International Scholarly Research Network
ISRN Signal Processing
Volume 2012, Article ID 306305, 9 pages
doi:10.5402/2012/306305

Research Article

Direct Recovery of Clean Speech Using a Hybrid Noise Suppression Algorithm for Robust Speech Recognition System

Peng Dai,[1] Ing Yann Soon,[1] and Rui Tao[2]

[1] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*
[2] *School of Electronic and Information Engineering, Beihang University, China*

Correspondence should be addressed to Peng Dai, daip0001@e.ntu.edu.sg

Received 9 November 2012; Accepted 28 November 2012

Academic Editors: L. Fan and A. M. Peinado

A new log-power domain feature enhancement algorithm named NLPS is developed. It consists of two parts, direct solution of nonlinear system model and log-power subtraction. In contrast to other methods, the proposed algorithm does not need prior speech/noise statistical model. Instead, it works by direct solution of the nonlinear function derived from the speech recognition system. Separate steps are utilized to refine the accuracy of estimated cepstrum by log-power subtraction, which is the second part of the proposed algorithm. The proposed algorithm manages to solve the speech probability distribution function (PDF) discontinuity problem caused by traditional spectral subtraction series algorithms. The effectiveness of the proposed filter is extensively compared using the standard database, AURORA2. The results show that significant improvement can be achieved by incorporating the proposed algorithm. The proposed algorithm reaches a recognition rate of over 86% for noisy speech (average from SNR 0 dB to 20 dB), which means a 48% error reduction over the baseline Mel-frequency Cepstral Coefficient (MFCC) system.

1. Introduction

The main objective of speech recognition is to get a higher recognition rate. However, lots of factors tend to degrade the performance of automatic speech recognition (ASR) system, such as environmental noise, channel distortion, and speaker variability [1, 2]. Generally, automatic speech recognition system consists of two parts, feature extraction and pattern matching. Therefore, methods which aim to improve the performance of ASR system can be mainly divided into two categories, the "model" approach and the "feature" approach. The "model" approach mainly focuses on improving the speech recognizer, where the speech features are classified into different patterns developed from the statistical properties of speech. As for "feature" approach, emphasis is put on improving the robustness of speech features. The method proposed by this paper belongs to this category.

Noise reduction or clean speech estimation is a straight forward "feature" approach to improve the performance of ASR systems. There are different ways to get the estimation. minimum mean square Error is one of the most important ones. Ephraim derived the short-time spectral amplitude (STSA) estimator using minimum mean square error (MMSE) in 1984 [3], which has become a standard approach for clean speech estimation in speech processing. The advantage of MMSE estimator is very obvious. It is mathematically optimized, which theoretically can get a good estimation of the clean speech. Besides, there is solid derivation making it easier to analyze. Originally, the MMSE-based algorithms were intended to be used for speech enhancement.

For speech recognition, several MMSE-based algorithms have been developed. Yu et al. in 2008 developed the MMSE estimator in the log-power domain [4]. The cepstral domain estimator appears also in 2008 [5]. Besides, different distortion models are developed for improving speech recognition system [5, 6]. Recently, some more complicated MMSE-based algorithms which require the so called stereo data input are proposed [7]. Admittedly, MMSE works well for speech enhancement and speech recognition. The main idea

of MMSE is to estimate the clean speech from the noisy speech. The success of MMSE in previous implementation reveals that it is one of the means to improve the performance of an automatic speech recognition system (ASR). However, it is not necessarily the only one. Mathematically, the recovery of clean speech from noisy corpus is a problem of solving a nonlinear function. The above mentioned MMSE approach can be treated as a kind of statistical approach to solve the function. However, there are other ways for nonlinear function root finding. In this paper, the iterative root finding approach is incorporated to recover the clean speech.

Unlike many other algorithms, the proposed algorithm does not need stereo data input, which makes it more robust to different conditions. It is because stereo data is impossible to get in practical situations. The novelty of this paper lies in that the two parts of MFCCs (c1∼c12, log-power) are processed separately. Direct solution of nonlinear system function is much easier than the statistical approach. Besides, compared with earlier MMSE-based algorithms, the proposed method does not need any additional training. The AURORA2 database is used for verification tests. It is a widely used, standard English database, which contains isolated digits as well as digit serials. Comparison is made against ordinary MFCCs, MMSE-STSA [3], Spectral Subtraction (SS) [8], Cepstral Mean and Variance Normalization (CMVN), the ETSI standard advanced front-end feature extraction algorithm (AFE) [9], and Mean Variance Normalization and ARMA filtering (MVA) [10]. Experimental results show the excellent performance of the proposed method.

The rest of the paper is organized as follows. In Section 2, the system model, nonlinear function, is presented. Section 3 discusses the details of the proposed algorithm, including detailed iteration steps of root finding algorithm, prior estimates for clean speech and noise, and the novel log-power subtraction method. The experimental speech databases, ASR systems and the additional comparison methods are described in Section 4. Conclusions are summarized in Section 5.

2. System Model

Following similar derivation procedure from [11], the clean speech waveform is denoted as x_t where t is the time index. It is assumed that x_t is corrupted by the independent additive noise waveform n_t and becomes the noisy speech waveform y_t as shown in

$$y_t = x_t + n_t. \tag{1}$$

The speech signal is cut into frames and transformed into frequency domain using DFT. Then (1) becomes

$$Y_{f,t} = X_{f,t} + N_{f,t}. \tag{2}$$

By assuming the additivity on the powers of the components in the frequency domain [12], the power spectrum of the noisy speech is given by

$$\left|Y_{f,t}\right|^2 = \left|X_{f,t}\right|^2 + \left|N_{f,t}\right|^2. \tag{3}$$

After applying Mel-filterbanks to the power spectra,

$$\sum_f W_f^l \left|Y_{f,t}\right|^2 = \sum_f W_f^l \left|X_{f,t}\right|^2 + \sum_f W_f^l \left|N_{f,t}\right|^2, \tag{4}$$

where W_f^l stands for the transfer function for the lth filter.

Define the log channel energy vectors as:

$$\mathbf{Y} = \begin{bmatrix} \log \sum_f W_f^1 \left|Y_{f,t}\right|^2 \\ \log \sum_f W_f^2 \left|Y_{f,t}\right|^2 \\ \vdots \\ \log \sum_f W_f^L \left|Y_{f,t}\right|^2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \log \sum_f W_f^1 \left|X_{f,t}\right|^2 \\ \log \sum_f W_f^2 \left|X_{f,t}\right|^2 \\ \vdots \\ \log \sum_f W_f^L \left|X_{f,t}\right|^2 \end{bmatrix},$$
$$\mathbf{N} = \begin{bmatrix} \log \sum_f W_f^1 \left|N_{f,t}\right|^2 \\ \log \sum_f W_f^2 \left|N_{f,t}\right|^2 \\ \vdots \\ \log \sum_f W_f^L \left|N_{f,t}\right|^2 \end{bmatrix}, \tag{5}$$

where $\log(\cdot)$ denotes the natural logarithm

Then (4) becomes

$$e^{\mathbf{Y}} = e^{\mathbf{X}} + e^{\mathbf{N}}. \tag{6}$$

Then changing (6) to the log-power domain,

$$\mathbf{Y} = \log\left(e^{\mathbf{X}} + e^{\mathbf{N}}\right), \tag{7}$$

then

$$\mathbf{Y} = \mathbf{X} + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}}\right), \tag{8}$$

where $\mathbf{1}$ stands for a vector with all elements equal to one.

Then, the MFCCs can be calculated by

$$\mathbf{C} = \mathrm{DCT}(\mathbf{Y}). \tag{9}$$

3. Algorithm Description

3.1. Iterative Root-Finding

3.1.1. Theoretical Analysis. In fact, with no additional constraints, and if $|Y_{f,t}|^2$ and $|N_{f,t}|^2$ are already known, according to (3) the clean speech can be estimated simply by

$$\left|\hat{X}_{f,t}\right|^2 = \left|Y_{f,t}\right|^2 - \left|N_{f,t}\right|^2, \tag{10}$$

where $|\hat{X}_{f,t}|^2$ is the clean speech estimate.

Equation (10) is just the basic idea of Spectral Subtraction (SS) [8]. However, (10) only exists when $|Y_{f,t}|^2 > |N_{f,t}|^2$. If $|Y_{f,t}|^2 \le |N_{f,t}|^2$, then $|Y_{f,t}|^2 - |N_{f,t}|^2 \le 0$. The

clean speech estimate becomes zero or negative, which is obviously wrong.

A traditional way to solve the above mentioned problem is to implement a threshold to guarantee the clean speech estimate to be positive:

$$\left|\hat{X}_{f,t}\right|^2 = \max\left(\left|Y_{f,t}\right|^2 - \left|N_{f,t}\right|^2, \varepsilon\right), \tag{11}$$

where the parameter ε is a small positive constant value.

Equation (11) is a very common way to implement Spectral Subtraction (SS) in speech recognition which will be denoted as SS in later discussion. Admittedly, (11) manages to increase SNR, which in return is a straight forward way to improve the performance of speech recognition systems.

However, there is a very serious problem caused by SS. Because of the threshold, ε, a certain portion of the recovered speech is forced to be corrected to ε. Figure 1 gives an example of the effect of SS on speech spectrogram. The blue area in Figure 1(b) is all equal to ε.

After processed by SS or other similar methods, the probability distribution of speech is greatly changed. For example, in Figure 2, it can be easily found out that the probability of speech power equal ε is greatly increased, which makes the pdf of the processed speech discontinuous.

Most state of the art ASR systems incorporate statistical methods to perform pattern recognition. HMM is one of the most popular ones. These statistical methods are all developed based on certain statistical model of the speech. In other words, a probability distribution is always assumed as the basis of recognizer derivation. SS like algorithms greatly changes the pdf of speech, which in return causes the performance of ASR systems to drop.

The proposed algorithm intends to achieve the clean speech in an iterative manner, which means the clean speech estimation $|\hat{X}_{f,t}|^2$ slowly converges to a better guess. There would not be a mass force assignment of the negative elements to a certain value. Thus the discontinuity problem is avoided.

3.1.2. Iterative Solution. The novelty of implementing iterative root finding algorithm is that unlike the Spectral Subtraction like approaches it manages to overcome the awkward $|Y_{f,t}|^2 \le |N_{f,t}|^2$ problem without causing discontinuity in the speech PDF. The statistical approach handles this by applying a series of mathematical operations which are not sensitive to the above mentioned problem. In power domain, the final expression is

$$\left|\hat{X}_{f,t}\right|^2 = G \times \left|Y_{f,t}\right|^2 = \left|Y_{f,t}\right|^2 - (1-G) \times \left|Y_{f,t}\right|^2 \tag{12}$$

which fundamentally avoids the possibility of $|Y_{f,t}|^2 \le |N_{f,t}|^2$. It is because the equivalent noise estimate is $(1-G)\times |Y_{f,t}|^2$, which is generated from only the current frame.

As described before, the iterative root finding algorithm can also handle $|Y_{f,t}|^2 \le |N_{f,t}|^2$ very well. Equation (8) can be reshaped to

$$\mathbf{X} + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}}\right) - \mathbf{Y} = 0, \tag{13}$$

where $\mathbf{Y}$ is the noisy speech vector, $\mathbf{X}$ is the parameter that needed to be recovered. If the noise vector $\mathbf{N}$ can be reasonably estimated, (13) becomes a nonlinear function about $\mathbf{X}$, which can be solved by iterative root finding algorithms.

Denoting

$$f(\mathbf{X}) = \mathbf{X} + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}}\right) - \mathbf{Y}, \tag{14}$$

then

$$f'(\mathbf{X}) = \frac{d[f(\mathbf{X})]}{d\mathbf{X}} = \mathbf{1} - \frac{e^{\mathbf{N}-\mathbf{X}}}{1 + e^{\mathbf{N}-\mathbf{X}}}. \tag{15}$$

According to Newton's method, given a function $f(\mathbf{X})$, its derivative $f'(\mathbf{X})$ and a first guess $\hat{\mathbf{X}}_0$, the solution to the function can be reached by

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \frac{f(\mathbf{X}_i)}{f'(\mathbf{X}_i)}, \tag{16}$$

where i is the iteration index.

For the iterative step

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \frac{\mathbf{X}_i + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}_i}\right) - \mathbf{Y}}{\mathbf{1} - \left(e^{\mathbf{N}-\mathbf{X}_i}/\left(1 + e^{\mathbf{N}-\mathbf{X}_i}\right)\right)} \tag{17}$$

it has to be noted that

$$\lim_{(\mathbf{N}-\mathbf{X}_i)\to+\infty} \mathbf{1} - \frac{e^{\mathbf{N}-\mathbf{X}_i}}{\mathbf{1} + e^{\mathbf{N}-\mathbf{X}_i}} = 0. \tag{18}$$

Therefore, a threshold β is adopted to guarantee the denominator to be non-zero. Then (17) is modified to

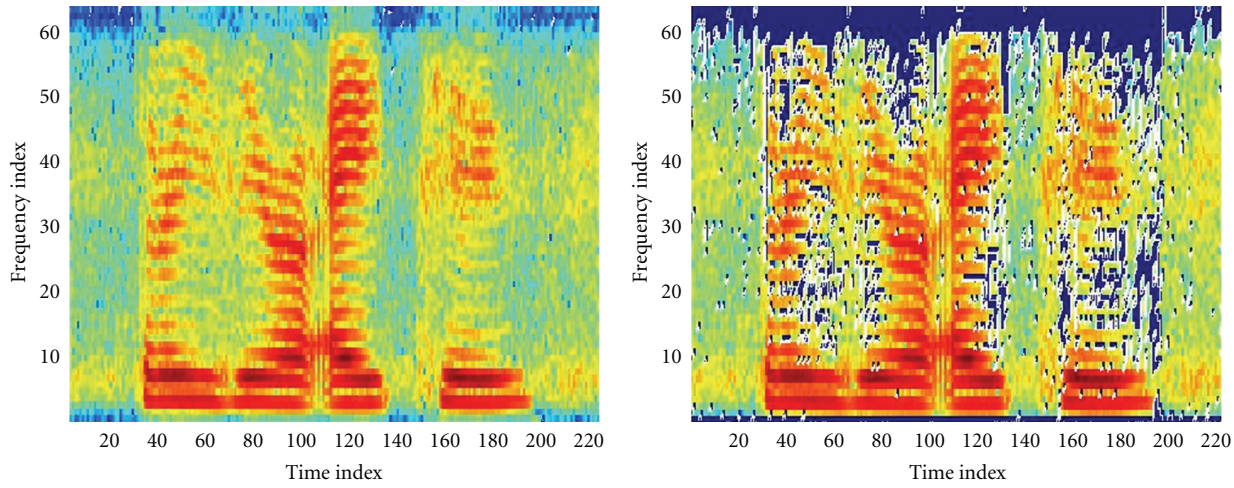
$$\mathbf{X}_{i+1} = \mathbf{X}_i - \frac{\mathbf{X}_i + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}_i}\right) - \mathbf{Y}}{\max\left(\mathbf{1} - \left(e^{\mathbf{N}-\mathbf{X}_i}/\left(1 + e^{\mathbf{N}-\mathbf{X}_i}\right)\right), \beta\right)}. \tag{19}$$

With a successful guess of the initial step, clean speech vector $\hat{\mathbf{X}}_0$ and noise vector $\hat{\mathbf{N}}$, the clean speech estimate $\hat{\mathbf{X}}$ can be satisfactorily approximated. Equation (19) can work very well even if $|Y_{f,t}|^2 \le |N_{f,t}|^2$. About the discontinuity problem, at extreme conditions where the threshold β works, the iteration becomes

$$\begin{aligned}\mathbf{X}_{i+1} &= \mathbf{X}_i - \frac{\mathbf{X}_i + \log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}_i}\right) - \mathbf{Y}}{\beta} \\ &= (1-\beta)\mathbf{X}_i - \frac{1}{\beta}\log\left(\mathbf{1} + e^{\mathbf{N}-\mathbf{X}_i}\right) + \frac{1}{\beta}\mathbf{Y}.\end{aligned} \tag{20}$$

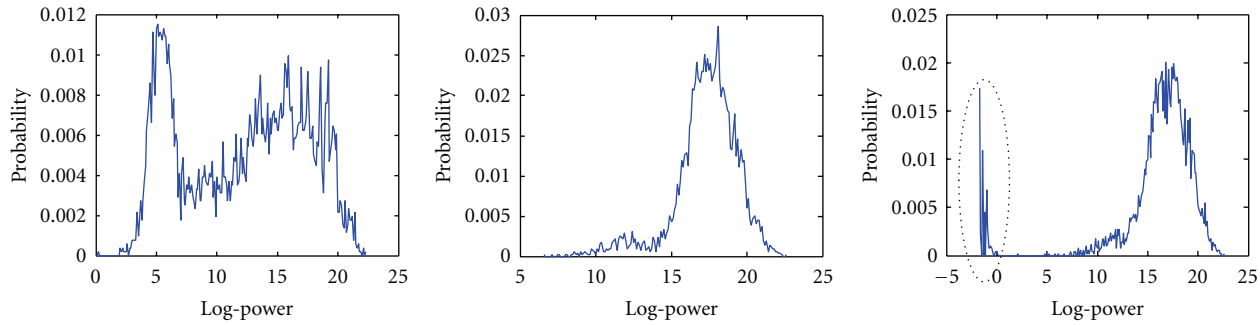
It can be easily seen that (20) would not cause mass assignment of the same value, which means the discontinuity problem will not appear.

3.2. Prior Estimates. In statistics, a minimum mean square error (MMSE) estimator is the approach which minimizes the mean square error (MSE), which is widely used in lots of areas in signal processing. In 1984, Ephraim and Malah derived the short time spectral amplitude (STSA) estimator using MMSE [3]. After that, MMSE has become a standard approach for enhancing the quality of speech. Therefore, it



(a) Original (b) Processed by SS

Figure 1: Spectrogram of digital string "3Z82".



(a) Clean (b) Noisy Speech (c) Processed by SS

Figure 2: Speech PDF of Mel channel log-power for digital string "3Z82".

is chosen to generate the prior estimate of the clean speech. The following equation shows the standard cost function for MMSE approach [3]:

$$\hat{X} = \arg\min_{\hat{X}} E\left[\left(X - \hat{X}\right)^2\right]. \tag{21}$$

By following MMSE-STSA [3] the clean speech estimate can be reached by

$$\begin{aligned}\hat{X}_{f,t} &= \Gamma(1.5)\frac{\sqrt{\nu_{f,t}}}{\gamma_{f,t}} M\left(-0.5; 1; -\nu_{f,t}\right) Y_{f,t} \\ &= \Gamma(1.5)\frac{\sqrt{\nu_{f,t}}}{\gamma_{f,t}} \exp\left(-\frac{\nu_{f,t}}{2}\right) \\ &\quad \times \left[\left(1+\nu_{f,t}\right) I_0\left(\frac{\nu_{f,t}}{2}\right) + \nu_{f,t} I_1\left(\frac{\nu_{f,t}}{2}\right)\right] Y_{f,t},\end{aligned} \tag{22}$$

where $\Gamma(\cdot)$ denotes the gamma function; $M(a;c;x)$ is the confluent hypergeometric function; $I_0(\cdot)$ and $I_1(\cdot)$ denote the zero and first order modified Bessel function; $\xi_{f,t}$ and $\gamma_{f,t}$ are the a priori and a posteriori signal-to-noise ratios (SNR), respectively:

$$\nu_{f,t} = \frac{\xi_{f,t}}{1+\xi_{f,t}} \gamma_{f,t}. \tag{23}$$

Then the clean amplitude estimate is transferred to log-power domain:

$$\begin{aligned}\hat{X}^l &= \log \sum_f W_f^l \left|\hat{X}_{f,t}\right|^2 \\ &= \log \sum_f W_f^l \left| \Gamma(1.5)\frac{\sqrt{\nu_{f,t}}}{\gamma_{f,t}} \exp\left(-\frac{\nu_{f,t}}{2}\right) \right. \\ &\quad \left. \times \left[\left(1+\nu_{f,t}\right) I_0\left(\frac{\nu_{f,t}}{2}\right) + \nu_{f,t} I_1\left(\frac{\nu_{f,t}}{2}\right)\right] Y_{f,t} \right|^2.\end{aligned} \tag{24}$$

Equation (24) will serve as the initial guess for the iterative approach.

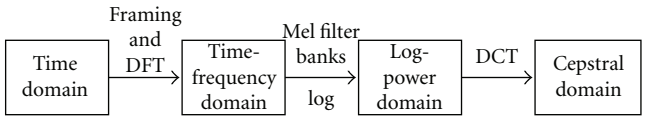


Figure 3: Different domains in MFCC scheme.

3.3. Log-Power Subtraction (LPS)

3.3.1. Algorithm Description. As is shown in Figure 3, there are mainly four different domains in the MFCC scheme. The proposed algorithm works in the log-power domain.

The clean speech estimate generated by the proposed algorithm in (19) is actually

$$\mathbf{X} = \begin{bmatrix} \log \sum_f W_f^1 \left| X_{f,t} \right|^2 \\ \log \sum_f W_f^2 \left| X_{f,t} \right|^2 \\ \vdots \\ \log \sum_f W_f^L \left| X_{f,t} \right|^2 \end{bmatrix}. \tag{25}$$

It is the clean speech log-power vector, in the log-power domain as described in Figure 3.

The MFCC static parameters can be divided into two parts, c1∼c12 and c0/log-energy. Strictly speaking, the proposed algorithm mainly focuses on c1∼c12. For log-energy, traditionally it should be calculated by

$$P_{\log} = \log\left(\sum_f \left| X_{f,t} \right|^2 \right). \tag{26}$$

The clean speech power estimate, $|X_{f,t}|^2$, cannot be perfectly recovered from (25) because of the Mel-filterbanks. Additional distortion will be introduced to the feature vectors. For c0, although it seems to work smoothly, the recognition results are just about "average". Therefore, a separate noise removing scheme is developed. At the iterative root finding part, an estimate for the noise is reached. The frame clean speech power can be estimated by

$$\hat{P}_c = \sum_f \left(\left| Y_{f,t} \right|^2 - \hat{N}_{f,t} \right). \tag{27}$$

Then the log-energy can be calculated by

$$P_{\log} = \log\left(\hat{P}_c\right) = \log\left[\sum_f \left(\left| Y_{f,t} \right|^2 - \hat{N}_{f,t} \right) \right]. \tag{28}$$

However, problem arises when $\hat{P}_c \leq 0$. Therefore, a weighting parameter is incorporated to reduce the chances of imaginary parts appearing. Then (28) becomes

$$\hat{P}_c = \sum_f \left(\left| Y_{f,t} \right|^2 - \alpha \hat{N}_{f,t} \right). \tag{29}$$

Furthermore, another parameter ε_0 is set the guarantee the log-energy not to be infinity. Therefore, the log-power part becomes

$$P_{\log} = \log\left(\hat{P}_c\right) = \log\left\{ \max\left[\sum_f \left(\left| Y_{f,t} \right|^2 - \alpha \hat{N}_{f,t} \right), \varepsilon_0 \right] \right\}. \tag{30}$$

3.3.2. Theoretical Analysis. The basic idea of log-power subtraction is similar to the Spectral Subtraction (SS) algorithm developed by Boll in 1979 [8]. Figure 4 shows the diagram of Spectral Subtraction algorithm.

Boll defined the SS in the magnitude domain in [8]. When adopted in speech recognition, SS is normally implemented in the power spectral domain.

Most of noise estimation algorithms are developed based on statistical models of clean and noisy speech, which makes the estimation at a specific point, $|N_{f,t}|^2$, more like an expectation or average of noise based on previous frames. Therefore, when used in spectral subtraction, lots of elements will become negative, especially in the non-speech period, which will lead to the problem described in Section 3. However, for the proposed LPS approach the effect of the above mentioned problem is to a certain extent avoided. It is because traditionally the log-power is calculated by (26). It is based on the sum of all the speech power elements in one frame. Mathematically, from (29), the following equation can be derived:

$$\begin{aligned} \hat{P}_c &= \sum_f \left(\left| Y_{f,t} \right|^2 - \alpha \hat{N}_{f,t} \right) = \sum_f \left| Y_{f,t} \right|^2 - \alpha \sum_f \hat{N}_{f,t} \\ &= F \times \left(\frac{1}{F} \sum_f \left| Y_{f,t} \right|^2 - \frac{1}{F} \alpha \sum_f \hat{N}_{f,t} \right), \end{aligned} \tag{31}$$

where F is the total number of frequency bins.

Equation (31) shows that LPS is equivalent to performing spectral subtraction after averaging all the elements in the current frame. Due to the averaging process, the whole spectral subtraction scheme become more robust since both speech and noise estimate are kind of expectations of the actual signal.

3.4. Implementation Details.

The proposed algorithm consists of two parts, iterative solution of the nonlinear function and log-power subtraction.

Figure 5 shows the block diagram of the proposed algorithm. The detailed parameter settings are $\alpha = 0.9$, $\alpha_{f,t} = 0.4$, $\beta = 0.8$, $\varepsilon_0 = 10^{-10}$, and the iteration is performed only once. Minimum Statistics (MS) is used for noise estimation [13].

4. Algorithm Evaluation

4.1. Experiment Setup

4.1.1. AURORA2 Database. The AURORA2 database is adopted to evaluate the performance of the proposed

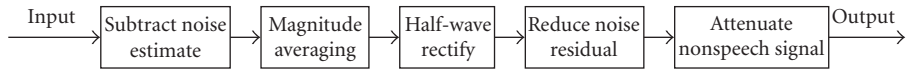


Figure 4: Diagram of spectral subtraction.

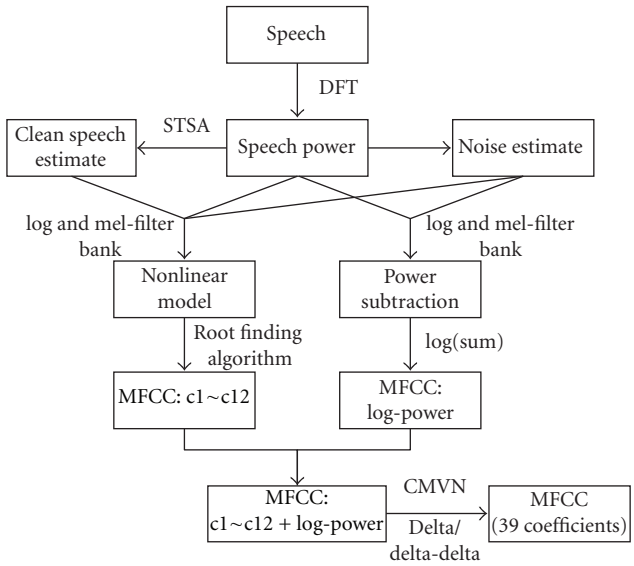


Figure 5: Diagram of the proposed algorithm.

method. The AURORA2 data is based on a version of the original TIDigits (as available from LDC) sampled to 8 kHz [14]. Noise is artificially added at several SNRs (20 dB, 15 dB, 10 dB, 5 dB, 0 dB, −5 dB). Set A and Set B are filtered with a G712 [15] characteristic filter, which simulates the response of filters found in the A/D interface of PCM transmission systems. Set C is filtered with MIRS filter to simulate a telephone system. There are two training conditions in AURORA2, clean training set and multi-condition training set. For clean training condition, the training set has no noise added and it consists of 8440 utterances recorded from 55 male and 55 female adults. 4004 utterances from 52 male and 52 female speakers are split equally into 4 subsets with 1001 utterances each, with all speakers being present in each subset. In the multi-condition training set, four types of noises have been added at various SNR levels [14, 16].

4.1.2. System Description. The proposed front-end feature extractor is modified from the MFCC model provided by Voicebox Toolkit [17]. The demo scripts from the AURORA2 database are used for training. In the evaluation experiments, log-energy (log E) together with c1 to c12 is used as the static feature vector, and then the delta and delta-delta features are calculated using the frame-differential.

The same recognizer is used for both the proposed front-end feature extraction algorithm and the baseline system for comparison. Each digit is modeled by a simple left-to-right 18 states (including two non-emitting states) HMM model, with 3 Gaussian mixtures per state. Two pause models are defined. One is "sil", which has 3 HMM states and models the pauses before and after each utterance. The other one is "sp", which is a single state model (tied with the middle state of "sil") and models the pauses among words.

Table 1: Detailed recognition rates (%).

Iteration No.	1	*2*	3	4	5
Clean	99.09	*99.08*	98.37	98.36	97.23
Avg 0–20	**85.70**	***85.76***	**85.66**	**80.07**	**76.17**
−5 dB	27.80	*27.85*	23.24	23.24	21.31

4.1.3. Comparison Targets. The proposed algorithm does not need stereo data input. Therefore, algorithms such as SPLICE [18] are not selected for comparison, since comparison between algorithms with and without clean speech input is unfair. Because the proposed method is developed based on MFCC, it is chosen to be the baseline. The diagram is given in Figure 6.

MMSE-STSA is the standard MMSE approach for mathematically recovering the clean speech. In this paper, the STSA algorithm is implemented with minimum statistics as the noise estimation part. The log-power subtraction approach is similar to SS, so it is chosen to show that in speech recognition log-power subtraction is much better than SS. MVA is a cepstral domain approach, which is chosen to show the superiority of the proposed algorithm in relevant area. Figure 7 shows the diagram of MVA.

The ETSI standard advanced front-end feature extraction algorithm (AFE) is also implemented for comparison [9].

4.2. Results and Discussion

4.2.1. Experimental Results. Experiments are conducted to show the speech recognition results of the proposed NLPS algorithm with different iterations. Detailed recognition results are given in Table 1. It can be easily found out that the optimal result comes at the second iteration.

Comparison is made against MFCC, MMSE-STSA [3], Spectral Subtraction (SS) [8], Cepstral Mean Variance Normalization (CMVN), AFE [9], and MVA [10].

There are two training conditions in the AURORA2 database demo, clean-training condition and multi-training condition. In the multi-training condition noisy speech together with the clean speech are used for training HMMs. Therefore, the recognition results from multi-training condition are very good. For most of the SNR levels, the recognition results are over 90%. It makes all of the above mentioned methods yields similar recognition results, about

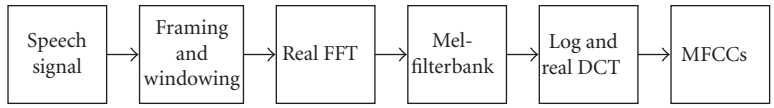


Figure 6: Diagram of MFCC.

Table 2: Detailed recognition rates (%).

SNR	Set A					Set B			Set C	
	Subway	Babble	Car	Exhibition	Station	Restaurant	Street	Airport	Restaurant	Street
Clean	98.83	99.09	99.14	99.29	98.83	99.09	99.14	99.29	98.93	99.15
Avg 0–20	**85.96**	**86.72**	**87.77**	**82.81**	**86.80**	**86.84**	**88.33**	**87.94**	**83.15**	**84.82**
−5 dB	29.23	26.18	31.40	29.37	29.63	29.50	31.26	30.64	21.80	26.72

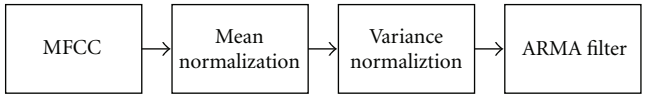


Figure 7: Diagram of MVA.

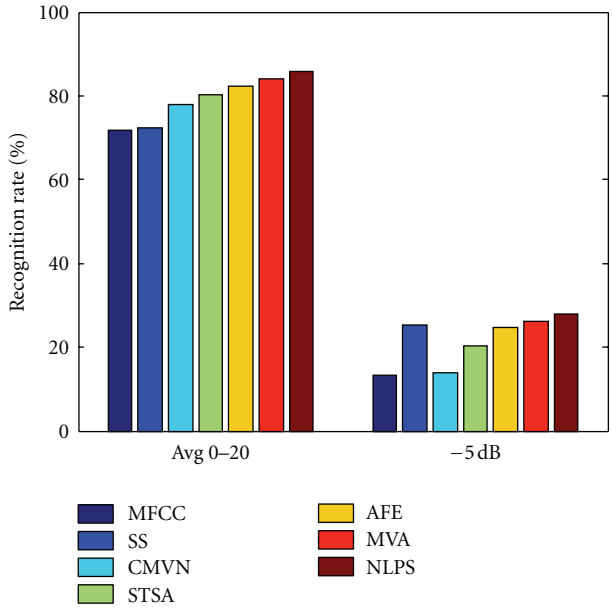


Figure 8: Experimental results.

92% on average. Actually, it is meaningless to make the recognition results increase from 92.1% to 92.5%. Besides, in real life preparing a noisy database for training HMMs is not realistic. It is because there are infinite types of noise and SNRs, which makes it difficult to generate an effective database for training. Moreover, if the noise encountered is very different from that in the database, bad results will be obtained. Therefore, only the clean training condition results are used for comparison. The experiment results are shown in Tables 2, 3, 4, and 5.

Table 3: Recognition results for different parts of the proposed algorithm.

	Clean	Avg 0–20	−5
CMVN	99.32	**77.78**	13.90
LPS	99.47	**73.07**	12.92
LPS + CMVN	99.07	**83.30**	20.78
Newton	99.20	**83.83**	25.65
Newton + CMVN	99.25	**84.96**	27.78
NLPS (Newton + LPS + CMVN)	*99.08*	***85.76***	*27.85*

Table 4: Recognition results for comparison targets.

	Clean	Avg 0–20	−5
MFCC	99.42	**71.80**	13.39
SS	99.32	**72.42**	25.26
CMVN	99.32	**77.78**	13.90
STSA	99.26	**80.12**	20.31
AFE	99.20	**82.23**	24.77
MVA	99.20	**84.15**	26.24

In Table 3, LPS stands for log-power subtraction. LPS + CMVN means only LPS and CMVN are implemented in the speech recognition system. Newton refers to the system with only Newton's iterative method. Newton + CMVN means both methods are implemented. NLPS is the final form of the proposed algorithm which involves the implementation of all the three methods, Newton's method, LPS and CMVN. The experimental results in Table 3 are given to show that the three parts of the proposed algorithm all helps to improve the performance of speech recognition system.

In the following discussion, MFCC denotes the traditional 13 Mel Frequency Cepstral Coefficients together with the corresponding velocity and acceleration parameters. Results are averaged over the noisy test sets with SNRs from 0 to 20 dB, denoted as Avg 0–20. Another point that has to be mentioned is that the clean set results of all the above mentioned algorithms are over 99%. It is also meaningless to attempt to achieve significant improvements at this level.

Table 5: Recognition results for comparison targets.

	Avg 0–20	Relative Imp.	SNR − 5	Relative Imp.
MFCC	**71.80**	19.4%	**13.39**	108.0%
SS	**72.42**	18.4%	**25.26**	10.3%
CMVN	**77.78**	10.3%	**13.90**	100.4%
STSA	**80.12**	7.0%	**20.31**	37.1%
AFE	**82.23**	4.3%	**24.77**	12.4%
MVA	**84.15**	1.9%	**26.24**	6.1%

Therefore, discussion will be carried out mainly for Avg 0–20 and SNR −5 dB. Figure 8 shows the experimental results at Avg 0–20 and SNR −5 dB.

4.2.2. Results Analysis. Experimental results in Table 3 show that the implementation of LPS and Newton's method greatly improves the recognition results. Besides, the two fundamental parts, LPS and Newton's method, both contribute to the excellent performance of the proposed algorithm. As shown in Table 3, Newton's method alone can reach a recognition rate of 83.83% at Avg 0–20. With the combination of LPS and CMVN, the performance of the speech recognition system is further improved, 84.96% for Newton + CMVN and 85.76% for the proposed NLPS algorithm.

Comparisons in Tables 3 and 4 show that the proposed algorithm significantly improved the performance of speech recognition system. The relative improvement ratios are shown in Table 5. Compared with the baseline MFCC system, the proposed algorithm achieves very impressive improvements, 19.4% in terms of Avg 0–20 and *108*% in SNR −5 dB. For CMVN and STSA, also very significant improvements are reached. In the level of Avg 0–20, the relative improvements are 10.3% over CMVN, and 7% over STSA. When it comes to SNR −5 dB, the improvements become much more significant, *100.4*% over CMVN and 37.1% over STSA.

In speech processing technique, there is a kind of awkward situation when speech enhancement algorithm sometimes cannot improve the speech recognition results even if it manages to improve speech quality in terms of human listening test. SS is just one of the above mentioned methods. Direct implementation of the SS in [8] yields terrible results. Therefore, in our evaluation test, the noise estimation part of SS is replaced by Minimum Statistics [19]. In terms of Avg 0–20, the relative improvement reaches 18.4%. At SNR −5 dB the relative improvement is 10.3%. The performance of SS can successfully support the novelty of the LPS method, which is an indispensible part of the proposed NLPS algorithm. As for MVA, admittedly it is a very successful algorithm. However, the proposed algorithm still yields better results. At Avg 0–20, a 1.9% improvement is reached. For SNR −5 dB, the relative improvement is 6.1%. For the European Telecommunications Standards Institute (ETSI) standard AFE, at Avg 0–20, a relative improvement of 4.3% is reached. For SNR −5 dB, the relative improvement is 12.4%.

5. Conclusion

In this paper, a novel algorithm for robust speech recognition system is presented with its detailed derivation, implementation, and evaluation. It is based on the direct solution of a nonlinear system model together with a novel log-power subtraction method. The novelty of the proposed algorithm lies in four parts. Firstly, the proposed method does not need any additional training process, which makes the computational burden very small. Besides, the proposed method is a blind approach, which means that the proposed method yields good performance at all SNRs and noise types. Another advantage of the proposed algorithm is its ability to adapt to changing environments. The adaptation can be made by simply changing the noise estimation part. Finally, the NLPS algorithm can be easily combined with other algorithm, such as the MVA discussed above. The proposed algorithm is implemented and evaluated with AURORA2 database. Comparison is made against STSA, SS, and MVA. Experimental results demonstrate significant improvement in the recognition accuracy.

References

[1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, Ny, USA, 1993.

[2] B. Gold and N. Morgan, *Speech and Audio Signal Processing—Processing and Perception of Speech and Music*, John Wiley & Sons, 2000.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust speech recognition using a cepstral minimum-mean-square-error- motivated noise suppressor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 1061–1070, 2008.

[5] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1654–1661, 2008.

[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1218–1233, 2006.

[7] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static

and dynamic features," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 218–233, 2004.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans Acoust Speech Signal Process*, vol. 27, no. 2, pp. 113–120, 1979.

[9] European Telecommunications Standards Institute (ETSI), ETSI ES 202 050 V1.1.5, 2007.

[10] C. Chia-Ping and A. B. Jeff, "MVA processing of speech features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[11] A. Acero, *Acoustical and environmental robustness in automatic speech recognition [Ph.D. thesis]*, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1990.

[12] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[13] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[14] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition system under noisy conditions," in *Proceedings of the 7th international conference on Information, communications and signal processing (ICICS '09)*, Paris, France, 2000.

[15] ITU-T, Recommendation G.712. Transmission Performance Characteristics for Pulse Code Modulation Channels, Geneva, Switzerland, 1996.

[16] R. G. Leonard, "A database for speaker independent digit recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84*, vol. 3, pp. 42–53, 1984.

[17] M. Brookes, Voicebox, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[18] J. Droppo, A. Acero, and L. Deng, "Evaluation of the SPLICE algorithm on the Aurora 2 database," in *Proceedings of the Eurospeech Conference, International Speech Communication Association*, Aalbodk, Denmark, September 2001.

[19] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of the European Signal Processing Conference (EUSIPCO '96)*, pp. 1182–1185, 1994.