NOISE ENHANCEMENT METHOD FOR SPEECH SIGNAL PROCESSING

Muhamad Arif Hashim and Roshidi Din
University Utara Malaysia, 06010 Sintok, Kedah

Abstract: Speech recognition is like biometric (eg. smartcard) that based on person voice (speech) which is convert to signal, recognize whether the intended person is genuine or not. Basically, signal consists two kind of information: message information and noise information and as well as speech signal, it contains speech information and noise information. Generally, noise information is not needed and causes trouble by disturbing the speech signal and adds errors to the signal. To overcome this problem, speech enhancement method is used to separate speech information and noise from speech signal using several kinds of techniques such as Kalman Filtering and Signal-to-Noise ratio (SNR) and enhance the intended feature (speech information). In this paper, our main intention on different method called noise enhancement method, which based on speech enhancement method to recognize the intended speech signal and learning noise behaviors and its usefulness. The relationship between noise and speech signal will be discuss including chronology of speech signal processing and types of noise that appears. We state steps on doing the noise enhancement method and some mathematical basis will be presented. Finally, we will discuss challenge that occur and conclude it based on based on the discussion. Hopefully, it will give a new direction of noise and speech's intention not just in information communication and technology field but also in speech research, sound engineering and other related fields.

Keywords: Speech recognition, Speech enhancement, Noise enhancement, Signal processing

INTRODUCTION

Speech signal is basically human voice (speech) that been convert into signal. This signal maybe an analog signal or digital signal but in this paper, we focusing only on digital signal. Digital signal is a combination between two different kinds of bits: 0 and 1, which in fact, can be corrupted due to effect of noise. In speech recognition area, the translated human voice (signal) is the main focus to recognize a person [1]. So, in the case of sending speech signal through a transmission channel, problems will arises due to effect of noise. As in the signal itself, it contains noise that based on the speech surrounding and it becomes more critical when noise from transmission channel affect it too. Mainly, speech model is applied to recognize a speech based on clean speech signal. So it gives a negative impact if speech signal contains noise. If we still decode the noisy speech signal, it will degrade the quality and intelligibility of the signal [3]. That is why we need to process the signal first, so that it suitable for speech model. Speech enhancement is one of the ways to overcome the problem of corrupted signal. In speech signal processing, the method convert the signal into spectral/spectrum power or spectrogram so that it can easily be preprocess and analyze by subtracting the speech signal from the noise and then enhancement the speech signal to improve it quality. This process is called feature extraction which extracting feature/properties/characteristic from signal. Then it will be match meaning extracted feature will be compare with speech database based on speech model such as Hidden Markov Model (HMM) or vector quantization (VQ) [1]. In analyzing the signal, the comparison can be according to phone syllable, word and sentence after modeling it like in Figure 2 [6].

There are two principal criteria for measuring the performance of a speech enhancement system. The *quality* of the enhanced signal measures its clarity, distorted nature, and the level of residual noise in that signal. The quality is a subjective measure that is indicative of the extent to which the listener is comfortable with the enhanced signal. The second criterion measures the *intelligibility* of the enhanced signal. This is an objective measure, which provides the percentage of words that could be correctly identified by listeners. The words in this test need not be meaningful. The two performance measures are not correlated. A signal may be of good quality and poor intelligibility and vice versa. Most speech enhancement systems improve the quality of the signal at the expense of reducing its intelligibility. Speech enhancement is concerning on processing of noise and corrupted speech to improve the quality and intelligibility of the signal [8]. It is an uncommon case in estimating signal when the human ear does not believe in a simple mathematical error criterion.

Below is a figure of block diagram of human speech production, which shows parts of body that related in making voice (Figure 1).

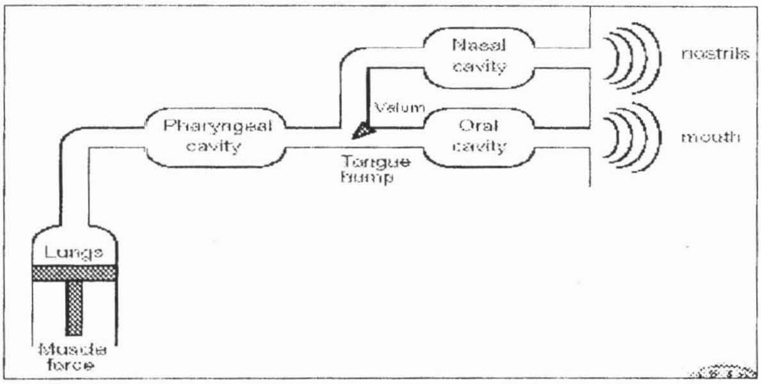


Figure 1: A Block Diagram of Human Speech Production [6]

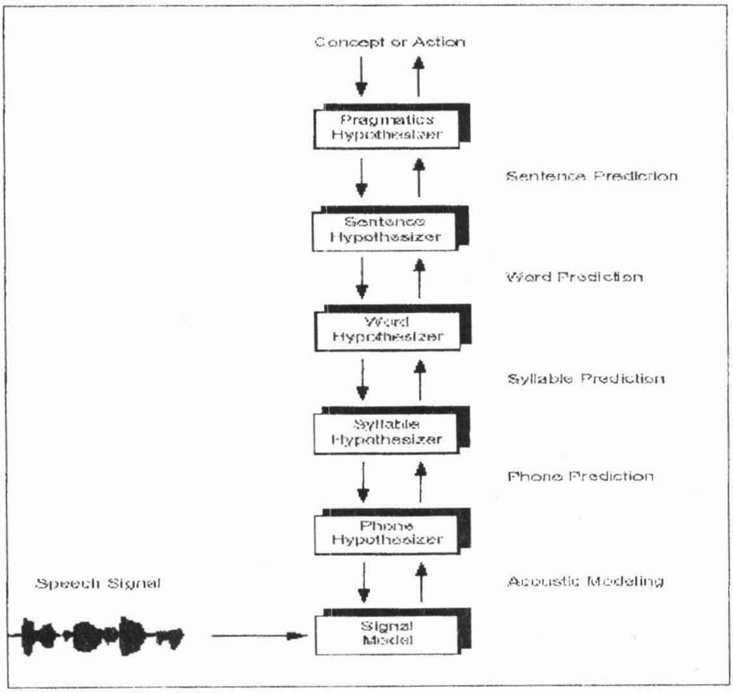
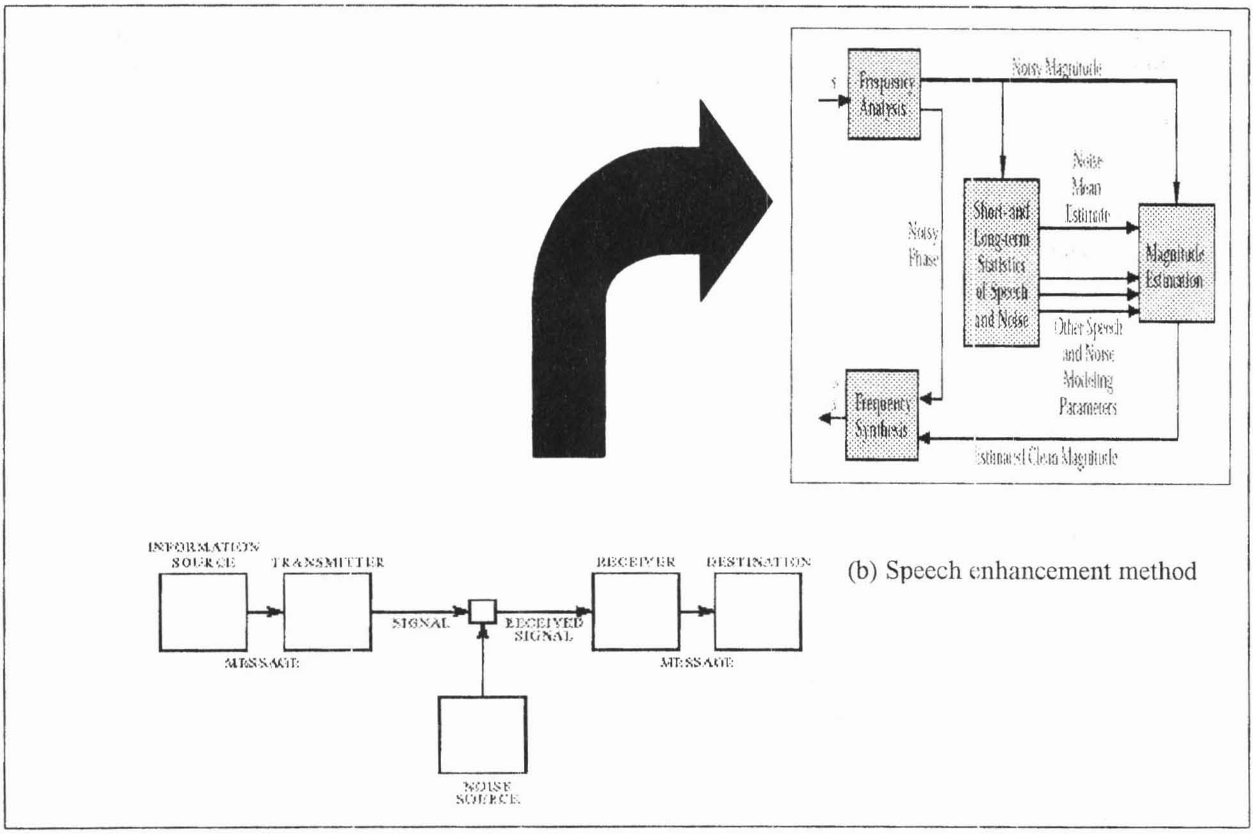


Figure 2: Predicting speech signal [6]

DISCUSSIONS

Chronology of Speech Signal Processing



(b) Speech enhancement method

(a) Model of communication system

Figure 3: Location of speech enhancement in communication system

From human speech production, a speech was created (Figure 1), we extend to a bigger picture where the speech signal is intend to be send to a destination. It went through digital communication system, which we adapted Shannon's systematic communication model (a), where the signal will be in form of digital bit. From information source, which in this case we consider as human speech production, the speech is send to transmitter to be convert into digital signal. The signal will be interrupt by noise when it through transmission channel. This interruption will corrupt the sequence of the signal bits and make it more corrupted than before. As in the speech itself contain noise that based from it's surrounding. It will be process for correction; in our case it will be speech enhancement (b), when it arrives at receiver. The sequence of bit is converting into spectral power or spectrogram to extract it features. Using the features, we can analyze and process for enhancement or recognition purpose.

We extend a little bit of the item in the chronology by explaining noises that occur along the process. Two types of noise that will occur: additive noise and convolutional noise. All condition of degradation of the signal due to ambient acoustic noise is called additive noise. The seriousness of problem will depend on the ratio of Signal-to-Noise ratio (SNR), but also depend on how the speech ' suit' with behaviors of the noise. It corresponds to addition of the time-domain speech and noise signals. While convolutional noise occurs due to changes in the speech signal due to changing room acoustics, changes in microphone etc. It corresponds to convolution with a time-domain signal. Based on literature of this two noise, convolution noise is more difficult to handle than additive noise due to the bursty nature of speech, which usually researcher observe the behaviors of noise during the speech pauses [2].

Noise Enhancement

Combination of multiple recognizers is consistently reported to outperform baseline recognition systems. For example, a hybrid speech recognition system based on the combination of acoustic and word information achieved better word recognition results than the baseline recognition systems. Choices of the level of the combination and the best feature streams to be combined together remain as key issues for successful combination. These choices are currently made through intuition and empirical comparison. An approach for selecting the level of the combination based on conditional mutual information of the feature streams given the underlying phoneme identity is the main focus.

Our approach demonstrate the following steps:

i) Converting speech signal to speech noise power spectrum.
ii) Extracting the noise spectrum from speech noise power spectrum using spectral subtraction.
iii) Measuring the noise spectrum using Signal-to-Noise ratio (SNR) and mutual information measurement based on spectrogram.

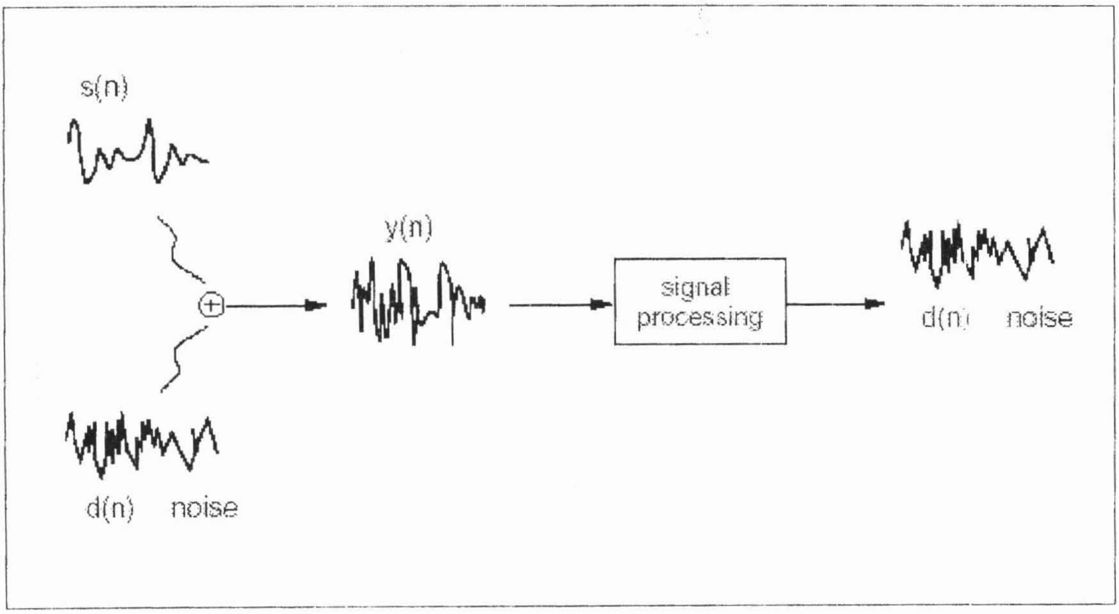


Figure 4: Diagram of proposed noise enhancement method

Thus the goal of speech enhancement is to find an *optimal estimate* (i.e., preferred by a human listener), $\hat{s}(n)$ given a noisy measurement,

$$y(n) = s(n) + d(n) \tag{1}$$

Whereas noise enhancement, we replace speech with noise, giving us an *optimal estimate* for noise, $\hat{d}(n)$ using function (1). The signal will be capture for example in spectrogram and analyze based on $y(n)$ which from it, we can see the pattern of speech $s(n)$ and noise $d(n)$.

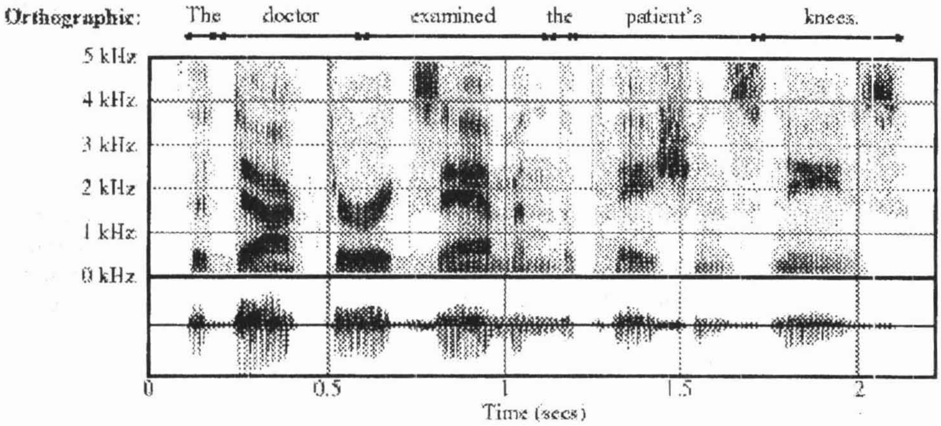


Figure 5: Example of speech spectrogram

Based on the Figure 5, the spectrogram shows a sample of time domain signal of speech based on phonetics. It measures and draws the analysis of every word based on the duration of the pronunciation of the word. From it, SNR and mutual information measurement can be perform to analyze its dependency and pattern of noise. SNR is a ratio of the amplitude of the desired signal to the amplitude of noise signals at a given point in time. Because many signals have a very wide dynamic range, SNRs are often expressed in terms of the logarithmic decibel scale.

Signal-to-Noise ratios are closely related to the concept of dynamic range. Where dynamic range measures the ratio between noise and the greatest un-distorted signal on a channel, SNR measures the ratio between noise and an arbitrary signal on the channel, not necessarily the most powerful signal possible.

Mutual information measures dependence between variables, for example between X and Y. The mutual information between two random variables X, representing the phoneme, and Y, representing the acoustic noise feature, is

$$I(X \mid Y) = \int \sum_{i=1}^{I} p(y \mid x_i) \log \frac{p(y \mid x_i)}{p(y)} dy$$

where I is the number of values that the phoneme set can take, x_i is the ith value of the phoneme. The mutual information is a natural measure of the dependence between random variables. It is always non-negative, and zero if and only if the variables are statistically independent. Thus the mutual information takes into account the whole dependence structure of the variables.

Finally, we hope to get a pattern of noise based from analysis and use it to verify speech signal especially in the case of verification of received speech through communication channel.

Challenges

Coarticulation: Coarticulation in speech is one of the most difficult problems for speech processing area. It usually defined as a change in the acoustic-phonetic content of speech segment due to anticipation or preservation of adjacent segments [5]. Current techniques for statistical modeling used by speech recognition systems (e.g., context-dependent phone models) are believed to be insufficient for capturing all coarticulatory effects. The degree of coarticulation is assumed to vary with contextual conditions, such as differences in speaking rate, stress, etc. In the past, coarticulation has been studied using only limited data sets and using acoustic phonetic methods such as formant analysis.

For the various conditions typically assumed to increase coarticulation (high speaking rate, unstressed syllables, and central/lax vowels), research shows that a corresponding increases in the conditional mutual information. Different kind of form of a speaker or a speech will create a different coarticulation. It brings difficulty in recognizing and verifying process due to different feature that create based on the articulation.

CONCLUSION

In this paper, we had clarified speech signal and its relationship with noise in speech recognition and speech signal processing. The process of separating or subtracting noise from speech signal is the main focus. Our proposed approach called noise enhancement that applies Signal-to-Noise ratio and mutual information measurement to analyze the behavior and pattern of noise in intention from it we can recognize genuinity of the received speech signal. This method promise a high potential compatibility in other information, communication and technology area too for example in extracting noise from digital images or cleansing caller and receiver voice for mobile communication.

REFERENCES

1. Campbell, J. 1997. Speaker Recognition: A Tutorial. *Proceedings of the IEEE.* vol. 85. no. 9

2. Compernolle, D. V. 1992. DSP techniques for speech enhancement. In *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pages 21--30.

3. Ephraim, Y. 1992. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526--1555.

4. Ho, D. K. C. Speech Enhancement: Concept And Methodology. 1998.
[Internet URL: http://meru.cecs.missouri.edu/mm_seminar/ho.pdf].

5. Kirchhoff, K. and Bilmesy, J. A. 1999. Statistical Acoustic Indications Of Coarticulation. *14th Int. Cong. of Phonetic Sciences.*

6. Picone, J. Fundametals Of Speech Recognition: A Short Course. 1996.
[Internet URL : http:// www.isip.msstate.edu/publications/courses/ isip_0000/lecture_notes.pdf]

7. Torkkola, K. 2003. Feature Extraction by Non-Parametric Mutual Information Maximization *Journal of Machine Learning Research 3.* 1415-1438

8. Wan, E. A. and Nelson A. T. 1998. Network for speech enhancement. Handbook of Neural Networks for Speech Processing. 1st Edition. Artech House.