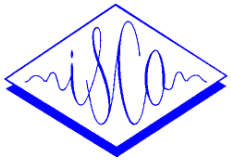

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEECH ENHANCEMENT : NEW APPROACHES TO SOFT DECISION

JOON-HYUK CHANG, NAM SOO KIM

School of Electrical and Computer Engineering,
Seoul National University, Seoul, Korea.
E-mail: {changjh, nkim}@snu.ac.kr

ABSTRACT

In this paper, we propose new approaches to speech enhancement based on soft decision. In order to enhance the statistical reliability in estimating speech activity, we introduce the concept of a global speech absence probability (GSAP). First, we compute the conventional speech absence probability (SAP) and then modify it according to the newly proposed GSAP. Moreover, for improving the performance of the SAP's at voice tails (transition periods from speech to silence), we revise the SAP's using a hang-over scheme based on hidden Markov model (HMM).

In addition, we suggest a robust noise update algorithm in which the noise power is estimated not only in the periods of speech absence but also during speech activity by noise and speech spectrum estimation based on soft decision. Also, for improving the SAP determination and noise update routine we present a new signal to noise ratio (SNR) concept which is called the predicted SNR in this paper. The prediced SNR is defined by the ratio between estimated speech and noise spectrum makes a further improvement the discrete cosine transform (DCT). Results from the test show that the proposed algorithm which is called the speech enhancement based on soft decision (SESD) yields better performance than the conventional methods.

1. INTRODUCTION

The problem of enhancing speech degraded by uncorrelated additive noise, when only the noisy speech is available, has been widely addressed in the past and it still provides an active field of research. Many approaches have been investigated in order to gain spectral enhancement [1]-[3].

In this paper, we introduce the global SAP (GSAP) which is a robust measure for the speech absence determined in each frame globally. Based upon a set of statistical models, we propose a robust way to obtain the GSAP. The LSAP computed in each frequency component is then combined with the GSAP in order to produce robust estimates. The modification is made in such a way that the SAP has the same value of GSAP in the case of speech absence while it is maintained to its original value when the speech is present. Moreover, a hang-over technique based on the hidden Markov model (HMM) is incorporated to achieve good quality particularly in speech tails [4]. The hang-over technique enables the SAP to increase gradually in speech tails and consequently improves the intelligibility of the enhanced speech. According to the newly derived SAP, we modify the spectral gain and update the noise power spectrum in both the speech activity and non-activity periods.

For gaining good performances in estimating both the SAP and the noise power, we propose the predicted SNR. This newly proposed concept also helps to accelerate the performance of the discrete cosine transform (DCT) known to be very effective in energy compaction [5] since we estimate the speech and noise spectrum, respectively, based on soft decision.

Results from the MOS test confirm that our approach yields much better performance than the speech enhancement method adopted in the IS-127 standard used for the North American CDMA digital PCS.

2. SOFT DECISION

We assume that a noise signal *n* is added to a speech signal *x*, with their sum being denoted by *y*. Taking Fourier transform gives us,

$$Y_k(t) = X_k(t) + N_k(t),\ k = 1, 2, \cdots, M \qquad (1)$$

where *k* denotes the *k*th frequency bin, *M* is the total number of frequency components, and *t* is the frame index in the time domain, respectively. The basic assumption adopted in a speech enhancement approach is described by the following two hypotheses:

$$H_0 : \text{speech absent} : \mathbf{Y} = \mathbf{N} \qquad (2)$$

$$H_1 : \text{speech present} : \mathbf{Y} = \mathbf{N} + \mathbf{X} \qquad (3)$$

in which $\mathbf{Y} = [Y_1, Y_2, ..., Y_M]$, $\mathbf{N} = [N_1, N_2, ..., N_M]$, and $\mathbf{X} = [X_1, X_2, ..., X_M]$, respectively.

The purpose of the speech enhancement technique is to estimate $\{X_k(t), k = 1, 2, ..., M\}$ given only $\{Y_k(t), k = 1, 2, ..., M\}$.

2.1. Local Soft Decision

We can compute the LSAP in the *k*th frequency bin conditioned on the current observations by

$$\begin{aligned} P(H_0|Y_k(t)) &= \frac{p(Y_k(t)|H_0)P(H_0)}{p(Y_k(t))} \\ &= \frac{p(Y_k(t)|H_0)P(H_0)}{p(Y_k(t)|H_0)P(H_0)+p(Y_k(t)|H_1)P(H_1)} \\ &= \frac{1}{1+\frac{P(H_1)}{P(H_0)}\Lambda(Y_k(t))} \end{aligned} \tag{4}$$

where $P(H_0)$ represents the *a priori* probability of speech absence and $\Lambda(Y_k(t))$ is the likelihood ratio in the *k*th frequency bin. Based on a complex Gaussian assumption of the clean speech and the noise spectra, the probability density functions (pdf's) conditioned on H_0 and H_1 are assumed to be

$$p(Y_k|H_0) = \frac{1}{\pi\lambda_{n,k}}\exp\{-\frac{|Y_k|^2}{\lambda_{n,k}}\} \tag{5}$$

$$p(Y_k|H_1) = \frac{1}{\pi[\lambda_{n,k}+\lambda_{s,k}]} \tag{6}$$

$$\cdot\exp\{-\frac{|Y_k|^2}{\lambda_{n,k}+\lambda_{s,k}}\} \tag{7}$$

in which $\lambda_{s,k}$ and $\lambda_{n,k}$ denote the variances of the clean speech and noise, respectively. Therefore, $\Lambda(Y_k(t))$ can be given as follows [4]

$$\begin{aligned} \Lambda(Y_k(t)) &= \frac{p(Y_k(t)|H_1)}{p(Y_k(t)|H_0)} \\ &= \frac{1}{1+\xi_k(t)}\exp\left[\frac{\gamma_k(t)\xi_k(t)}{1+\xi_k(t)}\right] \end{aligned} \tag{8}$$

where

$$\xi_k(t) \equiv \frac{\lambda_{s,k}(t)}{\lambda_{n,k}(t)}, \gamma_k(t) \equiv \frac{|Y_k(t)|^2}{\lambda_{n,k}(t)} \tag{9}$$

and $\xi_k(t)$, $\gamma_k(t)$ are called the predicted SNR and the *a posteriori* SNR, respectively. Robust estimation of the LSAP is attributed to the variances of speech and noise which must be obtained appropriately from not only the current observations but also the past observations. Detailed descriptions of the algorithm for estimating $\xi_k(t)$ and $\gamma_k(t)$ are given in the following section.

2.2. Global Soft Decision

GSAP is newly proposed and defined in a way similar to representing the LSAP, and it is computed in a frame globally. Applying Bayes rule, it is easily derived that

$$\begin{aligned} P(H_0|\mathbf{Y}(t)) &= \frac{p(\mathbf{Y}(t)|H_0)P(H_0)}{p(\mathbf{Y}(t))} \\ &= \frac{p(\mathbf{Y}(t)|H_0)P(H_0)}{p(\mathbf{Y}(t)|H_0)P(H_0)+p(\mathbf{Y}(t)|H_1)P(H_1)} \end{aligned} \tag{10}$$

where $\mathbf{Y}(t) = [Y_1(t), Y_2(t), \ldots, Y_M(t)]$. Since the spectral component in each frequency bin is assumed to be statistically independent, Eq. (10) can be converted to

$$P(H_0|\mathbf{Y}(t)) = \frac{1}{1+\frac{P(H_1)}{P(H_0)}\prod_{k=1}^{M}\Lambda(Y_k(t))}. \tag{11}$$

As shown in Eq. (11), the GSAP can be considered to be more statistically reliable than the LSAP for which only the data in the corresponding frequency bin are used.

2.3. Soft Decision including Hang-Over based on HMM

In [4], to prevent clipping of weak speech tails, the decision made on the current frame is modified by the use of an HMM-based hang-over technique. For a similar purpose, we propose to incorporate a hang-over scheme to our soft decision approach to obtain improved quality in speech tails. In the proposed algorithm, a Markov state is constructed for each frequency bin, and then used to take advantage of the all previous decisions to determine the SAP in the current frame. This differs form the hang-over scheme proposed by Sohn et,al [4] in that separate state is assigned to each frequency bin, which is considered more realistic since an indication of speech presence does not tell us that all the spectral components are present.

For each frequency bin, there are two states H_0 and H_1 which are equivalent to the hypotheses given by Eqs. (2) and (3). A sequence of states is governed by a first-order Markov process where the state transition possibilities are parameterized such that

$$a_{ij}^k = P(q_{t,k} = H_j|q_{t-1,k} = H_i), i,j = 0,1 \tag{12}$$

in which $q_{t,k}$ represents the state of the *k*th frequency bin at time *t*. Based on the assumption that the Markov process is time invariant and it reaches to its stationary state, we can obtain

$$P(q_{t,k} = H_i) = P(H_i), i = 0,1 \tag{13}$$

where $P(H_0)$ and $P(H_1)$ can be computed according to the stationarity equation given by

$$a_{01}^k P(H_0) = a_{10}^k P(H_1) \tag{14}$$

under the constraint, $P(H_0)+P(H_1) = 1$. Therefore, the total process can be fully parameterized by a_{01} and a_{10}, and we choose 0.2 and 0.1 respectively for their values.

Since the Markov state model depends on not only the current observation but also the past observations, we should substitute the *a posteriori* probability ratio $\Lambda(Y_k(t)) = P(H_1|Y_k(t))/P(H_0|Y_k(t))$ with $\Gamma(Y_k(t)) = P(q_{t,k} = H_1|\mathcal{Y}_k(t))/P(q_{t,k} = H_0|\mathcal{Y}_k(t))$ where $\mathcal{Y}_k(t) = \{Y_k(t), Y_k(t-1), \ldots, Y_k(1)\}$ is a sequence of observations in the *k*th frequency bin till the current frame *t*. For the calculation of $\Gamma(Y_k(t))$, we define the forward variable $\alpha_k(t,i) = p(q_{t,k} = H_i, \mathcal{Y}_k(t))$ and compute it recursively according to the well known forward procedure given as follows :

$$\alpha_k(t,i) = \begin{cases} P(H_i)p(Y_k(1)|q_{1,k} = H_i), \text{for } t{=}1 \\ \left(\alpha_k(t-1,0)a_{0j} + \alpha_k(t-1,1)a_{1j}\right) \\ \qquad \cdot p(Y_k(t)|q_{t,k} = H_i), \text{for } t \geq 2. \end{cases} \tag{15}$$

By using this result, we obtain the recursive formula for $\Gamma(Y_k(t))$ as follows:

$$\Gamma(Y_k(t)) = \frac{\alpha_k(t,1)}{\alpha_k(t,0)} = \frac{a_{01}^k + a_{11}^k \Gamma(Y_k(t-1))}{a_{00}^k + a_{10}^k \Gamma(Y_k(t-1))} \Lambda(Y_k(t)). \tag{16}$$

The revision of the SAP's using the proposed hang-over scheme is completed by replacing $\Lambda(Y_k(t))$ in Eq. (4) and Eq. (11) with $\Gamma(Y_k(t))$.

2.4. Soft Decision

As for the GSAP, we can have more robust estimates. By using the GSAP, a fixed gain may be equally applied to all frequency components during speech absence and we can obtain the natural characteristics of noise by reducing gain variances. For the purpose of gaining such a selective spectral suppression, we propose a way to modify the LSAP's such that

$$\begin{aligned} \widetilde{P}(H_0|Y_k(t)) &= P(H_0|\mathbf{Y}(t)) \cdot P(H_0|\mathbf{Y}(t)) \\ &\quad + P(H_1|\mathbf{Y}(t)) \cdot P(H_0|Y_k(t)) \end{aligned} \tag{17}$$

where $\widetilde{P}(H_0|Y_k(t))$ denotes the modified LSAP. Considering Eq. (17), it is not difficult to find out the modified LSAP replaces the LSAP in the case of speech presence and the GSAP during the speech absence.

Furthermore, in the case of transition periods from speech to silence $(0 < P(H_0|\mathbf{Y}(t)) < 1)$, $\widetilde{P}(H_0|Y_k(t))$ represents the LSAP compensated by the GSAP which has statistical reliability.

3. NOISE AND SPEECH POWER SPECTRUM ESTIMATION

We use the long-term smoothed power spectra of the background noise and clean speech as the estimates for $\{\lambda_{n,k}(t)\}$ and $\{\lambda_{s,k}(t)\}$, respectively.

$$\begin{aligned} \hat{\lambda}_{n,k}(t+1) &= \\ &\quad \zeta_n \hat{\lambda}_{n,k}(t) + (1-\zeta_n) E[|N_k(t)|^2|Y_k(t)] \\ \hat{\lambda}_{s,k}(t+1) &= \\ &\quad \zeta_s \hat{\lambda}_{s,k}(t) + (1-\zeta_s) E[|X_k(t)|^2|Y_k(t)] \end{aligned} \tag{18}$$

where $\hat{\lambda}_{n,k}(t)$, $\hat{\lambda}_{s,k}(t)$ are the estimates for $\lambda_{n,k}(t)$, $\lambda_{s,k}(t)$ and ζ_n and ζ_s are the parameters for smoothing under a general stationarity assumption of $\mathrm{N}(t)$ and $\mathrm{X}(t)$. Taking into account the uncertainty for speech absence or presence, the expectation of the power spectra for the speech and noise can be shown to be

$$\begin{aligned} &E[|N_k(t)|^2|Y_k(t)] \\ &= E[|N_k(t)|^2|Y_k(t), H_0]\widetilde{P}(H_0|Y_k(t)) \\ &+ E[|N_k(t)|^2|Y_k(t), H_1]\widetilde{P}(H_1|Y_k(t)) \end{aligned} \tag{19}$$

$$\begin{aligned} &E[|X_k(t)|^2|Y_k(t)] \\ &= E[|X_k(t)|^2|Y(t), H_0]\widetilde{P}(H_0|Y_k(t)) \\ &+ E[|X_k(t)|^2|Y_k(t), H_1]\widetilde{P}(H_1|Y_k(t)) \end{aligned} \tag{20}$$

where

$$\begin{aligned} &E[|N_k(t)|^2|Y(t), H_0] = |Y_k(t)|^2 \\ &E[|N_k(t)|^2|Y(t), H_1] = \\ &\qquad \left(\frac{\hat{\xi}_k(t)}{1+\hat{\xi}_k(t)}\right)\hat{\lambda}_k(t) + \left(\frac{1}{1+\hat{\xi}_k(t)}\right)^2 |Y_k(t)|^2 \\ &E[|X_k(t)|^2|Y(t), H_0] = 0 \\ &E[|X_k(t)|^2|Y(t), H_1] = \\ &\qquad \left(\frac{1}{1+\hat{\xi}_k(t)}\right)\hat{\lambda}_k(t) + \left(\frac{\hat{\xi}_k(t)}{1+\hat{\xi}_k(t)}\right)^2 |Y_k(t)|^2 \end{aligned} \tag{21}$$

From Eq. (19), it is noted that the noise power spectrum estimate is updated not only during the periods of speech absence but also when it is present.

4. PREDICTED SNR

In this section, we present a description of the predicted SNR which is a new concept for the description of speech and noise ratio. In Eq. (18), we estimated the noise and speech spectrum, respectively, by considering the uncertainty of speech absence. By these estimated values, we can compute the predicted SNR for the next frame as

$$\hat{\xi}_k(t+1) = \frac{\hat{\lambda}_{s,k}(t+1)}{\hat{\lambda}_{n,k}(t+1)} \tag{22}$$

and apply it when we compute the SAP's and update the noise and speech power spectrum for the next frame.

Table I
MOS results for the proposed enhancement algorithm (SESD)
and IS-127 enhancement technique

noise		babble			pink			white		
SNR(dB)		5	10	15	5	10	15	5	10	15
none		2.00	2.14	2.29	1.13	1.68	1.92	1.19	1.49	2.11
IS-127		2.60	3.09	3.54	2.34	2.89	3.51	2.31	3.04	3.38
global soft decision		2.85	3.43	3.61	2.51	3.13	3.63	2.44	3.11	3.60
SESD	·	2.95	3.45	3.70	2.51	3.16	3.76	2.45	3.17	3.60
	hang-over	2.99	3.45	3.79	2.65	3.20	3.78	2.47	3.17	3.61
	DCT	2.88	3.43	3.78	2.91	3.56	4.02	2.87	3.61	4.02
	DCT, hang-over	2.96	3.52	3.79	2.82	3.48	3.98	2.95	3.68	4.04

4.1. Performance Acceleration of the DCT

In [5], the advantages of the DCT are illustrated for speech enhancement. The fact is mainly due to the energy compaction property of the DCT which makes the speech energy particularly in voiced sounds be concentrated in only few coefficients of the spectrum.

We now present how to achieve the performance acceleration of the DCT by the proposed predicted SNR. In those DCT embedded environments, we can obtain emphasized LSAP values in low frequency bands in comparison with the Fourier transform case. The emphasized LSAP mechanism comes from not only the properties of the DCT but also the predicted SNR which is determined by Eqs. (19) and (20) that the noise and speech spectrum are estimated separately based on soft decision. Consequently, we can update the speech power spectrum in low frequency bands and the noise power spectrum in high frequency bands without loss of the speech energy. This accurate estimation of the noise spectrum helps to reduce the musical noise.

5. RESULT AND CONCLUSION

To evaluate the performance of the proposed speech enhancement for the human perception, we conducted mean opinion score (MOS) tests on a number of enhanced noisy speech samples. In Table 1, the MOS test results are listed in various noise conditions. The proposed approach shows better MOS results than the IS-127 standard for every cases. In summary, we proposed a new speech enhancement algorithm based on soft decision. The performance of the proposed approach is found superior to the enhancement technique of the IS-127 standard through a number of MOS tests.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 32, No. 6, pp. 1109-1121, Dec. 1984.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 28, pp. 137-145, Apr. 1980.

[3] D. Malah, R. Cox and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, AZ, Mar. 1999.

[4] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1-3, Jan. 1999.

[5] I. Y. Soon, S. N. Koh and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech communication*, Vol. 24, No. 3, pp. 249-257, 1998.