BASIC PROCEDURES OF SPEECH ENHANCEMENT IN AUTOMATIC SPEECH RECOGNITION SYSTEMS

Ing. Matúš Pleva
Department of Electronics and Multimedia Communications,
Technical University of Košice, Park Komenského 13, 041 20 Košice,
Slovak Republic.
Tel: +421-55-6022334
+421-55-6022298
E-mail: Matus.Pleva@tuke.sk

ABSTRACT

This is a short review about basic procedures of speech enhancement and voice pre-processing techniques. Also, this paper is a short description of basic speech enhancement techniques needed by the *Automatic Speech Recognition* (*ASR)* systems as a part of d*ialogue systems* or v*oice portals*. This area is very popular today, because voice communications consecutively substitute the text–based communication.

The *speech enhancement* is a part of speech pre-processing procedures, which prepared the voice signal to speech recognition module. Therefore the signal need to be denoised, hum and other disturbance filtered. Very useful is to detect the speech activity and the noise estimation. The basic rules and ideas used in this area are shortly presented.

1 INTRODUCTION

The basic structure of speech enhancement system is given in the figure. This figure highlights three stages that are necessary in such systems:

1. A spectral analysis/synthesis system
2. Noise estimation algorithm
3. A spectral gain computation algorithm

Let *x[n]* and *d[n]* denote the speech and the noise processes, respectively.

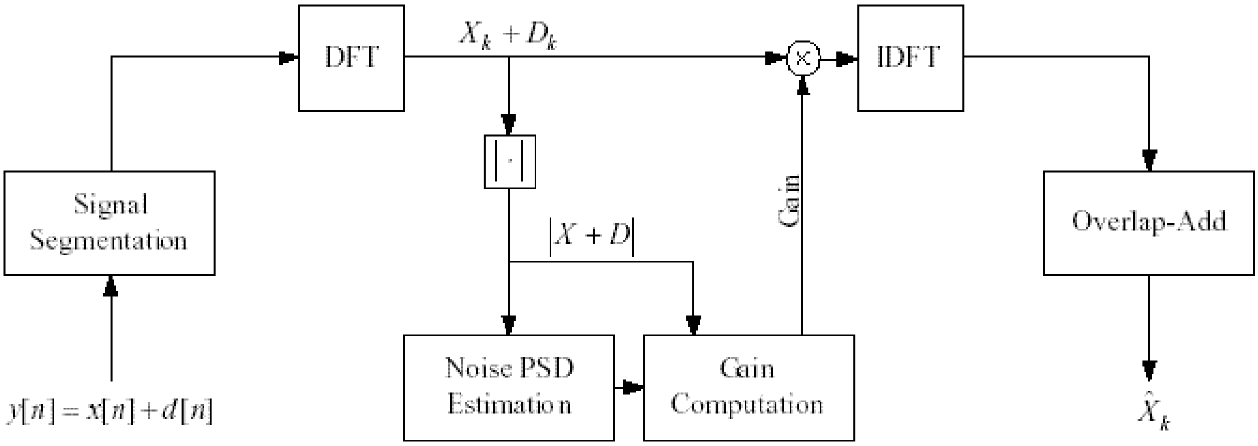
The observed noisy signal, *y[n]*, is given as:

$$y[n] = x[n] + d[n] \quad \text{where } 0 \le n \le N - 1$$

The objective of a speech enhancement block is to estimate *x[n]*. This is generally accomplished on a frame-by-frame basis by applying a unique gain to each of the frames of *y[n]*. These gains are computed (in either frequency or time domain) by minimizing or maximizing a cost function, [1]. For instance, Signal-to-Noise Ratio (SNR) is one such cost function. Defining a gain that maximizes the SNR of the output enhanced signal is one suitable criterion.

2 ASSUMPTIONS ON SPEECH AND NOISE PARAMETERS

In order to derive an estimator for A_k, the a priori probability distribution of the speech and noise Fourier expansion coefficients should be known. Generally, the speech and possibly the noise are neither stationary nor ergodic processes, thereby excluding the convenient possibility of obtaining the statistics of Fourier coefficients by examining the long-term behavior of each process. This problem can be resolved by assuming an appropriate statistical model (for example *Gaussian Statistical Model*).



3 APPROPRIATE ANALYSIS FRAME LENGTH

Long analysis frames are good, as they provide better statistical averages. However, due to time-varying characteristics of speech, a shorter window is sometimes desired. A convenient way to tackle this trade-off is to consider the durations of typical phonemes in speech.

A typical vowel (voiced phoneme) ranges between *50–400 ms*, while a plosive may last for about *10 ms*. Changes in the shape of the speech signal, whether gradual or abrupt, result from movements of the vocal tract articulators, which rarely stay fixed in position for more than *40 ms* at a time, [2]. Thus in most works, the analysis frame length, *T*, is usually chosen in the *16–40 ms range* (or about 128–320 samples, for speech sampled at 8 kHz).

Note that even if a smaller frame length were considered appropriate (to increase temporal resolution at the cost of spectral resolution), statistical independence would still be assumed.

4 INPUT CONDITIONING: HIGH PASS FILTER

Initial pre-processing is needed to condition the input signal against excessive low frequency and other background disturbances that might degrade the quality of a speech codec.

Low frequency disturbances include hum at 60/50 Hz and its harmonics at 120/100, 180/150 and 240/220 Hz. Therefore it is desired to pre-process input narrowband speech (sampling frequency of 8 kHz) with a high-pass filter that will eliminate low frequency noise.

5 VOICE ACTIVITY DETECTOR

Voice Activity Detector (or VAD) returns a ‘1’ in the presence of speech and ‘0’ in absence thereof. Conceptually, such a binary decision is based on some measured or extracted feature of speech compared against some pre-defined (or adaptively changing) thresholds (generally extracted from noise only frames).

A VAD must be robust as its accuracy is critical in noise suppression algorithms. Misclassifying speech as noise will erroneously remove speech or result in a poor estimate of noise.

Some of the early VAD algorithms relied on short-time energy, zero crossing rate, and LPC coefficients. Recent work employs Cepstral features, formant shape and a least-square periodicity procedure, [3].

6 INTUITIVE APPROACH TO NOISE SUPPRESSION

A typical speech file contains several pause (between words) and/or silence (speaker stops talking) segments. In speech enhancement, the noise variance is updated during such pauses or silence.

A VAD is used to discriminate between speech and silence/pause a VAD is employed. For denoising speech files corrupted with additive white Gaussian noise, one straightforward idea would be to subtract off the estimated noise variance, (D_k), from the power spectrum of the observed noisy signal, (Y_k), to obtain an estimate of the modulus of speech power spectrum (X_k). Mathematically this is represented as:

$$|X_k|^2 = |Y_k|^2 - |D_k|^2$$

However, there are limitations to this subtraction rule, and as such the basic problem has been tackled by deriving several fundamentally and theoretically justified noise suppression rules.

7 ANALYSIS WINDOW TYPE

Some of the commonly used windows in speech processing are symmetric (e.g., Hamming and Hanning windows) or asymmetric (such as the hybrid Hamming-Cosine window). The goal of asymmetric windows is to reduce the algorithmic delay in speech coders.

Ideally, the window spectrum should have a narrow main-lobe and small side-lobes. However, there is an inherent trade-off between the width of the main-lobe and the side-lobe attenuation. A wide main-lobe will average adjacent frequency components and large sidelobes will introduce contamination (or spectral leakage) from other frequency regions.

The main lobe for rectangular window is narrower than that of the Hanning window, while its side-lobes are higher. The third mentioned window type is the smoothed Trapezoidal window.

REFERENCES

[1] J. Tlučák, J. Juhár, L. Doboš, A. Čižmár: *Neural Network Based Speech Enhancement*. Radioengineering, Vol.8, No.4, pp.22-25, 1999.

[2] Tarun Agarwal, *Pre-Processing of Noisy Speech for Voice Coders*, Department of Electrical & Computer Engineering, McGill University, Montreal, MSc Thesis, Canada, 2002

[3] K. El-Maleh and P. Kabal, *Comparison of voice activity detection algorithms for wireless personal communications systems*, in Canadian Conf. on Elect. and Comp. Eng., vol. 2, (St. Johns, Canada), pp. 470–473, May 1997.