765.

Zhang, J., Dong, B., and Yan, Y. (2013). A Computer-Assist Algorithm to Detect Repetitive Stuttering Automatically. In *2013 International Conference on Asian Language Processing (IALP)*, (pp. 249–252).

Zhang, X., Demuynck, K. and Van hamme, H. (2010). Histogram equalization and noise masking for robust speech recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (pp. 4578–4581). Dallas, Texas: IEEE.