Audio Engineering Society

Convention Paper

Presented at the 112th Convention
2002 May 11–14 Munich, Germany

Single Channel Noise Reduction for Hands Free Operation in Automotive Environments

Stefan Schmitt, Malte Sandrock and Jochen Cronemeyer

DSPecialists GmbH, Rotherstraße 22, D-10245 Berlin, Germany, stefan.schmitt@dspe.de

ABSTRACT

The objective of this work is to establish a single channel noise reduction algorithm for speech enhancement integrated in DSP systems. The main emphasis is on spectral weighting. The chosen algorithm is based on a Minimum Mean Square Error Log Spectral Amplitude approach. One of the crucial tasks for good results, i.e. natural and intelligible speech in combination with well attenuated noise and low spectral distortion, is a balanced estimation and weighting of the noise magnitude spectrum.

INTRODUCTION

In modern hands free speech communication environments often occurs the situation that the speech signal is superposed by background noise (see Figure 1). This is particular the case if the speaker is not located as close as possible to the microphone. The speech signal intensity decreases with growing distance to the microphone. It is even possible that background noise sources are captured at a higher level than the speech signal. The noise distorts the speech and words are hardly intelligible.

In order to improve the intelligibility and reduce the listeners (FES) stress by increasing the signal to noise ratio a noise reduction or - in a wider sense – an also called speech enhancement algorithm is applied.

The objective of this work is to establish a model of a single channel speech enhancement algorithm using MATLAB as a base for a DSP software implementation. The main emphasis is on spectral weighting. The chosen algorithm with the best results is based on a *Minimum Mean-Square Error Log-Spectral Amplitude* (MMSE-LSA) approach.

NOISE REDUCTION PRINCIPLES

The requirements of a noise reduction system for speech enhancement are:

- Intelligibility and naturalness of the enhanced signal
- Improvement of signal-to-noise ratio
- Short signal delay
- Computational simplicity

The quality of the enhanced signal is a diverse issue, it may be characterised by the terms *intelligibility* and *naturalness*. There are several methods for performing noise reduction, but all can be regarded as a kind of filtering. In our application speech and noise are mixed to one signal channel. They reside in the same frequency band and may have similar correlation properties. Consequently the filtering will inevitably have an effect on both the speech and the noise. Therefore it is a very challenging task to distinguish between them. I.e. speech components can be detected as noise and thus will be suppressed as well. Especially fricatives and plusives are attenuated due to their noise-like properties. Furthermore the

residual noise characteristics should preserve the characteristics of the background noise in the recording environment. Typical single channel noise reduction algorithms add a synthetic noise, also called 'Musical Noise', which sounds artificial and has a disturbing effect on the listener.

Single channel noise reduction algorithms are based on the fact that the statistical properties of speech are only stationary over short periods of time whereas the noise often can be assumed to be stationary over much longer periods.

Another aim for the algorithm design is the limitation of the signal delay because of its annoying effect in dialog situations.

The noise reduction algorithms can be split into two groups: time domain algorithms and those utilising some kind of transform, e.g. Fourier Transform. Whereas the filter calculation for time domain solutions generally relies on the usage of correlation estimates, there is a large variety of algorithms operating in the frequency domain.

Noise reduction in frequency domain

The fundamental concept of a frequency domain solution is spectral weighting and block processing. The architecture of such a system is presented in Figure 2. It consists of three major components:

- the analysis/synthesis framework for time domain / frequency domain transformation
- the noise estimation
- the weighting function.

In a typical hands free situation (Fig. 1) the recorded time domain signal $x(k)$ is composed of the superposition of speech $s(k)$ and noise $n(k)$:

$$x(k) = s(k) + n(k) \qquad \textit{(equ. 1)}$$

The basic idea of spectral subtraction is to estimate the noise spectrum $N_{est}(n,\Omega_i)$ and to subtract it from the observed signal spectrum $X(n,\Omega_i)$:

$$Y(n,\Omega_i) = S_{est}(n,\Omega_i) = X(n,\Omega_i) - N_{est}(n,\Omega_i) \qquad \textit{(equ. 2)}$$

where n designates the current frame and Ω_i the frequency bin.

If the noise estimation equals the disturbing noise spectrum the output signal spectrum $Y(n,\Omega_i)$, also designated as speech estimation $S_{est}(n,\Omega_i)$, will be very similar to the noiseless speech spectrum $S(n,\Omega_i)$.

Because simple spectral subtraction shows limited performance in manner of speech quality, *equ. 2* is converted to

$$Y(n,\Omega_i) = X(n,\Omega_i)\cdot\left(1-\frac{N_{est}(n,\Omega_i)}{X(n,\Omega_i)}\right) \qquad \textit{(equ. 3)}$$
$$= X(n,\Omega_i)\cdot H(n,\Omega_i)$$

to enable further refinements by deriving a more sophisticated weighting function $H(n,\Omega_i)$ (see "Spectral Weighting"). Thus the enhanced output is given as the weighted input signal spectrum. The weighting function, which can be regarded as a kind of an adaptive frequency domain filter, is calculated of both estimated noise and current spectrum as depicted in Figure 2.

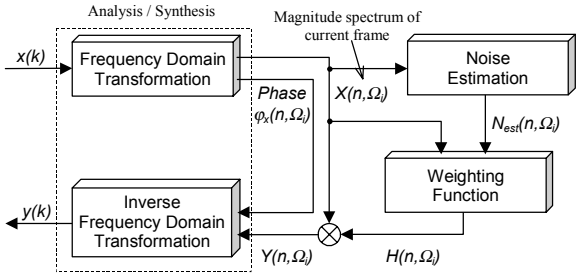

Fig. 2: Block diagram of the spectral weighting approach

Apart from rather low SNRs the magnitude of the noise will not be high enough to cause a considerable phase change of from s(k) to x(k). Since the human perception is almost not sensitive to distortions of the phase of the signal, it is sufficient to enhance the magnitude spectrum of the input signal $X(n,\Omega_i)$. Therefore the signal is transformed to the frequency domain, filtered and then transformed back to the time domain. The processing is organised block-wise, i.e. the input signal is partitioned into overlapping frames of equal size (see next chapter) and collected over time.

Estimating the noise spectrum $N_{est}(n,\Omega_i)$ is one of the major tasks of a noise cancelling system. Based on the above mentioned assumption that the noise part of the signal is stationary over longer periods of time than the speech part, an estimate of the noise is obtained by extracting slowly changing portions of the signal spectrum.

The output frame is obtained by applying the inverse frequency transformation to the weighted enhanced spectrum $Y(n,\Omega_i)$ and the noisy phase $\varphi_x(n,\Omega_i)$.

ANALYSIS/SYNTHESIS FRAMEWORK

Since in a single channel approach the estimation of the noise and the weighting function can only be derived in frequency domain, the time domain input signal has to be transformed. The transformations are performed by means of standard analysis and synthesis systems operating on a frame-by-frame basis. The frame-wise

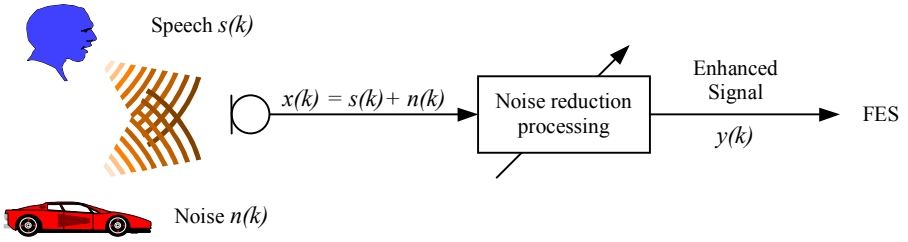

Fig. 1: Single channel noise reduction setup

processing is not only motivated by the availability of the DFT but also by the fact that speech is short time stationary. Therefore the weighting rule $H(n,\Omega_i)$ has to be adapted from one frame to another. For the analysis a data frame of L consecutive samples is taken in intervals of M samples. L equals the window size, M equals the distance of the frame borders of two neighbouring frames also called frame rate. The difference L minus M equals the

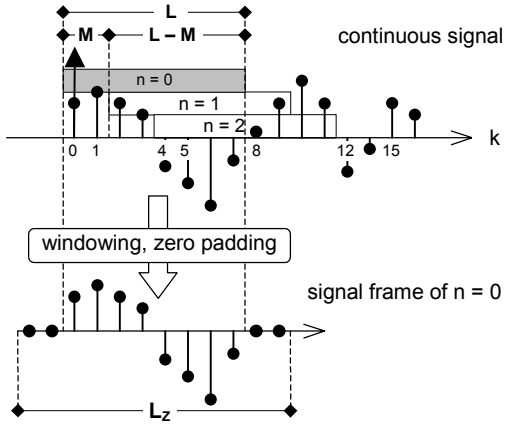

Fig. 3: Parameters of frame wise signal processing

overlap between consecutive frames. The data frame is multiplied by a hanning window and zero padded to the new frame length of L_Z, such that the samples are centred in the new frame. The initial zeros are introduced to further reduce the aliasing effect which can arise when filtering is performed in frequency domain. The analysis is finalised by the L_Z-point FFT. Consequently the magnitude spectrum of the current frame n is:

$$X(n,\Omega_i) = \left| \sum_{\nu=0}^{L_Z-1} x(n\cdot M+\nu)\cdot e^{-j\Omega_i \nu} \right|$$

$$\text{where} \quad \Omega_i = 2\pi \frac{i}{L_Z}, \quad i = 0,1,\ldots,L_Z-1 \qquad \text{(equ. 4)}$$

$$\text{and} \quad 1 \le M \le L_Z$$

The window length L should be chosen according to the speech properties: In general terms it should be of the same order as the time during which the speech can be considered as stationary. The overlap is set to 50% to 75% (of L).

The result $Y(n,\Omega_i)$ of the multiplication of the input signal spectrum with the real valued weighting function is transformed back into the time domain by an inverse FFT and the output signal $y(k)$ is synthesised by overlap and add.

NOISE ESTIMATION

For noise estimation there exist two approaches. The simplest form is the analysis during speech pauses. A typical human dialog consists of 40 % speech and 60 % speech pauses, where $x(k) = n(k)$. But this method has two obvious disadvantages: First, the changes in the noise spectrum during speech periods can not be detected, i.e. the noise has to be stationary over long time periods, and second, a voice activity detection (VAD) must be introduced to interrupt noise estimation between speech activity. One major difficulty in this case is the recognition of unvoiced phonemes.

The second possibility which is used in our approach is a minimum statistics algorithm also proposed by [7]. The algorithm is based on the observation that for each frequency band Ω_i the smallest value of $\hat{R}_{xx}(n,\Omega_i)$ (the power spectral density (PSD) estimate of a noisy speech signal) that is observed in a sufficiently large number of consecutive frames corresponds to the noise only. Consequently, by tracking these minima in a sliding window covering several frames, an estimate for the noise magnitude spectrum can be obtained. To get a reliable noise power estimation the frame size must be large enough to bridge speech activity.

At first the PSD estimates of the noisy speech signal $\hat{R}_{xx}(n,\Omega_i)$ are to be calculated. An effective way is offered by an exponential decaying window (first order recursive averaging):

$$\hat{R}_{xx}(n,\Omega_i) = \alpha_N \hat{R}_{xx}(n-1,\Omega_i) + (1-\alpha_N) X^2(n,\Omega_i) \qquad \text{(equ. 5)}$$

where α_N ($\alpha_N \in [0;1]$) is the smoothing factor. The higher this factor is, the more stable and smooth the estimate will be. On the other hand the ability to track sudden changes will decrease. When equ. 5 is used for the estimation of the noise PSD a relatively low factor α_N should be chosen (we achieved the best results with $\alpha_N = 0.85$).

The minimum noise power estimate $N_{min}^2(n,\Omega_i)$ of subband i is obtained by frame-wise comparison of the actual smoothed signal power estimate $\hat{R}_{xx}(n,\Omega_i)$ and some preceding PSD values $\hat{R}_{xx}(n-r,\Omega_i)$ with $r = 1, 2, 3, \ldots, R-1$, which are stored in a FIFO register. The introduction of that FIFO register is the special feature of the minimum statistics algorithm. The depth of the FIFO is given by R. (See Fig. 4 for the structure of the noise power estimation algorithm.)

If the actual subband power $X^2(n,\Omega_i)$ is smaller than the estimated minimum noise power $N_{min}^2(n,\Omega_i)$ the minimum noise power spectrum is updated immediately:

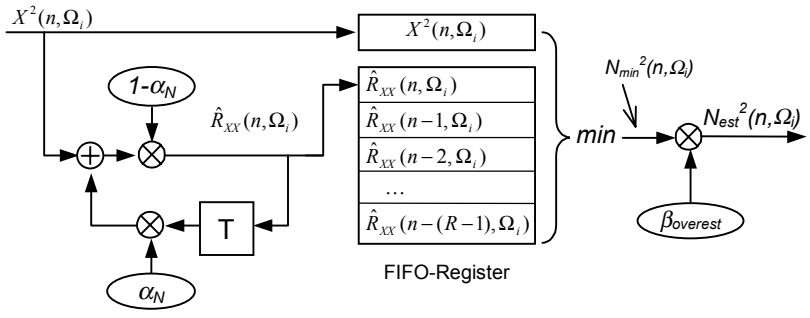

Fig. 4: Structure of the subband noise power estimation algorithm

$$N_{\min}{}^2(n,\Omega_i) = \min\{\hat{R}_{XX}{}^2(n-r,\Omega_i), X^2(n,\Omega_i)\} \quad \text{(equ. 6)}$$
$$\big| 1 \le r \le R-1$$

Thus, in case of decreasing noise power, we achieve a fast update of the minimum power estimate. In case of increasing noise power the update of noise estimates is delayed by *R* samples.

To compensate the bias of the minimum estimate the output noise power estimate $N_{est}{}^2(n,\Omega_i)$ is obtained by weighting the minimum noise power with the overestimation factor $\beta_{overest}$ with the best results for $\beta_{overest} = 1.5$:

$$N_{est}{}^2(n,\Omega_i) = \beta_{overest} \cdot N_{\min}{}^2(n,\Omega_i) \quad \text{(equ. 7)}$$

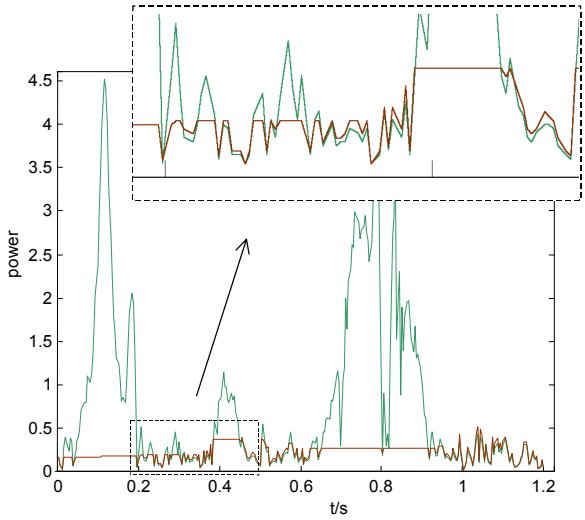

Fig. 5: Short time subband power (green/light grey) and estimated noise power (red/black) of a noisy speech signal

SPECTRAL WEIGHTING

There is a large number of algorithms for single channel noise reduction, most of them operating in the frequency domain. Among the primary ideas are *spectral subtraction* and *Wiener filtering*. A common disadvantage of these early algorithms is that their processed output signals suffer from *musical noise*. These artefacts are due to randomly distributed spectral peaks in the residual noise spectrum caused by overestimation or underestimation of the noise. This occurs if the actual noise spectrum differs considerably from the averaged magnitude spectrum. Among many proposals devoted to reduce the musical noise phenomenon the approach by Ephraim and Malah [4] and [5] can be considered as state of the art weighting rule, which reduces the musical noise drastically. Hence these algorithms have found wide spread use in many noise reduction applications.

Ephraim-Malah Weighting

Ephraim and Malah derived a *Minimum Mean-Square Error (MMSE) Short-Time Spectral Amplitude Estimator*. The main assumption is that speech and noise spectral components can be modelled as statistically independent gaussian random variables.

There are many further developments of the original idea. Here we focus on the *MMSE* rule by [4] and the derivative *Minimum Mean-Square Error Log-Spectral Amplitude (MMSE-LSA) Estimator* from [5] (further examined and described in [2] and [3]).

The gain function is determined by two values: An *a priori* signal to noise ratio

$$SNR_{prio}(n,\Omega_i) = \frac{R_{ss}(n,\Omega_i)}{R_{nn}(n,\Omega_i)} \quad \text{(equ. 8)}$$

which describes the unknown input signal SNR and an *a posteriori* SNR

$$SNR_{post}(n,\Omega_i) = \frac{|Y(n,\Omega_i)|^2}{R_{nn}(n,\Omega_i)} \quad \text{(equ. 9)}$$

which can be interpreted as the instantaneous SNR. These SNRs are to be estimated. The MMSE-LSA weighting rule is defined for discrete frequencies Ω_i and for a frame *n*. It minimises the mean squared error of the logarithmic spectra of the original undisturbed speech signal and the processed output signal

$$\varepsilon_{MMSE} = \mathrm{E}\left\{\left(\lg|S(n,\Omega_i)| - \lg|Y(n,\Omega_i)|\right)^2\right\} \quad \text{(equ. 10)}$$

with respect to the assumed distribution of the spectral magnitudes. The solution is given by

$$H_{mmse\,lsa}(n,\Omega_i) = \frac{SNR_{prio}(n,\Omega_i)}{SNR_{prio}(n,\Omega_i)+1} \cdot \exp\left(\tfrac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right) \quad \text{(equ. 11)}$$

$$\textit{with} \quad \nu = \frac{SNR_{prio}(n,\Omega_i)}{SNR_{prio}(n,\Omega_i)+1} \cdot SNR_{post}(n,\Omega_i) \quad \text{(equ. 12)}$$

An estimate for the *a posteriori* SNR can be obtained easily, because in equ. 9 the numerator equals the subtraction of the estimated noise spectrum from the input signal spectrum $X(n, \Omega_i)$. So it can be calculated as

$$SNR_{post}(n,\Omega_i) = \frac{|X(n,\Omega_i)|^2}{|N_{est}(n,\Omega_i)|^2} - 1 \quad \text{(equ. 13)}$$

Following the "Decision-Directed" estimation approach in [4] the *a priori* SNR can be derived as a weighted sum of the *a posteriori* SNR and the SNR computed with the speech power estimation of the previous output frame (the equations can be found in [2] and [3] too.):

$$SNR_{prio}(n,\Omega_i) = (1-\alpha_H)\cdot Q[SNR_{psost}(n,\Omega_i)] \quad \ldots$$
$$+\alpha_H \cdot \frac{|H(n-1,\Omega_i)X(n-1,\Omega_i)|^2}{|N_{est}(n,\Omega_i)|^2} \quad \text{(equ. 14)}$$

whereas *Q[x]* is given by *Q[x] = x if ≥ 0 and Q[x] = 0 if x < 0* to ensure, that this estimate is always greater or equal to zero. Since the weighting factor α_H should be selected close to one, the second term dominates in the equation for SNR_{prio} (equ. 14). The *a priori* SNR represents a smoothed version of the instantaneous SNR estimate. The higher α_H is, the more stable SNR_{prio} is.

In equ. 11 we note that the attenuation $H(n,\Omega_i)$ is mainly a function of the *a priori* SNR, whereas the *a posteriori* SNR acts as a correction term whose influence is limited to the case where SNR_{prio} is low. So the *a priori* SNR is the dominant parameter in the weighting function. Little attenuation $H(n,\Omega_i)$ is obtained, if SNR_{prio} is high, and high attenuation is obtained if SNR_{prio} is low. In the second case the attenuation $H(n,\Omega_i)$ increases for higher SNR_{post} values.

The estimation rule of the *a priori* SNR and the interaction with the *a posteriori* SNR are the key factors for reducing the musical tone artefacts by the MMSE-LSA weighting rule (e.g. see [2]). Musical tones appear when local bursts arise in the instantaneous noise spectrum which are larger than the average noise level. However applying equ. 13 such a peak will have an immediate effect on SNR_{post}, which will increase, whereas the effect on SNR_{prio} will be very weak because of the recursive smoothing. Thus the attenuation $H(n,\Omega_l)$ will increase so that the noise peak will be suppressed.

Concerning the application to speech signals the mechanism described above will cause suppression of the initial phase of important plosive parts of the speech as well. So a trade-off between speech distortion and musical noise level must be archived when choosing the factor α_H.

CONCLUSION

In the simulation the described noise suppression system shows good results concerning the naturalness of the speech at a sufficiently high amount of noise reduction. This is valid as long as all spectral components of the noise signal are below the related components of the speech signal. The musical noise artefacts are considerably reduced in comparison to a simulation model with a simple spectral subtraction algorithm. However, the musical noise is still audible and therefore future work will be done to further eliminate this phenomenon.

Our implementation operates at a sampling frequency of 8 kHz, The signal delay of 32 ms is caused mainly by the FFT transformation. The requirement of processing power is about 8 MIPS. We achieve a maximum noise reduction of 20 dB. The Figures 6a and 6b show the performance of the developed system. The upper figure is the spectrogram of a human speech recorded in a driving car. The second picture shows the result after the enhancement.

Sampling frequency	Delay	Max. reduction	Complexity
8000 Hz	32 ms	20 dB	~ 8 MIPS

REFERENCES

[1] M. Amiri, "Single Channel Speech Enhancement for Handsfree Operation in Automotive Environements", Diploma Thesis, DSPecialists/TU-Berlin, June 2001

[2] O. Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", IEEE Transactions on Speech and Audio Processing, April 1994.

[3] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in subbands", Proc. 4th European Conf. on Speech Communication and Technology, Vol. 2, Sept. 1995, Madrid

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. on Acoustics, Speech, and Signal processing", vol. ASSP-32(6), pp. 1109-1121, December 1984

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. on ASSP-33(2), pp. 443-445, April 1985

[6] S. Gustafsson, "Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction", Dissertation, ABDN Band 11, Wissenschaftsverlag Mainz, 1999

[7] R. Martin, "Spectral Subtraction Based on Minimum Statistics", EUSIPCO-1994, Edinburgh, Sept. 1994

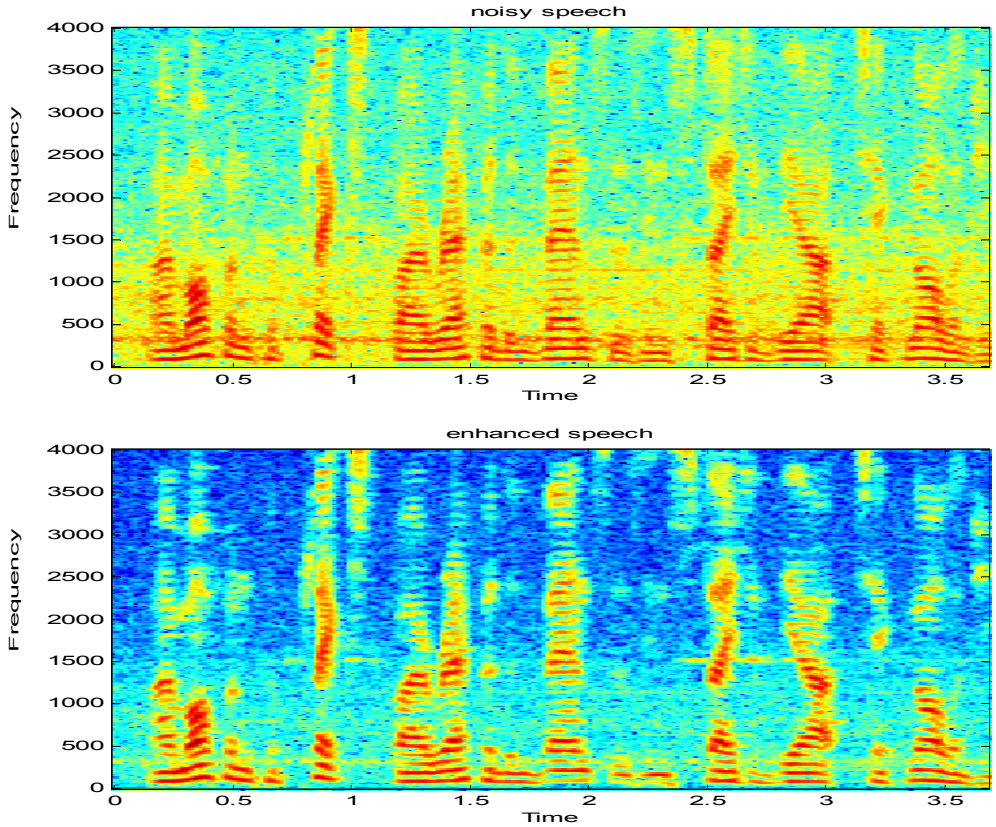

Fig. 6: Spectrogram of a human speech recorded in a driving car. Fig. 6b shows the result after the enhancement