A Speech Enhancement Approach Using Piecewise Linear Approximation of An Explicit Model of Environmental Distortions

Jun Du[1,2], *Qiang Huo*[1]

[1]Microsoft Research Asia, Beijing, P. R. China
[2]University of Science and Technology of China, Hefei, Anhui, P. R. China
unuedjwj@ustc.edu, qianghuo@microsoft.com

Abstract

This paper presents a speech enhancement approach derived by using a piecewise linear approximation (PLA) of an explicit model of environmental distortions. PLA is a generalization of two traditional approaches, namely vector Taylor series (VTS) and MAX approximations. Formulations are described for both maximum likelihood (ML) estimation of noise model parameters and minimum mean-squared error (MMSE) estimation of clean speech. Evaluation experiments are conducted to enhance speech signals corrupted by several types of additive noises. Compared to the traditional MAX-approximation based approach, our PLA-based speech enhancement approach achieves better performance in terms of two objective quality measures, namely segmental SNR and log-spectral distortion.

Index Terms— speech enhancement, piecewise linear approximation, distortion model.

1. Introduction

Enhancing noisy speech captured by a single microphone for improving listening experience by human has been a long standing research problem in the past several decades. Among many approaches developed over the years (e.g., [11, 4, 2] and the references therein), we are interested in a class of speech enhancement algorithms which are derived from three key elements, namely a statistical reference clean speech model pretrained from some training data, a noise model with parameters estimated from the noisy speech to be enhanced, and an *explicit* distortion model characterizing how speech is distorted (e.g., [4, 5, 6, 1]).

In particular, for the approach described in [1], it is assumed that in the time domain, the "corrupted" speech $y^{\mathrm{t}}(l)$ is subject to the following *explicit* distortion model:

$$y^{\mathrm{t}}(l) = x^{\mathrm{t}}(l) + n^{\mathrm{t}}(l) \qquad (1)$$

where independent signals $x^{\mathrm{t}}(l)$ and $n^{\mathrm{t}}(l)$ represent the l^{th} sample of clean speech and additive noise, respectively. Then in frequency domain, we have

$$\mathbf{y}^{\mathrm{f}} = \mathbf{x}^{\mathrm{f}} + \mathbf{n}^{\mathrm{f}} \qquad (2)$$

where $\mathbf{y}^{\mathrm{f}}$, $\mathbf{x}^{\mathrm{f}}$ and $\mathbf{n}^{\mathrm{f}}$ represent the spectra of noisy speech, clean speech and additive noise, respectively. By ignoring correlations among different frequency bins, the distortion model in the log-power-spectral domain can be expressed *approximately* as

$$\exp(\mathbf{y}^{\mathrm{l}}) = \exp(\mathbf{x}^{\mathrm{l}}) + \exp(\mathbf{n}^{\mathrm{l}}) \qquad (3)$$

where $\mathbf{y}^{\mathrm{l}}$, $\mathbf{x}^{\mathrm{l}}$ and $\mathbf{n}^{\mathrm{l}}$ are log-power spectra of noisy speech, clean speech and noise, respectively. The nonlinear nature of the above distortion model makes statistical modeling and inference of the above variables difficult, therefore certain approximations have to be made. In [1], a so-called MAX approximation is used, which was developed originally in [13] for robust automatic speech recognition (ASR). In this paper, we propose to use a more accurate approximation, namely a piecewise linear approximation (PLA) of the above nonlinear distortion model, to derive a new speech enhancement algorithm. It is noted that we had also used the above PLA approximation to derive a feature compensation approach for robust ASR, whose results were reported in [3].

The rest of the paper is organized as follows. In Section 2, we introduce our PLA-based method for speech enhancement. In Section 3, we present evaluation results. Finally, we conclude the paper in Section 4.

2. Our Approach

2.1. System Overview

A block diagram of our speech enhancement system is illustrated in Fig. 1. In the training stage, as in [1], a Gaussian mixture model (GMM) with diagonal covariance matrices is trained from clean speech using log-power spectra features. Let's use $\{\omega_m, m = 1, 2, \cdots, M\}$ to denote the set of M mixture coefficient weights. In the enhancement stage, by ignoring the correlations among different frequency bins, we can do feature compensation in the log-power-spectral domain for different frequency bins independently. In the following subsections, we elaborate on several modules in Fig. 1.

2.2. Feature Extraction

First, we apply a short-time Fourier analysis to the input signal by computing the DFT of each overlapping windowed frame:

$$y^{\mathrm{f}}(k) = \sum_{l=0}^{L-1} y^{\mathrm{t}}(l)h(l)e^{-j2\pi kl/L} \qquad k = 0, 1, \cdots, L-1. \quad (4)$$

where k is the frequency bin index, $h(l)$ denotes the window function (Hamming window here). Then log-power spectra are defined as

$$y^{\mathrm{l}}(k) = \log|y^{\mathrm{f}}(k)|^2 \qquad k = 0, 1, \cdots, K-1 \qquad (5)$$

where $K = L/2 + 1$. For $k = K, \cdots, L-1$, $y^{\mathrm{l}}(k)$ may be obtained using symmetry $y^{\mathrm{l}}(k) = y^{\mathrm{l}}(L-k)$. The relations among $y^{\mathrm{t}}(l)$, $y^{\mathrm{f}}(k)$, $y^{\mathrm{l}}(k)$, and phase information $\angle y^{\mathrm{f}}(k)$ are shown in the feature extraction module of Fig. 1.

This work has been done when the first author was an intern at Microsoft Research Asia, Beijing, China.

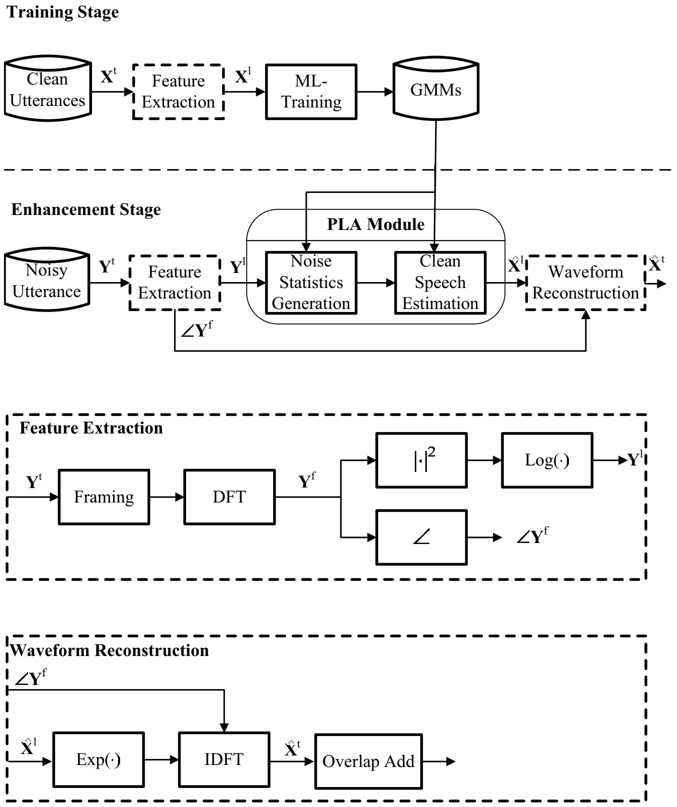
Training Stage
Clean Utterances
$\mathbf{X}^{\mathrm{t}}$
Feature Extraction
$\mathbf{X}^{\mathrm{l}}$
ML-Training
GMMs
Enhancement Stage
PLA Module
Noisy Utterance
$\mathbf{Y}^{\mathrm{t}}$
Feature Extraction
$\mathbf{Y}^{\mathrm{l}}$
Noise Statistics Generation
Clean Speech Estimation
$\hat{\mathbf{X}}^{\mathrm{l}}$
Waveform Reconstruction
$\hat{\mathbf{X}}^{\mathrm{t}}$
$\angle\mathbf{Y}^{\mathrm{f}}$
Feature Extraction
$\mathbf{Y}^{\mathrm{t}}$
Framing
DFT
$\mathbf{Y}^{\mathrm{f}}$
$|\cdot|^2$
Log(·)
$\mathbf{Y}^{\mathrm{l}}$
$\angle$
$\angle\mathbf{Y}^{\mathrm{f}}$
Waveform Reconstruction
$\angle\mathbf{Y}^{\mathrm{f}}$
$\hat{\mathbf{X}}^{\mathrm{l}}$
Exp(·)
IDFT
$\hat{\mathbf{X}}^{\mathrm{t}}$
Overlap Add

Figure 1: A block diagram of our speech enhancement system.

2.3. Waveform Reconstruction

After we obtain the estimation of the log-power spectrum of clean speech, $\hat{x}^{\mathrm{l}}(k)$, from PLA module, the reconstructed spectrum $\hat{x}^{\mathrm{f}}(k)$ is given by

$$\hat{x}^{\mathrm{f}}(k) = \exp\{\hat{x}^{\mathrm{l}}(k)/2\}\exp\{j\angle y^{\mathrm{f}}(k)\} \ , \tag{6}$$

where the phase information $\angle y^{\mathrm{f}}(k)$ is derived from the original noisy speech. Then a frame of speech signal, $\{\hat{x}^{\mathrm{t}}(l); l = 0, 1, \cdots, L-1\}$, is reconstructed by computing inverse DFT (IDFT) of the current frame of spectrum as follows:

$$\hat{x}^{\mathrm{t}}(l) = \frac{1}{L}\sum_{k=0}^{L-1}\hat{x}^{\mathrm{f}}(k)e^{j2\pi kl/L} \ . \tag{7}$$

Waveform for the whole utterance can then be synthesized by using a traditional overlap-add procedure as described in [8] where the same Hamming window as in speech analysis step is used for waveform synthesis.

2.4. PLA Module

Let's assume the noise feature vector $\mathbf{n}^{\mathrm{l}}$ follows a Gaussian PDF (probability density function) with a mean vector $\boldsymbol{\mu}_{\mathbf{n}}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{n}}$ respectively. We have studied two ways of estimating $\{\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\}$. The first approach simply takes the sample mean and covariance of the relevant features from the first several (10 in our experiments) frames of the noisy speech utterance. The second approach uses a maximum likelihood (ML) estimation of $\{\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\}$ from the whole noisy speech utterance with T frames of observations, which can be solved by using EM algorithm iteratively (e.g., [15, 10]). The EM updating formulas are as follows:

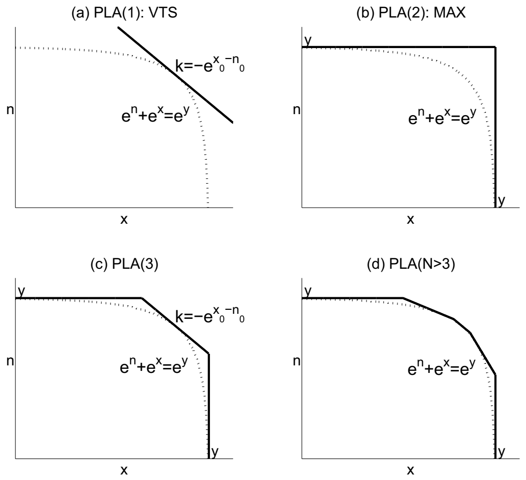

Figure 2: An illustration of several special cases of PLA.

$$\overline{\boldsymbol{\mu}}_{\mathbf{n}} = \frac{\sum_{t=0}^{T-1}\sum_{m=1}^{M} P(m|\mathbf{y}_t^{\mathrm{l}})E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m]}{\sum_{t=0}^{T-1}\sum_{m=1}^{M} P(m|\mathbf{y}_t^{\mathrm{l}})} \tag{8}$$

$$\overline{\boldsymbol{\Sigma}}_{\mathbf{n}} = \frac{\sum_{t=0}^{T-1}\sum_{m=1}^{M} P(m|\mathbf{y}_t^{\mathrm{l}})E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}(\mathbf{n}_t^{\mathrm{l}})^{\top}|\mathbf{y}_t^{\mathrm{l}}, m]}{\sum_{t=0}^{T-1}\sum_{m=1}^{M} P(m|\mathbf{y}_t^{\mathrm{l}})} - \overline{\boldsymbol{\mu}}_{\mathbf{n}}\overline{\boldsymbol{\mu}}_{\mathbf{n}}^{\top} \tag{9}$$

where

$$P(m|\mathbf{y}_t^{\mathrm{l}}) = \frac{\omega_m p_{\mathbf{y}}(\mathbf{y}_t^{\mathrm{l}}|m)}{\sum_{l=1}^{M}\omega_l p_{\mathbf{y}}(\mathbf{y}_t^{\mathrm{l}}|l)} \ . \tag{10}$$

In the above equations, $p_{\mathbf{y}}(\mathbf{y}_t^{\mathrm{l}}|m)$ is the PDF of the noisy speech $\mathbf{y}_t^{\mathrm{l}}$ for the m^{th} component of the compensated noisy speech mixture of densities, $E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m]$ and $E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}(\mathbf{n}_t^{\mathrm{l}})^{\top}|\mathbf{y}_t^{\mathrm{l}}, m]$ are the relevant conditional expectations, t is the frame index.

Given the noisy speech and noise estimation, the minimum mean-squared error (MMSE) estimation of clean speech can be calculated as

$$\hat{\mathbf{x}}_t^{\mathrm{l}} = E_{\mathbf{x}}\left[\mathbf{x}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}\right] = \sum_{m=1}^{M} P(m|\mathbf{y}_t^{\mathrm{l}})E_{\mathbf{x}}\left[\mathbf{x}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m\right] \tag{11}$$

where $E_{\mathbf{x}}\left[\mathbf{x}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m\right]$ is the conditional expectation of $\mathbf{x}_t^{\mathrm{l}}$ given $\mathbf{y}_t^{\mathrm{l}}$ for the m^{th} mixture component. Finally, the estimated clean-speech features in the log-power-spectral domain are converted to time domain through waveform reconstruction.

To implement the above feature compensation approach, the key technical issues become how to calculate $p_{\mathbf{y}}(\mathbf{y}_t^{\mathrm{l}}|m)$, $E_{\mathbf{x}}\left[\mathbf{x}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m\right]$, $E_{\mathbf{x}}\left[\mathbf{x}_t^{\mathrm{l}}(\mathbf{x}_t^{\mathrm{l}})^{\top}|\mathbf{y}_t^{\mathrm{l}}, m\right]$, $E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}|\mathbf{y}_t^{\mathrm{l}}, m]$, and $E_{\mathbf{n}}[\mathbf{n}_t^{\mathrm{l}}(\mathbf{n}_t^{\mathrm{l}})^{\top}|\mathbf{y}_t^{\mathrm{l}}, m]$, respectively. The complete set of formulas can be found in [3].

As we discussed in [3], both the MAX approximation and the first-order vector Taylor series (VTS) approximation (e.g., [12]) can be treated as special cases of PLA as illustrated in Fig. 2.

Table 1: Comparison of segmental SNR (SegSNR) of noisy speech with that of enhanced speech by “NO_MAX” approach.

Input	AWGN		Babble Noise		Restaurant Noise		Street Noise	
SNR[dB]	Noisy	NO_MAX	Noisy	NO_MAX	Noisy	NO_MAX	Noisy	NO_MAX
20	10.19	12.03	10.66	11.14	11.29	11.28	10.27	10.78
15	5.60	8.53	6.07	6.68	6.58	7.10	5.65	6.30
10	1.12	4.87	1.60	2.83	2.04	2.99	1.16	2.59
5	-3.26	1.86	-2.76	-0.03	-2.39	-0.81	-3.23	-0.53
0	-7.41	-0.75	-6.90	-2.80	-6.60	-4.61	-7.39	-3.65
-5	-11.20	-3.32	-10.75	-4.73	-10.58	-8.17	-11.19	-6.52
-10	-14.54	-5.28	-14.16	-6.86	-14.17	-11.64	-14.53	-9.29

Table 2: Comparison of log-spectral distortion (LSD) of noisy speech with that of enhanced speech by “NO_MAX” approach.

Input	AWGN		Babble Noise		Restaurant Noise		Street Noise	
SNR[dB]	Noisy	NO_MAX	Noisy	NO_MAX	Noisy	NO_MAX	Noisy	NO_MAX
20	4.79	2.53	3.02	1.96	2.67	2.07	3.29	2.34
15	7.86	3.48	4.57	2.89	4.57	3.19	5.44	3.46
10	11.41	4.37	6.62	4.08	7.31	4.58	8.38	4.71
5	15.31	5.53	9.25	5.42	10.65	6.22	11.93	6.12
0	19.48	6.60	12.43	7.35	14.46	8.26	15.87	7.53
-5	23.85	8.03	16.12	9.46	18.59	10.34	20.13	9.25
-10	28.37	9.39	20.17	12.21	22.96	13.40	24.61	11.50

3. Evaluation Experiments

3.1. Experimental Setup

In our experiments, the following three enhancement methods are compared:

- **NO_MAX**: the reference speech enhancement system as described in [1], where noise re-estimation is NOT conducted and MAX approximation is used for MMSE estimation of clean speech spectrum;
- **NO_PLA**: noise re-estimation is NOT conducted and PLA(3) approximation is used for MMSE estimation of clean speech spectrum;
- **MAX_PLA**: MAX approximation is used for noise re-estimation and PLA(3) approximation is used for MMSE estimation of clean speech spectrum.

The noise signals used in our evaluation include synthetic additive white Gaussian noises (AWGN) and three other types of noise recordings extracted from Aurora2 database [9], namely *babble* (crowd of people), *restaurant*, and *street*. A total of 800 clean-speech sentences from 40 males and 40 females (10 sentences per speaker) are selected from TIMIT database [7] for training the clean speech GMM with 32 Gaussian components. Another 1680 clean-speech sentences from 112 males and 56 females (10 sentences per speaker) in TIMIT database are used to generate the set of testing sentences for each combination of noise type and SNR (as measured in Aurora2 database). All the speech samples from TIMIT database are down-sampled to 8KHz. For speech analysis, the frame length is set as $L = 256$ and the frame shift $R = 128$. When applicable, the number of EM iterations for noise re-estimation is set as 4.

The performance is evaluated by two objective quality measures and informal listening tests. The first objective quality measure is segmental SNR (SegSNR, in dB) defined as follows (e.g., [14, 2]):

$$\mathrm{SegSNR} = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{C}_1 \left\{ 10\lg \frac{\sum_{l=0}^{L-1}[x^{\mathrm{t}}(l+tR)]^2}{\sum_{l=0}^{L-1}[x^{\mathrm{t}}(l+tR)-\hat{x}^{\mathrm{t}}(l+tR)]^2} \right\} \tag{12}$$

where T denotes the number of frames in the signal, and

$$\mathcal{C}_1(z) = \min\left[\max(z, -20),\ 30\right] \ . \tag{13}$$

The above operator $\mathcal{C}_1(\cdot)$ confines the SNR at each frame to the perceptually meaningful range between 30dB and -20dB, which prevents the segmental SNR measure from being biased in either a positive or negative direction due to a few silence or unusually high SNR frames, because they do not contribute significantly to the overall speech quality.

The second objective quality measure is log-spectral distortion (LSD, in dB) defined as follows (e.g., [2]):

$$\mathrm{LSD} = \frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{L/2+1}\sum_{k=0}^{L/2}\left[10\lg\frac{\mathcal{C}_2(x_t^{\mathrm{f}}(k))}{\mathcal{C}_2(\hat{x}_t^{\mathrm{f}}(k))}\right]^2\right\}^{\frac{1}{2}} \tag{14}$$

where

$$\mathcal{C}_2(z_t^{\mathrm{f}}(k)) = \max\left[|z_t^{\mathrm{f}}(k)|^2,\ 10^{-50/10}\max_{t,k}(|z_t^{\mathrm{f}}(k)|^2)\right] \tag{15}$$

is the clipped power spectrum such that the dynamic range of the log-spectrum is confined to about 50dB.

3.2. Evaluation Results

Table 1 summarizes a comparison of SegSNR of noisy speech with that of enhanced speech by “NO_MAX” approach, where “Input SNR” denotes the global SNR of the testing utterances as measured in Aurora2 database [9]. A similar comparison is made in terms of LSD in Table 2. It is observed that NO_MAX approach achieves significant performance improvement in all cases with different noises and SNR levels. This is especially

Table 3: Comparison of segmental SNRs (SegSNR) of enhanced speech by NO_MAX, NO_PLA, MAX_PLA approaches respectively.

Input	AWGN			Babble Noise			Restaurant Noise			Street Noise		
SNR[dB]	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA
5	1.86	2.12	3.04	-0.03	-0.02	1.22	-0.81	-0.68	0.52	-0.53	-0.40	1.77
0	-0.75	-0.51	0.86	-2.80	-2.47	-1.19	-4.61	-4.40	-2.68	-3.65	-3.41	-0.17
-5	-3.32	-2.95	-1.11	-4.73	-4.40	-3.35	-8.17	-8.02	-5.14	-6.52	-6.23	-2.07
-10	-5.28	-5.15	-2.31	-6.86	-6.37	-5.37	-11.64	-11.40	-7.37	-9.29	-8.94	-3.62

Table 4: Comparison of log-spectral distortion (LSD) of enhanced speech by NO_MAX, NO_PLA, MAX_PLA approaches respectively.

Input	AWGN			Babble Noise			Restaurant Noise			Street Noise		
SNR[dB]	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA	NO_MAX	NO_PLA	MAX_PLA
5	5.53	5.42	5.67	5.42	5.36	5.15	6.22	5.98	5.58	6.12	5.97	5.41
0	6.60	6.50	6.40	7.35	7.27	6.79	8.26	7.92	7.20	7.53	7.36	6.60
-5	8.03	7.79	7.52	9.46	9.26	8.99	10.34	9.98	8.73	9.25	8.93	7.91
-10	9.39	9.35	8.25	12.21	11.68	11.41	13.40	12.86	10.83	11.50	10.95	8.89

true for AWGN, because the relevant modeling assumptions are more accurate in this case.

Table 3 summarizes a comparison of SegSNR of enhanced speech by “NO_MAX”, “NO_PLA”, “MAX_PLA” approaches, respectively. A similar comparison is made in terms of LSD in Table 4. Here only the results for those cases with SNR below 10dB are given because the reference method “NO_MAX” is good enough for cases with SNR above 10dB and there is no big performance difference among three approaches compared. By comparing “NO_PLA” with ”NO_MAX”, it is observed that both SegSNR and LSD measures are improved in all cases when the more accurate PLA is used in MMSE estimation of clean speech. By comparing “MAX_PLA” with “NO_PLA”, it is also observed that both SegSNR and LSD measures are improved in almost all cases when noise re-estimation is conducted. Apparently, “MAX_PLA” achieves the best performance.

We have also compared spectrograms of noisy speech and enhanced speech with different approaches. Compared with “NO_MAX”, we found that “MAX_PLA” can suppress better the residual noise after speech enhancement. This is confirmed by informal listening tests as well.

To understand the computational complexity of the proposed “MAX_PLA” approach, a timing experiment is conducted on a “Pentium-4” PC with a 3GHz clock by using three testing sentences with different lengths of 1s, 1.93s, and 7.47s, respectively. The total *User CPU Time* for enhancing the above three noisy sentences are 2.78s, 5.52s, and 21.33s, respectively. The main overhead comes from the noise re-estimation part.

4. Summary

We have presented a new speech enhancement approach derived by using a piecewise linear approximation (PLA) of an explicit model of environmental distortions. Formulations are described for both ML estimation of noise model parameters and MMSE estimation of clean speech. Evaluation experiments are conducted to enhance speech signals corrupted by several types of additive noises. Compared to the traditional MAX-approximation based approach, our PLA-based speech enhancement approach achieves better performance in terms of segmental SNR and log-spectral distortion.

5. References

[1] D. Burshtein and S. Gannot, “Speech enhancement using a mixture-maximum model,” *IEEE Trans. on Speech and Audio Processing*, Vol.10, No.6, pp.341-351, 2002.

[2] I. Cohen and S. Gannot, “Spectral enhancement methods,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, Y. Huang (Eds.), Springer 2008, pp.873-901.

[3] J. Du and Q. Huo, “A feature compensation approach using piecewise linear approximation of an explicit distortion model for noisy speech recognition,” *Proc. ICASSP*, 2008, pp.4721-4724.

[4] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, Vol.67, pp.1526-1555, 1992.

[5] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Processing*, Vol.40, No.4, pp.725-735, 1992.

[6] Y. Ephraim, D. Malah, and B.-H. Juang, “On the application of hidden Markov models for enhancing noisy speech,” *IEEE Trans. on ASSP*, Vol.37, No.12, pp.1846-1856, 1989.

[7] J. S. Garofolo, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, NIST Technical Report, 1988.

[8] D. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. on ASSP*, Vol.32, No.2, pp.236-243, 1984.

[9] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” *Proc. ISCA ITRW ASR*, 2000, pp.181-188.

[10] D.-Y. Kim, C.-K. Un, and N.-S. Kim, “Speech recognition in noisy environments using first-order vector Taylor series,” *Speech Communication*, Vol.24, pp.39-49, 1998.

[11] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, Vol.67, No.12, pp.1586-1604, 1979.

[12] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” *Proc. ICASSP*, 1996, pp.733-736.

[13] A. Nádas, D. Nahamoo, and M. A. Picheny, “Speech recognition using noise-adaptive prototypes,” *IEEE Trans. on ASSP*, Vol.37, No.10, pp.1495-1503, 1989.

[14] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.

[15] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, “Integrated models of signal and background with application to speaker identification in noise,” *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.2, pp.245-257, 1994.