OLAKUNLE O ANIFOWOSE

Department for Electrical and Computer Engineering

A Review of Speech Enhancement Techniques

Temple University, Philadelphia, PA

[kunleanney@temple.edu](mailto:kunleanney@temple.edu)

**Problem Statement**

Speech enhancement involves the reduction of additive background noise. Speech enhancement is a very special case of signal estimation as speech is non-stationary, and the human ear---the final judge---does not believe in a simple mathematical error criterion. Therefore subjective measurements of intelligibility and quality are required. The goal of speech enhancement is to find an optimal estimate preferred by a human listener of a speech denoted by s(t) when given a noisy version y(t)=s(t)+n(t). Speech quality in cellular radio system are plagued with degradation problems usually cause by acoustical background noise, quantization noise due to encoding at the source, residual bit errors after channel decoding ,residual bit errors after channel decoding and band pass limitation of the speech signal. Speech Enhancement Depends on good signal processing technique, human perceptual factor, speech quality and intelligibility are dependent on short term spectral amplitude and insensitive to spectral phase. Applications of speech enhancement include improving speech quality or intelligibility in multimedia and wireless communications and speech recognition.

**Introduction**

There are several techniques used in speech enhancement to address the problem of degraded speech, for the purpose of analysis in this paper. I will be considering three techniques

1. Speech Noise Reduction using sequential spectrum detection based on modified DFT pair
2. Noise reduction based on sinusoidal speech coding system
3. Speech enhancement using conditional estimations

**1). Speeh noise reduction using sequential spectrum detection based on modified DFT pair.**

In this method the noise added signal is decomposed into frequency signal using a modified DFT. The amplitude spectra for speech and noise are estimated using level detector at every sampled point. A speech spectrum is then obtained sequentially by subtracting the estimated noise spectrum from that of the signal. The noise reduced frequency signals are then summed using a modified IDFT to obtain a speech with reduced noise. The top level diagram of the system can be seen below



Figure 1(System Diagram)

The modified DFT is a simplified version of the DFT. It only requires real value operation and eliminates the imaginary content. It can be represented mathematically as

While its inverse (MIDFT) can be represented as

In order to avoid spectral leakage due to the spectrum side-lobe of a truncation function for DFT analysis, a window function is used on the MDFT which can then be rewritten as

Level Detector

This method makes use of two level detectors, one for calculating the amplitude spectral of the noisy speech and the other for estimating the noise amplitude spectra present in the speech. For the level detector to detect the speech and noise spectra’s certain assumptions are made

* The noise spectrum is spread in wide frequency band and its rapidity of change is very slow
* The speech spectrum is located in lower frequency band and it varies rapidly in comparison to noise

Signal Level Detector

The power spectrum is computed for each sample level and averaged. The power spectrum can be represented as

The power spectrum is then compared with the past estimate of the detector for the amplitude spectra. Below is the schematic for the signal level detector:



Figure 2(Signal level Detector)

If is larger than, the switch (SW) is assigned to “a” which means the output of the detector presently is given by

If is smaller than however, the switch (SW) is assigned to “b” which means the output of the detector presently is given by

Where is a very large time constant.

The signal level detector outputs rapidly change when the amplitude of an input frequency signal increases and output slowly change through the smoothing circuit when the amplitude of the input frequency signal decreases.

Noise level detector

The average power spectrum is computed and used to compute the estimated amplitude noise spectra as :



Figure 3(Noise level Detector)

Where is a time constant which can take on two different values. .If the average power spectrum is greater than the past estimated noise spectrum output of the detector. The larger time constant will be selected and if the average power spectrum is smaller than the past estimated noise spectrum output of the detector the smaller time constant will be selected. The output of the noise level detector doesn’t change rapidly it has a relatively steady value.

Obtaining noise reduced signal.

The concept of spectral subtraction is used here. The idea is to subtract the noise at each sample in the spectral domain. The noise subtraction can be defined as

where are the amplitude spectra of the noise added signal and the noise respectively and is the amplitude spectra of the noise reduced signal. Since the noise and signal level detector might output different level during speech pause the noise subtraction co-efficient is used. The noise reduced spectra is then used to compute the frequency signal as can be seen below

.

Using the MIDFT all the noise reduced frequency signal are summed to produce a noise reduced signal which can be represented as

**2). Noise reduction based on sinusoidal speech coding system**

This technique seeks to reduce substantial amount of noise without adding any noticeable artifacts into the speech signal through a series of step outline in the block diagram below.

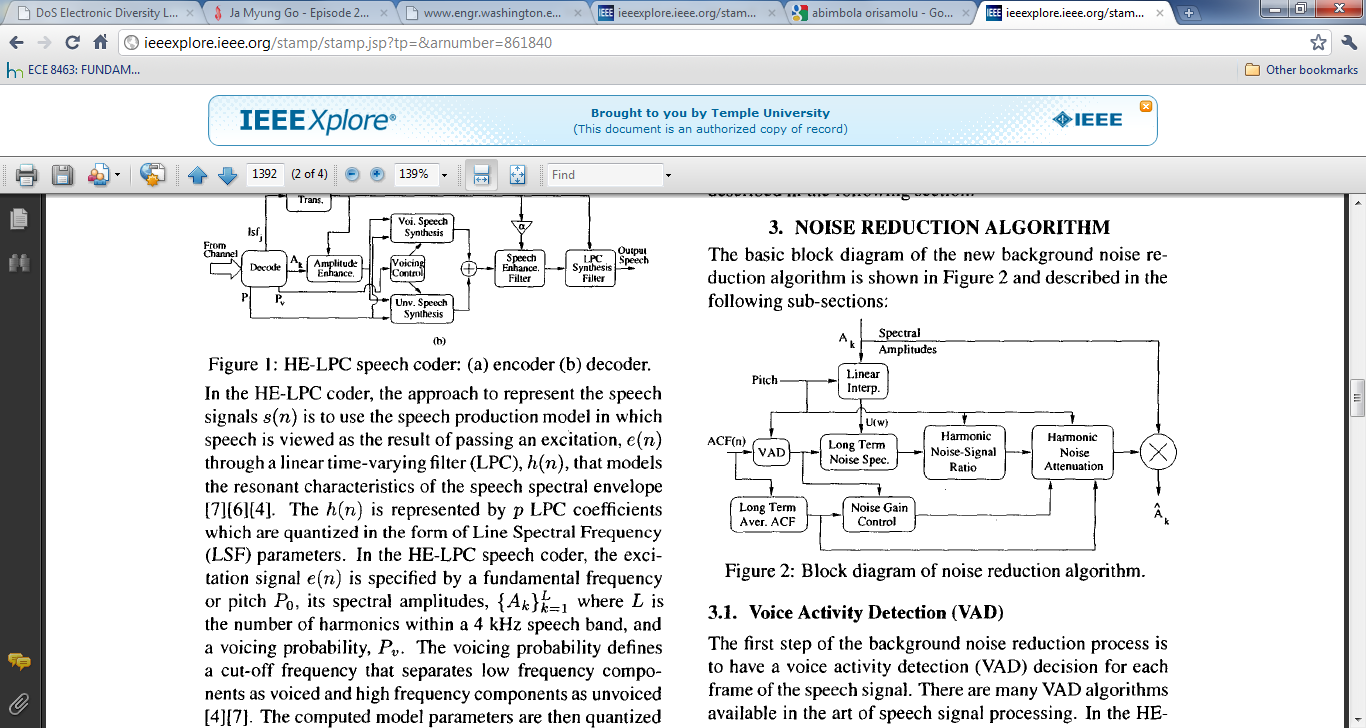


Figure 4(Block diagram of noise reduction algorithm)

1).Voice activity detection

For each frame of the speech features are calculated and used to classify each frame as speech or non-speech depending on if the feature of that speech frame exceeds a certain threshold.

2).Long term noise spectrum estimation

The noise spectrum is adjusted based on the output of the voice activity detector. It can be mathematically represented as

W ranges from o to , is the long term magnitude spectrum for the speech frame, is a constant that is typically around 0.95 and vad =0 is when the voice activity detector gives a non-speech classification. is the magnitude spectrum that is derived from the harmonic residual spectral amplitudes. The harmonic residual spectral amplitudes are computed as follows:

where is the kth harmonic spectral amplitude, is the fundamental frequency of the current speech signal, is the LPC residual spectrum, is a window function, and ,and ; are the upper and lower limits for the kth harmonic. The harmonic spectral amplitudes are then interpolated to form a fixed dimension magnitude spectrum, U(W)as :

; k where 1≤k≤L and L is the total number of harmonics within the 4kHz speech brand.

3).Noise Gain Control

The voice activity detector is used to control the noise reduction gain from frame to frame.

Where ∆ is a constant factor (typically ∆ = 0.1) and

Where min is the lowest noise attenuation factor (typically min= 0.4)

4).Harmonic Noise attenuation

The estimation Noise to Signal Ratio (ENSR) is calculated for each harmonic lobe and used to obtain the attenuation factor. The ENSR for the kth harmonic can be calculated using the equation below

) is the estimated noise spectrum for the speech frame and is the window function.

;

are the lower and upper limits of the kth harmonic and be computed using the equation below

is the fundamental frequency for the corresponding speech sequence. A noise attenuation factor will be computed using the noise gain control and harmonic noise attenuation factor for each harmonic.

If is less than 0.1,it will be set as 0.1. is a constant factor that can be set as

is the long term average background noise energy computed as:

Where is a constant factor of around 0.96 and is the average background noise energy for the current frame of speech signal being processed,. The noise attenuation factor for each harmonic will be used to scale the harmonic amplitudes.

**3). Speech enhancement using conditional estimations**

Conditional Bayesian estimation can be applied in stages to enhance a speech to deliver high quality at the receiver end of a communication device. The different stages of speech enhancement using conditional estimation are

1. Noise Reduction
2. Error Concealment
3. Bandwidth Extension

Below is a schematic showing a communication system that uses conditional estimation for enhancing speech signals

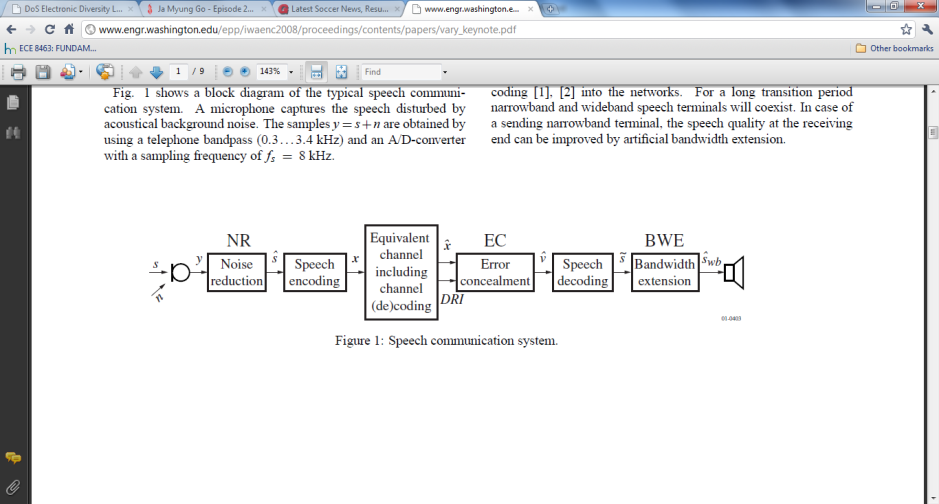


Figure 5 (Speech communication system)

A).Noise reduction stage

The objective in this stage is to produce an enhanced speech signal with reduced background noise before it is fed to the speech encoder. It required knowledge of the noisy signal a prior statistical knowledge about the speech and noise. Below is a block diagram of the implementation of the noise reduction.

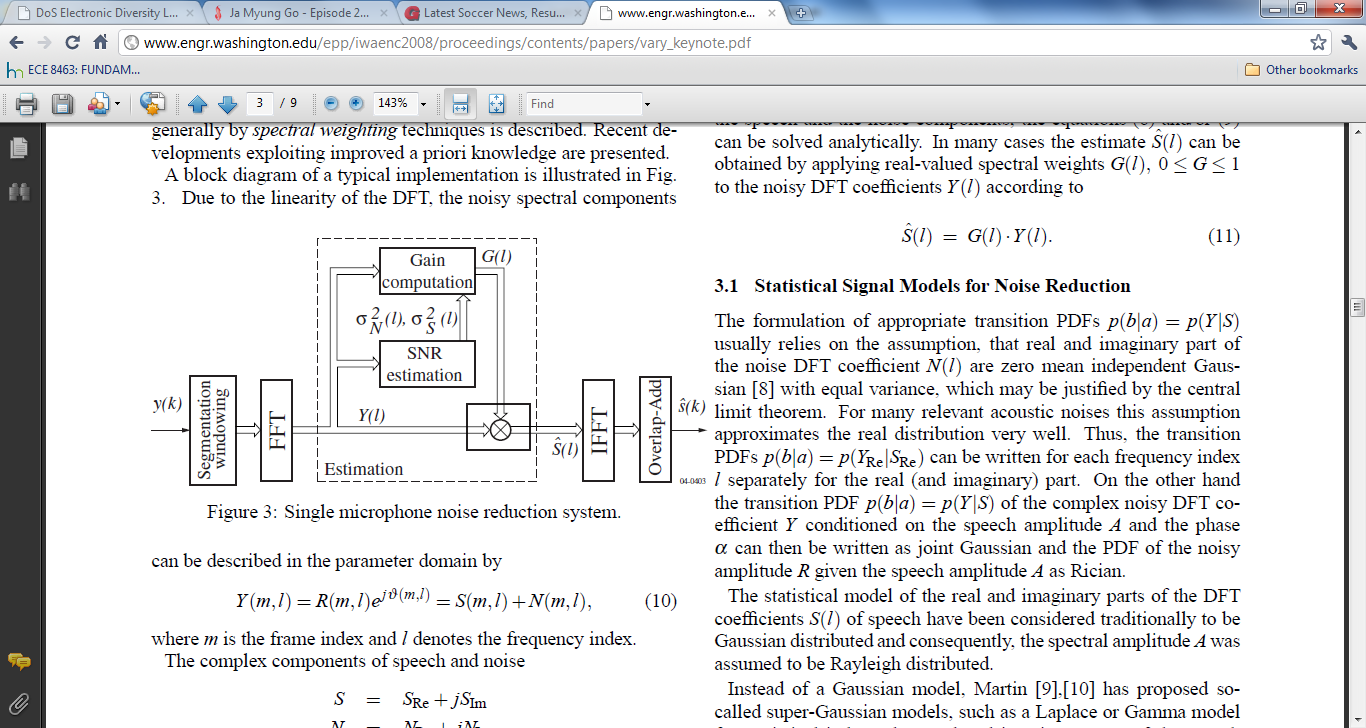


Figure 6(SIngle microphone noise reduction system)

The input speech y(k) has a segmentation window applied to it. Its DFT co-efficient is then computed. The noise added signal takes on this form in the frequency domain:

where L is the frame size. This system uses a frame size of 256 and a sampling frequency of 20kHZ. To compute the next DFT the window is shifted by Q=112 samples.

Where Y,S and N are the noise added speech, speech and

noise respectively. The speech co-efficient consists of an amplitude A and a phase α.

.

The SNR estimation block calculates a priori SNR (ζ) and a posteriori SNR ( for each DFT bin using an estimation of the noise power spectral density obtained by minimum statistics.

is the estimate of the instantaneous frequency and time dependent power spectral density of the speech.

In the estimation block in the diagram above the Noise spectral variance and the speech variance are obtained using a minimum statistic algorithm. The minimum statistic algorithm makes two assumptions. Firstly the speech and noise are usually statistically independent, zero mean and that the power of noisy speech signal frequently decays to the power level of the disturbing noise. Motivated by the central limit theorem, real and imaginary part of the noise DFT coefficients are very often modeled as zero mean independent Gaussian with equal variance. This is true for many relevant acoustic noises. The variance of the noise DFT coefficient σ2Nis assumed to split equally into real and imaginary part. The PDF of the noisy spectrum Y conditioned on the speech amplitude A and phase α can then be written as joint Gaussian

The PDF of the noisy amplitude R given the speech amplitude A is rician

denotes the modified Bessel function of zeroth order.

The statistical model of the real and imaginary parts of the DFT coefficients S(l) of speech have been considered traditionally to be Gaussian distributed and the spectral amplitude A is assumed to be Rayleigh distributed. Instead of a Gaussian model, a super-Gaussian models, such as a Laplace or Gamma model is used for statistical independent real and imaginary parts of the speech coefficients. The PDF of the Fourier component of speech can either be modeled as a Laplace density

Or as a gamma density .

The PDF of the Fourier component of the noise can be modeled as

Using the multiple distribution of the noise and speech signal, the noise DFT component can be extracted leaving the DFT co-efficient of a noise reduced signal. The inverse of the noise reduced DFT is taken and added in overlaps to obtain a noise reduced speech. Once the noise reduction stage is complete its output is passed through a speech encoder before being passed through a channel which transforms the speech on a frame by frame basis into a parameter v. The output of the channel will then undergo a second stage of enhancement called error concealment.

B). Error Concealment

The objective in this enhancement stage is to improve data quality in a noisy transmission environment by reducing the effects of residual bits error which have not been eliminated by channel decoding. The approach used here is a softbit speech decoding which can be implemented as part of a robust speech decoder. This method makes use of bit reliability information provided by the demodulator. In a mobile radio system a speech coder is applied to preserve the quality level of speech over wide ranges of channel characteristics. It is very important that coding residual bit errors are avoided to prevent degradation of speech quality. The soft bit being used can be interpreted as a joint knowledge of a hardbit and its estimated bit error probability. The softbit speech decoding process consists of four steps

* Calculation of parameter transition probabilities
* Computation of a posteriori probabilities
* Estimation of codec parameters
* Conventional speech decoding

Bit error probability estimates can be modeled in two ways

1. As a fading channel with coherent BPSK: We assume a fading channel with additive white Gaussian noise (constant power spectral density) and coherent binary phase shift keying (BPSK) demodulation. The bit error probability of the hard decided bit can be formulated as

with and is the real-valued received value.

1. Transmission scheme with channel coding: The reliability information can be provided by a soft-output Bahl channel decoder which provides a log-likelihood value of

is the input sequence of the channel decoder and the hardbit is simply

Its corresponding bit error probability is

Step 1: Parameter Transition Probabilities

Once the bit error probability is computed we can obtain the conditional bit probability for the transition of of a transmitted bit to the known received bit as

And if the equivalent channel is memoryless the parameter transition probability reads

This term includes the channel characteristics and provides the probability of a transition from any possibly transmitted bit combination, to the known received bit combination . In real-world applications the assumption of a memory-less equivalent channel can be a coarse approximation, even if an interleaving scheme is employed. However, the achievable error concealment is still very effective.

Step 2: A posteriori probabilities

For the estimation of a speech codec parameter at the receiver, a posteriori probabilities providing information about any possibly transmitted bit combination are required.

1) Approach with no A Priori Knowledge: If there is no a priori knowledge available about the regarded speech codec parameter, it has to be assumed that the quantizer output symbols are uncorrelated and equally likely. In a case like this only the channel dependent information can be exploited in terms of softbits. The required probabilities are

With the normalization constant

*2) Approach with 0th Order A Priori Knowledge (AK0):* If there is 0th order *a priori* knowledge available the Bayes rule yields *a posteriori* probabilities

with

If all quantized parameter value are equally likely then approach 1 and 2 will yield the same result.

Step 3: Codec Parameter Estimation

Once the A Posteriori term is computed. The parameter value can be estimated using either the mean square (MS) criterion or the maximum a posteriori (MAP) estimator.

1. MAP estimation: The MAP estimator follows the criterion

Regardless of the type of speech codec parameter, a MAP estimation will minimize the probability of an erroneous decoded parameter. The optimum decoded parameter in a MAP equals on of the codebook/quantization table entries.

1. MS estimation: The optimum decoded parameter in a mean square sense equals

After the codec parameter estimation the speech is decoded and passed to the final speech enhancement stage Bandwidth Extension.

C). Bandwidth Extension

The algorithm for this can be divided into two tasks:

* Extension of the spectral envelope of the speech signal
* Extension of the spectral envelope of the excitation signal.

The schematic below is a block diagram of the algorithm

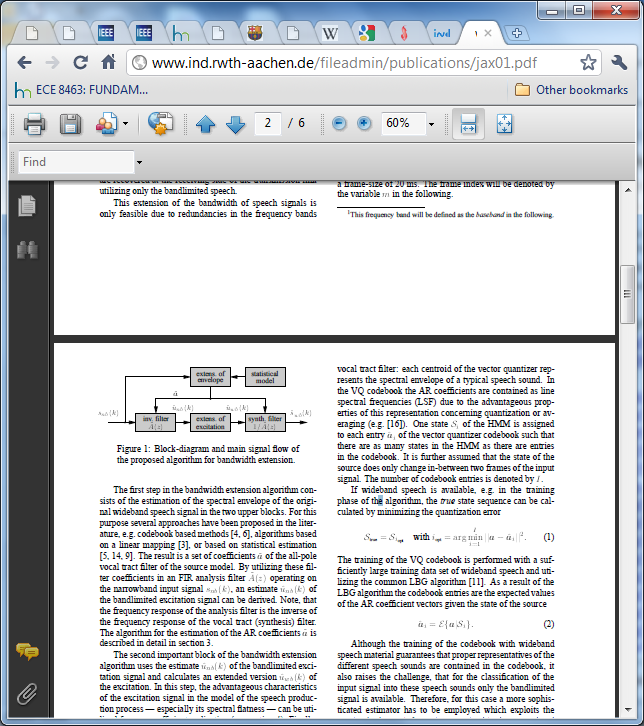


Figure 7(Block diagram for bandwith extension)

The first step in the bandwidth extension algorithm is to estimate the spectral envelope of the original wideband speech signal. The HMM Markov model is used. The states of the HMM are defined by the levels of a vector quantizer (VQ) of the co-efficients sets of the wideband auto-regressive vocal tract filter. Each centroid of the vector quantizer represents the spectral envelope of a speech sound. Each state S of the HMM is assigned to each entry of auto-regressive filter such there are as many states in the HMM as there are entries in the vector quantization codebook. The wideband speech material ensures that there is proper representations of different speech sound in the codebook. However only the band limited signal is available for classifying input signal into these speech sounds.

A more sophisticated estimator is employed to address this problem. It has the following steps

* Reducing the dimensionality of the estimation problem by extracting a limited number of features from each frame of the narrowband speech signal
* The extracted features are compared with a pre-trained statistical model of speech production.
* The current speech frame is then either classified into one of the trained speech sounds or the auto regressive (AR) co-efficient are estimated directly.

The second step of the bandwidth extension algorithm uses the spectral estimate of the band limited excitation signal and calculates an extended version of the excitation. This then creates an excitation signal with a spectral flatness that can be utilized for a very efficient realization. The estimated wideband excitation signal is then fed into and all pole synthesis filter to create the enhanced output speech signal.

**Performance Assessment of the three different methods**

The goal of speech enhancement is to produce intelligible speech of high quality by human perception. There are other ways to measure quality of speech such as signal to noise ratio(SNR) but ultimately the best test of a speech enhancer is how it‘s output is perceived by humans.

**1).For the Sequential spectrum detection based on modified DFT pair**

For this experiment two speeches from the multilingual speech Database 2002 sampled at 8kHZ and digitized by 16 bits are used. Four types of noise were used, a white noise, a pink noise, a babble noise and a factory noise. The first two noise types are stationary and the last two noises and non-stationary. The resulting noise reduced signals obtained using this technique and noisy signal were evaluated by having 11 male students listen to the signals without prior knowledge of the speech. They were given a scale from +2 to -2.

Table 1(Evaluation Standard)

|  |  |
| --- | --- |
| +2 | Noise reduced signal is better than noisy signal |
| +1 | Nose reduced signal is relatively better than noisy signal |
| 0 | They sound the same |
| -1 | Nose reduced signal is relatively worse than noisy signal |
| -2 | Nose reduced signal is worse than noisy signal |

For the pink and babble noise the evaluation scores were greater than zero. The method however failed to eliminate factory noises. The method effectively reduces the noise level depending on what kind of noise was present in the speech, however it introduces artificial sound in the noise reduced speech. This proves the technique is not robust enough to deal with different kinds of noise. Their results were compared with those obtained using a modified spectral subtraction method. According to their evaluation the modified spectral subtraction method also generated artificial sounds which were perceived as relatively husky leading to a worse rating for this method.

**2). For the sinusoidal speech coding system reduction technique**

To evaluate the effectiveness of this system, objective and subjective tests were carried out.

The objective performance of the system was determined by computing the signal to noise ratio (SNR)

r(n) s considered to be the reference signal or the best achievable version of the speech. For this experiment 8 female and 8 male sentence pairs were used to compute the signal to noise ratio. Two noise types were considered, a 15db car noise and a 25dB Babble noise. Three different SNRS were computed. The first being the SNR of the clean speech with noise reduction applied to it. The second being a speech signal with additive noise applied to it. The third being the noise added signal after the noise reduction technique has been applied to it. The results can be seen in the table below

Table 2(Signal to noise ratio output)

|  |  |  |  |
| --- | --- | --- | --- |
| Noise Type | SNR1 | SNR2 | SNR3 |
| 15 dB Car | 48.5dB | 4.2dB | 8.7dB |
| 25 dB Babble | 48.5dB | 15.5dB | 19.6dB |

As expected the noise added signal after noise reduction has an higher SNR than the noise added signal without noise reduction.

For the subjective test fourteen listeners listened to 8 sentence pairs for 4 speakers. According to the paper all listeners preferred the enhanced signal to the noise added signal. The algorithm was not used on several noise types making it difficult to determine the robustness of this technique when confronted with various noise models.

**3).For the speech enhancement using conditional estimation**

There wasn’t enough information provided to discern the level of performance of this algorithm; however statistics are becoming more and more important in speech processing. Theoretically the algorithm is very sound and should be very robust since the noise portion of the signal can be modeled as probability distribution function meaning various noise types can be included in the model.

**Conclusions**

There are several noise reduction techniques that can be used to improve noise added signal to make it acceptable by human perception. The decision on what algorithm to use depends on how robust you need it to be in the presence of various additive noises. The best measure of the performance of the noise enhancement system is human opinion. This branch of speech processing when implemented correctly has many applications. It can be used in in multimedia and wireless communications and speech recognition. Several noise types need to be modeled and used to optimize the various systems for best performance in any speech environment.

# References

Nakamashi,I. and Fujii.,K “Speech Noise Reduction Using Sequential Spectrum Detection Based on Modified DFT pair,” *2009 international symposium on Intelligent Signal Processing and Communication System*. 65-68, December 7-9 2009.

Yeldener,S, and Rieser J *A Background Noise Reduction Technique Based on Sinusoidal Speech Coding Systems*, COMSAT Laboratories, Maryland, USA, 2000.

Vary,P, “Speech Enhancement by Conditional Estimation” Institute of Communication Systems and Data Processing, Aachen, Germany.

Vary, P. and Lotter, T., “Noise Reduction by Maximum A Posteriori Spectal Amplitude Estimation with Supergaussian Speech Modeling,” Institute of Communication Systems and Data Processing, Aachen, Germany, pp. 83‑86.

Vary, P. and Lotter, T., “Noise Reduction by Joint Maximum A Posteriori Spectal Amplitude and Phase Estimation with Supergaussian Speech Modeling,” Institute of Communication Systems and Data Processing, Aachen, Germany, pp. 1457‑1460.

Jax, P and Vary,P., “Enhancement of Band-Limited Signals,”.