SPEECH ENHANCEMENT BASED ON SNR-DEPENDENT EMPIRICAL STATISTICAL ESTIMATION IN LOG-SPECTRAL MAGNITUDE DOMAIN

Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura

Nagoya University, Japan

dat@sp.m.is.nagoya-u.ac.jp

Abstract

We present a data-driven speech enhancement system based on empirical statistical estimations of speech in the log-spectral magnitude domain, where the enhancement filter is trained at each SNR index. We use a estimation method called SNRGMM, which were developed in our previous work, to cluster the training data and learn the enhancement filter at each SNR index. This measurement is later used in processing to switch noisy speeches to a nearest trained filter. For the enhancement filter training, instead of conventional the code-book dependent piecewise transform, we develop a general statistical estimations based on empirical joint or cumulative density. The empirical minimum mean square error (EMMSE), maximum a posterior probability (EMAP) and cumulative histogram equalization (CHEQ) methods are implemented and investigated in this work. A simulation mode for the cases, when a pair of noisy and clean speech databases are unavailable, is also considered. In the experimental evaluation, we use the AURORA2 for learning the enhancement filters and apply them to the AURORA2 Japanese version. The experiment results show improvements in both SNR and ASR performances of the proposed methods. Among the empirical estimators, the CHEQ shows better results in ASR with the approximately 62 percents relative improvements in the clean training and 41 percents relative improvement in the multi condition training. Moreover, the CHEQ method is shown to be effective with smaller number of samples in training database.

1. Introduction

Noise reduction is an important problem of speech processing, where the statistical methods are frequently used. Among this group of methods, both the model-based and data-driven approaches have been studied. The model-based approaches [1] require an explicit in model assumptions and are not sufficient for some real conditions, when we are dealing with long time reverberation or some types of noise, which contain other sources. The data-driven approaches have been studied mainly for speech recognition [2]. The SNR-dependent cepstral normalization [2] uses instantaneous SNR, which is estimated in the same manner as in the model-based approaches and therefore, has the same mentioned above limitations. Later, the codebook-dependent piecewise transform has been more frequently used [2], where the codebook is based VQ or GMM state. The codebook-dependent based on GMM training is also applied for speech enhancement [3]. Though the cepstral normalizations are shown to be superior in ASR, there remains some drawbacks. First of all, the VQ or GMMM codebook theoretically can not guarantee to cluster the training noisy speech data to have an identical noise power level if the SNR measurements are different and therefore the conditional Gaussian assumption might be failed. Moreover, the codebook-dependent piecewise transform can be generalized into actual non-linear if we known the joint distribution. In this work, we proposed a data-driven speech enhancement system, where the SNR measurement is used to cluster or simulate the training data in order to learn the enhancement filters. Furthermore, instead of codebook-dependent piecewise transform, we use the empirical estimators, which are derived as a general non linear transform from noisy to speech data using the joint density or cumulative density estimation. For the SNR estimation, a measurement called SNRGMM, which has been developed in our previous work, is used. In order to cover the waveform reconstruction, the algorithm is done on the log-spectral magnitude domain. In section2, we describe the data clustering or simulation by using the SNRGMM index. Section 3 develops the empirical statistical estimators for the enhancement filtering. Section 4 evaluates the proposed method on AURORA2J database and section 5 summarizes the work.

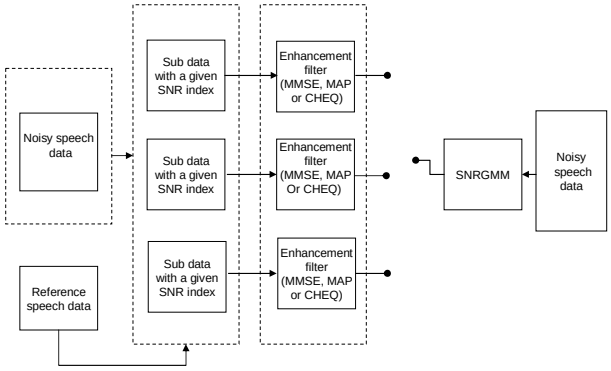

Figure 1: Proposed speech enhancement

2. SNR based data clustering for training the enhancement filters

2.1. Data clustering

The block diagram of SNR based data clustering is in the left side of Figure 1. Each collected noisy speech utterance is passed into SNR estimation and then the training data is clus-

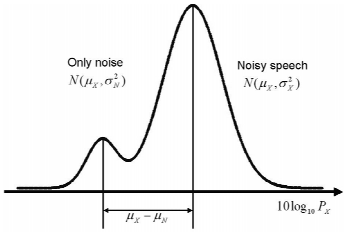

Figure 2: SNRGMM estimation method

tered into SNR-dependent sub-data. Further more, the average noise level, given from SNR estimation, is also pre-normalized to 0dB. The reason of the use of utterance's SNR but not instantaneous SNR is explained by fact of that, this measurement is more and consistent and robust. In fact, the non-stationary noise reduction is done by the empirical estimators, which will be described in next section. NRGMMHowever, the SNR-based clustering should decrease the variance of the estimation and is critically important. For the SNR estimation, we use a segmental SNR estimation method, called S, which has been developed in our previous work [4] for the real conditions, when neither signal nor noise reference nor the speech activity is available.

2.2. SNRGMM

The basic concept of the proposed method is fitting the distributions of noise and noisy speech in the log-power domain from actual noisy speech signal, using the GMM modeling and EM estimation. Denoting the segmental SNR in a notation of expectations as

$$SNR_{seg} = \left\langle 10\log_{10}\frac{P_S}{P_N}\right\rangle = \langle 10\log_{10} P_S\rangle - \langle 10\log_{10} P_N\rangle\,, \tag{1}$$

the segmental SNR is given through fitted distributions. When the original SNR is high, the SNRGMM defined in (1) can be approximated by the total to noise ratio, which is the expectation distance and noted by

$$TNR_{seg} = \left\langle 10\log_{10}\frac{P_X}{P_N}\right\rangle = \langle 10\log_{10} P_X\rangle - \langle 10\log_{10} P_N\rangle\,, \tag{2}$$

$$TNR_{seg} = \mu_X - \mu_N, \tag{3}$$

where $\mu_X > \mu_N$. Under low SNR conditions, this estimation yields a significant error. For that case, a compensation mode is proposed by denoting the SNR as a nonlinear moment of a random variable of the local TNR which is noted by

$$SNR_{seg} = \left\langle 10\log_{10}\left(\frac{P_X}{P_N} - 1\right)\right\rangle. \tag{4}$$

The local TNR is a difference of two Gaussian distributed random variables and therefore is also assumed to be a Gaussian,

$$10\log_{10}\frac{P_X}{P_N} \sim \mathrm{N}\left(x, \mu_X - \mu_N, \sigma_X^2 - \sigma_N^2\right). \tag{5}$$

By using an asymptotic expansion

$$\ln\left(e^r - 1\right) \approx r - 0.7e^{-r} - 0.9e^{-2r} - e^{-3r}, \tag{6}$$

the SNRGMM estimation is given as

$$SNR_{seg} = \frac{10}{\ln 10}\left\{\begin{array}{l} m - 0.7\exp\left[-\left(m - \frac{d}{2}\right)\right] \\ -0.9\exp\left[-2\left(m - d\right)\right] \\ -\exp\left[-3\left(m - \frac{3d}{2}\right)\right] \end{array}\right\}, \tag{7}$$

where

$$m = \frac{\ln 10}{10}\left(\mu_X - \mu_N\right), \quad d = \left(\frac{\ln 10}{10}\right)^2\left(\sigma_X^2 - \sigma_N^2\right). \tag{8}$$

Table 1 shows experimental evaluation on AURORA2 database for SNRGMM estimation. The SNRGMM, yields more accurate and robust estimation with less standard deviation than the conventional methods using VAD [5] or a raised cosine function (NISTSNR) [6]. Note that, one more output of estimation is the average noise level defined by mu_N and this measurement is used to normalize the noisy speech in training database.

Table 1: *Averaged estimation errors evaluated on AURORA2*

SNR	NISTSNR	VADSNR	SNRGMM
20dB	4.0±1.9	0.3±0.5	0.4±0.3
15dB	4.3±2.2	1.2±0.5	0.4±0.6
10dB	4.9±2.1	3.0±1.7	1.1±0.9
5dB	5.8±2.8	4.9±1.5	1.5±1.0
0dB	7.8±3.0	6.5±1.9	2.1±1.7
-5dB	11.2±3.1	10.1±2.8	3.1±1.9

2.3. Simulation mode

In real conditions, a pair of noisy and clean speech data (or recorded at closed talking microphone) might not be available. In these cases, we can use a noise sample to simulate this pair data at each SNRGMM index. Given a clean speech and a random noise sample, we first estimate the SNRGMM of their addition, then a weight coefficient, for multiplying the noise sample, is calculated by the given and initial SNRGMM. In our system, we repeat this procedure three times to better fit the SNR estimation. Note that, this simulation is quick and convenience since SNRGMM can be estimated without knowledge of VAD. We note that, this mode is able to train the enhancement filter in even more wide band of SNR than the real collected data.

3. Empirical estimation for speech enhancement

As it was mentioned above, the conventional data-driven approaches uses the codebook-dependent piecewise transform, where the conditional Gaussian distribution of noisy feature inside each codebook [2] is assumed. Another direction, which is investigated in this work, is the non-parametric learning of the joint or cumulative distributions and the implementation of statistical estimations in general, which can be considered as a use of a "continuous" code-book. Note that, the SNR clustering described in previous section, is a pre-normalization, which should decrease the variances of the estimation. The training for proposed system is in Figure 1. For each SNRGMM index, a pair data of noisy and clean speech is transformed into STDFT domain, and then into phase and log-spectral magnitude. The empirical joint density is estimated in each sub-band log-spectral magnitude domain.

3.1. Joint density estimation

The histogram is simplest form of the density estimation and from our experience, it can be in some cases, successfully used to implement empirical statistical estimation for speech enhancement. However, the histogram has various drawbacks caused by sensitivity to the number and orientation of bins as well as by the discontinuity. The told above make histogram unsuitable to practical developments In this work we use a Gaussian kernel for density estimation [7]. Denote the kernel density of a vector $\mathbf{x}$ as

$$\hat{p}(\mathbf{x}) = \frac{1}{N(2\pi h^2)^{\frac{p}{2}}} \sum_{i=1}^{N} K_h(\mathbf{x}, \mathbf{x}_i), \tag{9}$$

where N, p and h are total number of examples, dimension and bandwidth respectively. The Gaussian kernel is

$$K_h(\mathbf{x}, \mathbf{x}_i) = e^{-\frac{1}{2h^2}(\|\mathbf{x}-\mathbf{x}_i\|)^2}. \tag{10}$$

The bandwidth choice is given by the recommendation in [7] and it gives $h = 1.25$

3.2. Joint density estimation based estimators

In contrast to the codebook-dependent piecewise transformation, we derive a general non-linear transformation from noisy to clean feature (log-spectral magnitude), using the joint density. Both minimum mean square error (MMSE) and maximum a posterior probability (MAP) estimations are implemented. The empirical MMSE (EMMSE) estimation yields the conditional expectation, which is expressed as

$$\hat{S} = E[S|X] = \frac{\int_{-\infty}^{\infty} Sp(S, X)\, d}{\int_{-\infty}^{\infty} p(S, X)\, dS}, \tag{11}$$

where, the joint density $p(X, S)$ is estimated in each sub-band using the Gaussian kernel described in section 3.1. The estimation (11) is implemented by numerical method, where the infinitive boundaries are changed to a finite ones, which are taken from training. A discrete set of X and resulted $\hat{S}$ are saved in memory to implement an empirical transform by interpolation.

$$\hat{S} = g(X) \tag{12}$$

Alternatively, the empirical MAP (EMAP) estimation is denoted by

$$\hat{S} = \arg\max_{S} p(S, X). \tag{13}$$

Analogously, an interpolation based transform is given for further use in processing. One drawback of the joint density estimation is the requirement of quite large number of training data. In next section, we develop the alternative estimation method based on one dimensional cumulative density of noisy and clean speech. This method is shown to be able to apply with quite small number of samples.

3.3. Cumulative histogram equalization

The CHEQ estimation finds a non-linear transform which maps from the cumulative distribution function of the noisy feature to the clean speech

$$\hat{S} = g(X) = F_S^{-1}(F_X(X)). \tag{14}$$

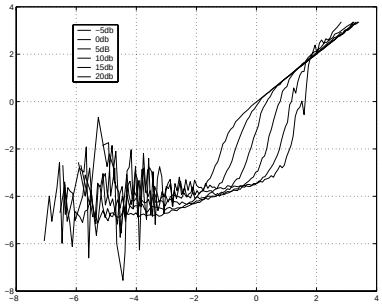

Figure 3: Example of EMMSE filters in sub-band f=1kHZ

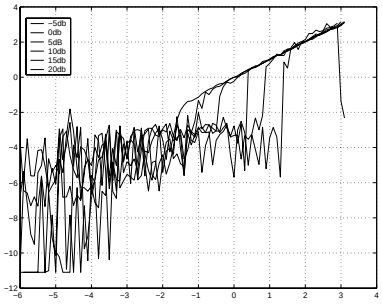

Figure 4: Example of EMAP filters in sub-band f=1kHz

Note that, the cumulative histograms are also given by using Gaussian kernel as in section 3.1. Like in section 3.2, an interpolation based transform is derived in each sub-band log-spectral magnitude domain. The advantage use of CHEQ can be listed as follows: (1) the monotonic form of CDF and cumulative histogram makes the estimation to be unique and smoothed, (2) the cumulative density can be estimated with small number of data. Examples of EMMSE, EMAP, and CHEQ filtering in the log spectral magnitude is shown in Figure 3,4 and 5, respectively. The figures show that, the CHEQ enhancement filters are more smooth than EMMSE and EMAP and it might be the reason why the sound made by this method has less musical noise level.

4. Experiments

The experiment are performed using English and Japanese versions of AURORA2 database. We use AURORA2 English version [8] to learn the enhancement filters, and evaluate the proposed enhancement methods on AURORA2J [9]. The databases have the same noise conditions as the speakers and languages are independent. In the enhancement training, the total noisy speech data are used, ignoring the original SNR index The noisy speech data is clustered into six sub-data according to the SNRGMM index of 20dB, 15dB, 10dB, 5dB, 0db and -2dB. The last index is explained by a bias of SNRGMM estimation under negative true SNR. For each noise condition and given sub-data index, the EMMSE, EMAP and CHEQ enhancement filters are trained using the joint or cumulative density of noisy and clean speech in each sub-band log-spectral magnitude domain. As for the ASR feature extraction, we use a 25ms hamming window length and 10 ms window shift for the STDFT. In the evaluation, the noisy speech is indexed by the SNRGMM estimation and is switched to the nearest filter set in the meaning of SNRGMM. In the simulation mode, only the noise sam-

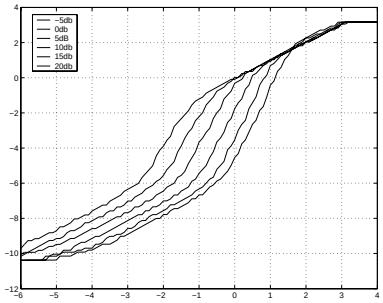

Figure 5: Example of CHEQ filters in sub-band f=1kHz

ple from AURORA2 data is used. We use more sub-classes in terms of 10 SNRGMM indexes to simulate the noisy speech data for training. The enhancement filters is then trained and applied as in the clustering mode. For the CHEQ filtering, one more set of enhancement filters are trained using a smaller number of samples training (CHEQS). For that case, only 50 utterances are used to train the enhancement filter at each SNRGMM index. The enhanced speech data is tested in both the SNR improvement and ASR performances. Speech recognition experiments are performed on the digit recognition task [8]. The digit HMMs are the standard complex back-end models of 16 states and each state has a 20 components Gaussian mixture with diagonal covariance matrix. The training process is carried out at each front-end before training. The experiment shows the significant improvement of proposed systems in both measurements. Among the estimation methods, the CHEQ are performed best at the speech recognition with approximate 62 percents in clean training and 41 percents of relative improvement in multi-conditional training, while the EMMSE is better in the SNR improvement with approximately 8dB of SNR improvement. Unlike in the model based approaches, the empirical MAP performed worse than EMME and CHEQ and it can be explained by the getting errors of the maximizing operation across spiked histograms. In other words, for the data driven approaches, the EMMSE and CHEQ is more preferable to use. The proposed system with CHEQ method overcomes the advance ETSI front-end [10] where the overall results of relative improvements are 59.09 and 36.08 percents for clean and multi-condition trainings, respectively. Note that, the SPLICE-like algorithm [2], is implemented by authors, in the log-spectral magnitude domain and it yields the overall results of 60.36 and 38.56 percents respectively. Note that, the simulation mode performers better than the clustering mode and it can be explained by the use of more set of SNR-dependent training data. One more thing is that the CHEQ produces less musical noise in the enhanced speeches than EMMSE and EMAP and is also superior to the model-based approaches in the sound quality, expecially for the restaurant, subway and exhibition noise conditions.

5. Conclusions

We present a data driven speech enhancement system based on SNR-dependent emprirical statistical estimators. This method uses the estimated SNR and noise level in utterance duration, to normalize and cluster the training data into SNR-dependent sub-data in order to learn the enhancement filters. The empirical statistical estimators are derived using the non-parametric learning of the joint or cumulative density. This method can be considered as a continous codebook-dependent transform versus the conventional discrete codebook- dependent piecewise transform. In future work, we intend to develop a system with an on-line SNRGMM measurement, which is applicable for the real-time processing.

Table 2: *SNR improvement: evaluation on AURORA2J, [dB]*

Learning mode	EMMSE	EMAP	CHEQ	CHEQS
Clustering mode	8.03	6.16	7.27	7.35
Simulation mode	8.43	7.69	8.20	8.15

Table 3: *Relative ASR performance of clean training: evaluation on the AURORA2J*

Learning mode	EMMSE	EMAP	CHEQ	CHEQS
Clustering mode	61.85%	50.44%	62.78%	62.07%
Simulation mode	62.32%	52.33%	62.98%	62.85%

Table 4: *Relative ASR performance of multi-condition training: evaluation on the AURORA2J*

Learning mode	EMMSE	EMAP	CHEQ	CHEQS
Clustering mode	23.79%	20.24%	32.78%	32.12%
Simulation mode	25.65%	28.32%	41.06%	40.17%

6. Acknowledgements

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for CC Society, Course Management System under Ubiquitous Computing Environment, 2004.

7. References

[1] Y.Emphraim "Statistical model based speech enhancement systems", *IEEE Proc.*, vol. 80, pp. 1526-1555, 1992.

[2] A.Acero "Acoustical and Environmental Robustness in Automatic Speech Recognition", *Kluwer Academic Publishers*, 1993.

[3] D.Burshtein, and S.Gannot, "Speech Enhancement Using a Mixture-Maximum Model" *IEEE Trans. ASSP*, Vol. 10, No. 6, pp. 341-351, 2002

[4] T.H.Dat, K.Takeda and F. Itakura "Robust SNR estimation of noisy speech based on Gaussian mixture modeling on log-power domain", *in ISCA ITRW Robust,*2004.

[5] A.Korthauer,"Robust estimation of the SNR of noisy speech for the quality evaluation of speech databases," *in IEEE RMSRAC,*, 1999.

[6] NIST Speech Quality Assurance (SPQA) Package http://www.nist.gov/speech/tools/index.htm.

[7] T. Hastie, R. Tibshinari, J. Friedman " The Elements of Statistical Learning" *Springer*, 2001.

[8] H.Hirsch, D.Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in ISCA ITRW ASR*, 2000.

[9] http://sp.shinshu-u.ac.jp/CENSREC

[10] ESTI standard document,*ETSI ES201 108 v1.1.2 (2000-04)*, 2000.