International Journal of Electrical and Electronics Engineers
Vol. No.8 Issue 01, January-June 2016
www.arresearchpublication.com
IJEEE
ISSN 2321 - 2055

VARIOUS METHODS FOR SPEECH INTELLIGIBILITY ENHANCEMENT: A BRIEF SURVEY

Shaveta Verma[1], Anil Garg[2]

[1]PG Scholar, [2]Associate Professor, ECE Department, Maharishi Markandeshwar University,(India)

ABSTRACT

Speech is essential, most effective, decisive, reliable and common medium to interact in real time systems. The speech signal is always degraded by noise in speech communication. Enhancing the speech signal which is degraded by noise is done by many speech enhancement algorithms. There are so many different applications of speech still so far from reality due to the lack of dynamic and predictable noise removal techniques for improving and retaining the intelligibility of speech signals. In this paper, attempt has been made towards evaluating the methodologies for Mel cepstral based speech intelligibility methods, real-time voice conversion methods and many other approaches and analyzing the drawbacks and impact on noise-suppression speech.

Keywords: Deep autoencoder, Dynamic range compression, Glimpse proportion, Laryngectomees, Wavelet thresholding.

I. INTRODUCTION

Enhancement of speech is enduring challenge in the enclosure of speech and speech processing. The dominant part of the interaction among humans is being done through speech communication [1]. So, the problem is with the growing number of operations in real-life situations, listeners are unavoidable to hear speech in noise. In order to accomplish the speech enhancement some of the previous techniques are proposed to remove noise from speech signal. Humans change their speaking way when in noisy environment and the speech produced is loud known as Lombard speech [2, 3]. By enhancing the clean speech signal the intelligibility of speech in noise increases, by modifying the Mel Cepstral coefficient on the basis of Glimpse proportion measure [3]. In case of library, people face problems while speaking in silent conditions as the noise disturb rest of the members and some people known as laryngectomees, whose larynx has been removed. NAM and EL speech is produced on real-time DSP systems to improve intelligibility [4]. Noise can also be reduced by evaluating noise-dependent method based on glimpse proportion and noise-independent method based on dynamic range compression [5]. Evaluating the intelligibility benefits in known noise conditions to modify speech [6]. Noisy speech signals are also enhanced by optimization techniques [7, 8]. Deep autoencoders are trained to minimize the noise from speech signal [9]. Multi-layer deep neural networks are used for enhancement of speech signal [10]. Degradation of speech signal due to noise is problem emerging in mobile phone systems. The main task of speech enhancement is to improve intelligibility of the signal, recognized by human listeners. A hybrid filter is proposed to filter the contaminated signals [11]. By comparing the envelopes of critical-band amplitude of processed speech or noisy signal, intelligibility of speech distorted by both stationary and non-stationary noises can be predicted [12]. An improved speech enhancement algorithm established on a novel expectation-maximization (EM) framework operates well when the speech is gnarled by the non-stationary noises [13]. Two new channel variable step size- forward backward (VSS- FB) algorithms used for speech enhancement and noise reduction [14]. Another process in which additive noise and late reverberation can be jointly taken into account in the near-end speech enhancement process and this process improves speech intelligibility, during speech signals are vulgarized simultaneously by additive noise and reverberation [15].

II. LITERATURE REVIEW

In this section, we are discussing the research work explaining various methods for speech intelligibility enhancement in speech processing field.

2.1 Speech Intelligibility Enhancement By Modifying Mel Cepstral Coefficient With Glimpse Proportion Using Lombard And Clean Speech Signals

By enhancing the clean speech signal it is expedient to increase the intelligibility of speech in noise signal. In this paper the effects of modifying the spectral envelope of synthetic speech is determined according to the environmental conditions [2]. To accomplish this, mel cepstral coefficients are modified according to an intelligibility measure i.e. the Glimpse Proportion measure. This method is then evaluated against a topline voice trained with Lombard speech as well as natural speech and baseline synthetic voice trained only with normal speech [2]. The intelligibility of these voices was calculated at three different levels when mixed with speech-shaped noise and with a competing speaker. As compared to normal voices, the Lombard voices, both synthetic and natural were more intelligible in all conditions. In case of speech shaped noise, the schemed modified voice was as intelligible as the Lombard voice which is synthetic in nature without demanding any recordings of Lombard speech which is basically hard to obtain. However, prevailing the case of opposing talker noise, the proposed modified noise was less intelligible as compared to Lombard synthetic voice [2]. Another method to enhance speech intelligibility from synthetic noisy speech signal using Hidden Markov Model. Keeping the speech energy fixed Mel Cepstral coefficients are altered by using Glimpse proportion. The acoustic analysis unfolds that how the conversation is changed and usually increased within the region 1-4 KHz, especially pertaining to vowels, nasals as well as approximants [3]. Natural speech almost completely good as HMM-generated synthetic speech in quiet listening environment. Synthetic speech which is not modified reduce intelligibility to a much greater extent as compared to unmodified natural speech in noisy conditions. So, by modifying synthetic speech using acoustic features or statistical models, generated synthetic speech is more intelligible in noise as compared to natural speech known as Lombard speech. Lombard speech is generate by a talker who is continuously listening to noise [3]. So, Glimpse proportion measure is used and maximized by extracting Cepstral coefficients. Glimpse model is based on information extracted from spectro-temporal regions where speech is less distorted. HMM-generated speech in noise when modified gives intelligibility gain same as when Glimpse proportion measure correlates with subjective scores of natural speech in noise gives similar results [2]. In that case fundamental frequency and spectral tilt are modified to compete with Lombard speech properties. In all these scenarios, Glimpse proportion performed better when compared to other measures. By comparing the levels of distortion and speech glimpse proportion predict only the effect of additive distortion but for noisy speech signals it cannot predict the effect of modifying speech on intelligibility. Results employing speech-shaped noise masker indicate that modified speech is intelligible as compared to synthetic voice trained on normal speech than altered to Lombard speech. Intelligibility gains are moderate for both the proposed process and for alteration to Lombard speech [3].

2.2 To Increase Speech Intelligibility by Digital Signal Processing

A digital signal processor (DSP) implementation for silent speech enhancement and electrolaryngeal speech enhancement established on real-time statistical voice conversion (VC). As in case of library people have trouble speaking in quiet conditions as the sound effect annoy others and people also facing problems to produce a natural voice after undergoing through surgery to remove speech organs. New technologies have been developed to break these boundaries so non-audible murmur (NAM) have been proposed as a silent speech interfaces so that people are capable of talking while keeping silent and electrolaryngeal (EL) speech is composed by an alternative speaking method for laryngectomees whose larynx has been removed through surgery to treat laryngeal cancer. After all the sound quality of NAM and electrolaryngeal speech experience lack of naturalness. VC has proven to be one of the encouraging approaches to address this problem and this has been strongly implemented on devices with plenty of computational resources. Two speech enhancement systems established on real-time VC, one from NAM to whispered voice and other from electrolaryngeal speech to a natural voice are to be implemented based on several methods for preserving conversion accuracy and reducing computational cost [4].

2.3 Improving Intelligibility In Noise By Different Methods

Some noise independent methods followed by dynamic range compression based on different spectral shaping techniques and noise dependent methods based on glimpse proportion measure. Noise- independent methods like spectral tilt flattening and formant enhancement, boosting of the consonant-vowel power ratio and manipulation of duration and prosody they work on the phenomena observed in human speech production [5]. Either for natural or TTS voices it remains unknown to what extend it manipulate the spectro-temporal characteristics of the masker is useful in noise-dependent methods. In this paper it was observed that the best results of the modifications on natural speech was provided by noise-independent spectral shaping with Dynamic Range Compression (SSDRC) despite noise-dependent, did not perform as well. It was also noticed that the noise-dependent approach when applied to TTS voice was as intelligible as a Lombard-adapted voice but not intelligible as natural speech in some stationary noise conditions. In this paper it appraise seven TTS voice styles i.e. two-dependent methods, glimpsed-optimised text-to-speech and speech intelligibility index optimisation (TTSGP and OptSII), two-independent methods, spectral shaping with dynamic range compression and extended spectral shaping (SS-DRC and SSE-DRC) and two combinations of these (TTSGP-DRC and TTSGP-SS-DRC) methods. When TTSGP and TTSGP-DRC are compared, the effect of DRC have been seen as increased gain on stop and fricatives and gain reduction on vowels. Therefore, DRC is reallocating energy of frames to such a degree to increase loudness of the unvoiced parts of speech. Noise-independent unimodal spectral gain combined with DRC i.e. TTR-SS-DRC is one of the most effective strategy in SSN. Using a stationary masker, a noise-independent approach adequately increased intelligibility and the performance exploit in case of the competing talker [5]. A technique to enhance natural and synthetic speech which desires to improve intelligibility in noise. The present study correlate the benefits of speech modification algorithms [6]. In this paper the result of the large-scale evaluation of speech production strategies are constructed in way to increase intelligibility in noise except any change in overall signal-to-noise ratio. Plain natural speech is always more intelligible than synthetic speech but modified synthetic speech decreased this deficit by a significant amount [6].

2.4 Speech Enhancement By Using Optimization Techniques

A technique based on natural flower pollination process. Its changing characteristics can be passed on to arrange new optimization algorithms named Flower Pollination algorithm inspired by flower pollination method. The main desire of a flower is basically reproduction which is done by pollination. In this paper, ten test functions are used to verify the new algorithm and the performance of particle swarm optimization and genetic algorithm are being compared with this algorithm. Simulation result shows that the flower pollination algorithm is more efficient as compared to both PSO and GA. The reasons that FPA is more efficient are flower constancy and long-distance pollinators. Pollinators like honeybee, bats, birds they can travel long distance and can in quire into larger areas and flower constancy guarantees the convergence of same species more quickly. In this algorithm it was assumed that each flower produce only one gamete for simplicity. Despite, by assigning each plant with multiple flower or having each flower with multiple gametes have many advantages like multi-objective optimization, graph colouring and image compression [7]. The flower algorithm is extended to deal with multi-objective optimization engineering problems. It uses weighted sum method including arbitrary weights. Multi-objective optimization has other difficult issues like inhomogeneity, time complexity and dimensionality as compared to single-objective optimization. To outline the Pareto front precisely for single-objective optimization is very moderate and there is no assurance that these explanation points will assign uniformly on the front. There are entirely a few access to negotiate multi-objectives using algorithms so that single-objective optimization problems have been tested. Maybe the straightforward way is to practice a weighted sum to fuse all multiple objectives within a complex single objective. In order to achieve Pareto front exactly same with solutions constantly distributed on the front, random weights have to be used that can be strained from a uniform distribution. For a set of test functions, the multi-objective flower pollination algorithm exactly find the Pareto fronts. Arithmetic experiments and design criterion have displayed that multi-objective flower pollination algorithm is very efficient along with exponential convergence rate [8].

2.5 Effect of Different Denoising Filtering Techniques on Speech Enhancement

In this, it introduced a detailed denoising process in training the deep autoencoder (DAE). Despite, the DAE was trained by applying only clean speech for speech enhancement and noise reduction. In training the DAE, greedy layer-wised pretraining is still selected plus fine tuning strategy [9]. In pretraining, each layer in the DAE is trained as one hidden layer neural autoencoder (AE) employing noisy-clean speech pairs as input and output. In fine tuning stage, in initial system parameters are fixed as the parameters achieved from pretraining stage. When noisy speech is given, the competent DAE is used as a filter for speech estimation. Speech enhancement experiments were done and performance of noise reduction, perceptual evaluation of speech quality criteria (PESQ) and speech distortion are evaluated. The DAE is also compared with minimum mean square error algorithm, then the proposed deep autoencoder (DAE) maintain remarkable performance [9]. A multi-layer deep architecture occupying a regression-based speech enhancement scheme using deep neural networks (DNN). In DNN learning mechanism, a huge set of training data provide a powerful modeling capability to measure the complicated nonlinear mapping observed from noisy speech signals to desired clean speech signals [10]. The DNN is efficient of capturing context information which is used to improve the performance of speech which is separated from background noises observed in speech enhancement algorithms. The DNN-based algorithm tend to gain significant improvement when compared with log minimum mean square error algorithm [10]. An adaptive fuzzy wavelet filter that is established on a fuzzy inference system for upgrading speech signals and as well as bettering the accuracy of speech recognition. The main ambition of an automated speech recognition system is to boost the recognition rate though for a communication system is to enhance the signal-to-noise ratio of distorted speech. The primitive wavelet thresholding algorithm has been completely used for noise filtering. In this suggested method, adaptive wavelet thresholds are achieved and composed according to the fuzzy rules about the presence of speech in contaminated signals. In adaptive fuzzy wavelet filter, the link between speech and noise are compiled into seven fuzzy rules used applying on four linguistic variables, which are used to resolve the state of a signal. A hybrid filter is a mixture of an adaptive fuzzy wavelet filter and the spectral subtraction method which are used to filter the corrupted signals. A hybrid filter is arranged to enhance the performance by using amplified voice activity detector when the signal-to noise ratio is less than 5 dB. This method effectively increases the speech recognition rate and SNR [11].

III. CONCLUSION

In this paper, various methods has been studied that are used for speech intelligibility enhancement. Some techniques perform well in non-stationary environments and some in stationary environment because it is challenging to estimate different types of noise and there time variations and therefore complete noise cancellation is unattainable. It is concluded that by modifying the Mel cepstral coefficient, noise reduces and glimpse proportion measure becomes effective and by comparing the levels of distortion and speech signal, SNR predicts the effect of additive noise but it does not predict the effect that modifying speech has on intelligibility of noisy signal. As an optimization criterion glimpse proportion measure need to limit the distortion generated by modifications. So, by improving the glimpse proportion measure, intelligibility of the modified speech can be predicted. For real-time voice conversion systems, the experimental results prove that DSP systems have the effectiveness to significantly enhance NAM and EL speech and run in real-time with less degradation. A hybrid filter is used to filter the corrupted signals and this method effectively increases the SNR and speech recognition rate.

REFERENCES

[1] Philipos C. Loizou, *speech enhancement: theory and practice* (CRC Press, Taylor and Francis, Boca Raton, FL, 2007).

[2] Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise, *SAPA-SCALE Workshop on Statistical and Perceptual Explorer, 2012.*

[3] Cassia Valentini-Botinhaoa, Junichi Yamagishia, Simon Kinga and Ranniery Maia, Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion, *Computer Speech and Language, 28,* 2014, 665-686.

[4] Takuto Moriguchi, Tomoki Toda, Motoaki Sano, Hiroshi Sato,Graham Neubig, Sakriani Sakt1, Satoshi Nakamura, A Digital signal processor implementation of Silent/Electrolaryngeal Speech Enhancement based on Real-Time Statistical Voice Conversion, *INTERSPEECH,* 2013, 3072-3076.

[5] Cassia Valentini-Botinhao, Elizabeth Godoy, Yannis Stylianou, Bastian Sauert Simon King and Junichi Yamagishi, Improving Intelligibility in noise of HMM- generated speech via noise-dependent and independent methods, *International Conference on Acoustics, Speech and Signal Processing,* 2013.

[6] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou,Bastian Sauert, Yan Tang, Evaluating the intelligibility benefit of speech modification in known noise conditions, *Speech Communication*, *55*, 2013, 572-585.

[7] Xin-She Yang, Flower Pollination Algorithm For Global optimization, *Unconventional Computation and* Natural Computation in Computer Science, 7445, 2013, 240-249.

[8] Xin-She Yang, Mehmet Karamanoglu, Xingshi He, Multi-objective Flower Algorithm for Optimization, *Procedia Computer Science*, *18*, *ICCS* 2013, 861-868.

[9] Xugang Lu, Yu Tsao, Shigeki Matsuda, Chiori Hori, Speech Enhancement based on Deep Denoising Autoencoder, *ICASSP*, *INTERSPEECH* 2013, 436-440.

[10] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, An experimental study on Speech Enhancement based on Deep Neural Networks, *IEEE SIGNAL PROCESSING LETTERS, 21(1),* 2014, 65-68.

[11] Chih-Chia Yao, Ming-Hsun Tsai and Yuan-Tain Chang, On the use of Adaptive Fuzzy Wavelet Filter in the Speech Enhancement, *JOURNAL OF COMPUTERS, 9(11),* 2014, 2501-2513.

[12] Jesper Jensen and Cees H.Taal, Speech Intelligibility Prediction based on Mutual Information, *IEEE Trans. Audio, Speech, Lang. Process., 22(2), February 2014, 430-440.*

[13] Daniel P.K.Lun, Tak-Wai Shen and K.C.Ho, A Novel Expectation-Maximization Framework for Speech Enhancement in Non-Stationary Noise Environment, *IEEE Trans. Audio, Speech, Lang. Process., 22(2), 2014, 335-346.*

[14] Redha Bendoumia, Mohamed Djendi, Two-channel variable-step-size forward and backward adaptive algorithms for acoustic noise reduction and speech enhancement, *ELSEVIER, Signal Processing, 108, 2015 226-244.*

[15] Richard C. Hendriks, João B. Crespo, Jesper Jensen, and Cees H. Taal, Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation Under an Approximation of the Short-Time SII, *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 23(5), May 2015, 851-862.*