



Voice Activity Detection Using Frame-Wise Model Re-Estimation Method Based on Gaussian Pruning with Weight Normalization

Masakiyo Fujimoto, Shinji Watanabe, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan

{masakiyo, watanabe, nak}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection (VAD). Our previous work, switching Kalman filter-based VAD, sequentially estimates a non-stationary noise Gaussian mixture model (GMM) and constructs GMMs of observed noisy speech signals by composing pre-trained silence and clean GMMs and sequentially estimated noise GMMs. However, the composed models are not optimal, because they do not fully reflect the characteristics of the observed signal. Thus, to ensure the optimality of the composed models, we investigate a method for re-estimating the composed model. Since our VAD method works under the frame-wise sequential processing, there are insufficient re-training data for re-estimation of whole model parameters. To solve this problem, we propose a model re-estimation method that involves the extraction of reliable information using Gaussian pruning with weight normalization. Namely, the proposed method re-estimates the model by pruning non-dominant Gaussian distributions in expressing the local characteristics of each frame and by normalizing Gaussian weights of remaining distributions.

Index Terms: voice activity detection, switching Kalman filter, Gaussian pruning, Gaussian weight normalization

1. Introduction

Voice activity detection (VAD) that automatically detects a period containing a target speech signal from a continuously observed signal plays crucial role as regards progress on speech processing technology. VAD has various applicable fields in speech-oriented technology, e.g., speech enhancement, speech coding, and the front-end processing of speech recognition.

VAD usually consists of two parts: a feature extraction part and a discrimination part. Higher-order statistics [1], long-term spectral divergence [2], and a periodic to aperiodic component ratio of speech [3] have been proposed as robust discriminative features. On the other hand, statistical model-based methods have been proposed as non-speech / speech discriminators [4, 5]. To additionally improve robustness for non-stationary noise environments from the conventional model-based approach, we have proposed a switching Kalman filter (SKF)-based VAD [6].

SKF-based VAD consists of the following three steps:

- (1) *Noise estimation step:* the parameters of a noise Gaussian mixture model (GMM) are estimated based on SKF by using the mean and the co-variance parameters of silence and clean speech GMMs, which are estimated in advance by using clean speech corpora.
- (2) *Composition step:* two internal states of non-speech (silence + noise) and speech (clean speech + noise) in the noisy speech model are constructed by composing silence and clean speech GMMs with the estimated noise GMM as shown in Fig. 1.

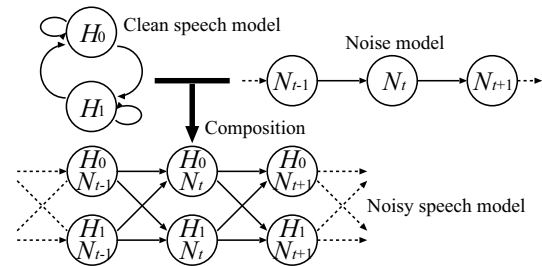


Figure 1: Non-speech / speech state transition model with noise dynamics. The symbols H_0 , H_1 , and N_t denote a silence state, a clean speech state, and a noise state sequence, respectively.

- (3) *Discrimination step:* speech absence or activity in the observed signal is discriminated by using the likelihood ratio of the observed signal between non-speech and speech GMMs.

In this framework, the characteristics of the observed signal are not directly reflected in the composition step, unlike the noise estimation. Namely, the composition step lacks optimality in terms of the consideration of the observed signal. Thus, one of the most important factors for our proposed approach is how to ensure the optimality of the noisy speech model by incorporating the characteristics of the observed signal. One of the standard approaches for this purpose is to employ a parameter re-estimation method in the noisy speech model by using the observed signal at the composition step. However, since VADs strongly require the smallest latency in many applications, a re-estimation method with one frame sample is desired. In this situation, it is almost impossible to re-estimate whole Gaussian parameters. To satisfy the requirement, we propose a model re-estimation method based on a Gaussian pruning technique by selecting dominant Gaussian distributions of non-speech and speech GMMs depending on the observed signal of the current frame. This method is reasonable, because in sequential processing, only partial aspects of whole speech characteristics are assumed to appear in each frame. Thus, contributions made by a small number of Gaussian distributions would be more dominant than the others in expressing the local characteristics of the observed signal in each frame. On the other hand, non-dominant Gaussian distributions may be unimportant and even harmful. Under these assumptions, we investigate a way of improving the optimality of the GMM by using Gaussian pruning-based model re-estimation. Then, we also apply Gaussian weight normalization to the remaining dominant Gaussian distributions in each frame. This normalization method can enhance likelihood given by each remaining Gaussian distribution, and improves effectiveness of the proposed Gaussian pruning.

The Gaussian pruning techniques are usually used to reduce computational complexity without degrading speech recogni-

tion accuracy [7, 8, 9]. Most Gaussian pruning techniques reduce the number of Gaussian distributions by merging similar distributions using some clustering methods. These methods involve the prior processing of speech recognition, and usually characteristics of the observed signal are not reflected in the pruning result. On the other hand, our proposed method involves posterior processing of VAD, and improves the optimality of GMM by using characteristics of the observed signal. The proposed method does not aim to reduce computational complexity, thus making it different from conventional pruning techniques.

2. Review of SKF-based VAD

This section briefly reviews each step of SKF-based VAD.

2.1. Noise estimation step

The mean and co-variance parameters of noise GMM are sequentially estimated and updated at each frame with a SKF in the log Mel-spectral domain. In this step, parts of parameter sets of SKF are obtained by the pre-trained silence and clean speech GMMs. Furthermore, the observed signal \mathbf{O}_t which is L -dimensional vector of the log Mel-spectra at current t -th frame is utilized in the SKF, thus the SKF can reflect the characteristics of observed signal in the estimation results. The details of noise estimation step are described in [6].

2.2. Composition step

In the log Mel-spectral domain, the composition of noisy speech model is derived by the following equations:

$$\boldsymbol{\mu}_{O,t,j,k} = \boldsymbol{\mu}_{S,j,k} + \log(1 + \exp(\boldsymbol{\mu}_{N,t,j,k} - \boldsymbol{\mu}_{S,j,k})) \quad (1)$$

$$\boldsymbol{\Sigma}_{O,t,j,k} = \mathbf{F}_{t,j,k} \boldsymbol{\Sigma}_{N,t,j,k} \mathbf{F}_{t,j,k}^T + \boldsymbol{\Sigma}_{S,j,k} \quad (2)$$

$$\mathbf{F}_{t,j,k} = \text{diag}\{\partial \boldsymbol{\mu}_{O,t,j,k} / \partial \boldsymbol{\mu}_{N,t,j,k}\}, \quad (3)$$

where $\boldsymbol{\mu}_{S,j,k}$, $\boldsymbol{\Sigma}_{S,j,k}$, $\boldsymbol{\mu}_{N,t,j,k}$, $\boldsymbol{\Sigma}_{N,t,j,k}$, $\boldsymbol{\mu}_{O,t,j,k}$, and $\boldsymbol{\Sigma}_{O,t,j,k}$ denote the mean vectors and the diagonal co-variance matrices of the k -th Gaussian distribution in a silence ($j = 0$) or clean speech ($j = 1$) GMM, a noise GMM, and a noisy speech GMM, respectively. The operations $\log(\cdot)$ and $\exp(\cdot)$ are independently applied to each vector element, and the vector $\mathbf{1} = \{1, \dots, 1\}^T$.

This step merely composes the silence and the clean speech GMMs with the estimated noise GMM, thus the characteristics of observed signal are not fully reflected to the composed noisy speech model.

2.3. Discrimination step

When the noisy speech model is given, speech absence or activity is discriminated by using the likelihood ratio test (LRT) as follows:

$$q_t = \begin{cases} \text{Speech absence} & R_t < \text{Threshold} \\ \text{Speech activity} & \text{otherwise} \end{cases} \quad (4)$$

where q_t and R_t denote the estimated state and likelihood ratio at the t -th frame, respectively.

In the LRT, R_t is given as ratio of forward probability $\alpha_{j,t}$, i.e., $R_t = \alpha_{1,t} / \alpha_{0,t}$ instead of ratio of a model likelihood $b_j(\mathbf{O}_t)$. This method is called as the HMM-based hang-over scheme [4]. The forward probability $\alpha_{j,t}$ is derived as

$$\alpha_{j,t} = \sum_{i=0}^{1} a_{i,j} \cdot \alpha_{i,t-1} \cdot b_j(\mathbf{O}_t) \quad (5)$$

$$b_j(\mathbf{O}_t) = \sum_{k=1}^K w_{j,k} \cdot \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,k}, \boldsymbol{\Sigma}_{O,t,j,k}), \quad (6)$$

where $a_{i,j}$ and $w_{j,k}$ denote the state transition probability of state i (preceding state) to j (present state) and Gaussian weight of a silence or clean speech GMM, respectively.

3. Model re-estimation based on Gaussian pruning with weight normalization

3.1. Problem of frame-wise model re-estimation

Noisy speech GMMs consist of the parameters derived by (1) and (2). In this method, the parameters of the noise GMM are optimally estimated by SKF, because the estimation scheme of the Kalman filter ensures the optimality of the estimation result by reflecting the characteristics the observed signal. However, the parts of parameter sets of SKF which are provided by the silence and clean speech GMMs are not optimal for the observed signal, because the parameters are trained by using speech corpora, which are usually different from the observed signal. Therefore, the parameters derived by (1) and (2) are not fully optimal for the observed signal.

The optimization of noisy speech model is an important factor for SKF-based VAD. Thus, to ensure the optimality of noisy speech model, the model parameters should be re-estimated by reflecting the characteristics of the observed signal. In many cases, the mean and the co-variance parameters of a Gaussian distribution are updated with a maximum likelihood or a maximum *a posteriori* estimator. Here, since VADs strongly require the smallest latency, a re-estimation method with one frame sample is desired. However, it is almost impossible to re-estimate whole Gaussian parameters with one frame sample. To satisfy the requirement, we propose a model re-estimation that is different from the parameter re-estimation method with re-training data. The proposed model re-estimation method consists of two techniques. One is a Gaussian pruning technique by selecting dominant Gaussian distributions of non-speech and speech GMMs depending on the observed signal of the current frame. The other is Gaussian weight normalization technique which enhances the likelihood given by remaining dominant Gaussian distribution after the Gaussian pruning.

3.2. The aim of Gaussian pruning

Each noisy speech GMM includes K Gaussian distributions. Each Gaussian distribution contained in each GMM represents various noisy speech signal characteristics. Furthermore, the observed signal also has various characteristics. For the effective likelihood calculation, usually, various Gaussian distributions are needed to cope with various characteristics included in all (or several) frame sequences of the observed signal. For example, Fig. 2(a) shows the correspondence of the characteristics of several frames to Gaussian distributions. As shown in Fig. 2(a), the GMM of the figure consists of four Gaussian distributions, and we can see that all the Gaussian distributions are needed to represent the characteristics of several frames. Thus, we must utilize various Gaussian distributions that can satisfactorily represent all the characteristics of the observed signal for effective likelihood calculation.

However, the above approach can be applied only in the case of a likelihood calculation with the whole frame sequence of the observed signal. If the likelihood calculation is restricted to a local (specific) characteristic of a current frame, Gaussian distributions that are not dominant in expressing the local characteristic may be unimportant and even harmful. For example, Fig. 2(b) shows the correspondence of the characteristics of the current frame to Gaussian distributions. As shown in Fig. 2(b), the likelihood calculation may be performed satisfactorily by using only two Gaussian distributions, $k = 2$ and $k = 3$, which

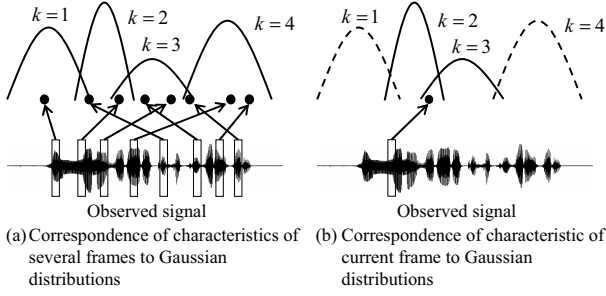


Figure 2: Example of correspondence of frame characteristics to Gaussian distributions

are dominant distributions for expressing the local characteristic of the current frame. Other Gaussian distributions, $k = 1$ and $k = 4$, are not dominant distributions for expressing the local characteristic, thus these Gaussian distributions may be unimportant for the likelihood calculation. Therefore, since only parts of the entire speech characteristics appear in each frame in the sequential processing, the contributions of a small number of Gaussian distributions are more dominant than the others. Under these assumptions, we investigate the effective likelihood calculation by pruning non-dominant Gaussian distributions.

3.3. Gaussian weight normalization

The Gaussian weights of the remaining distributions are normalized as the sum of the Gaussian weights is equal to 1 in accordance with the following equation:

$$\hat{w}_{t,j,n} = \frac{w_{j,n}}{\sum_{n'} w_{j,n'}}, \quad (7)$$

where n and $\hat{w}_{t,j,n}$ denote the Gaussian index and the normalized Gaussian weight of the remaining distributions at the t -th current frame, respectively.

This normalization is an important factor in Gaussian pruning, because it enhances the likelihood of the remaining distributions. For example, when the Gaussian weights of before the pruning are given as $w_{j,k} = \{0.25, 0.35, 0.15, 0.25\}$ and the two Gaussian distributions, $k = 2$ and $k = 3$, remain after the pruning, the Gaussian weights after the pruning are normalized as $\hat{w}_{t,j,2} = 0.7$ and $\hat{w}_{t,j,3} = 0.3$. By using these values, the likelihoods of Gaussian distributions $k = 2$ and $k = 3$ can be effectively enhanced.

3.4. Gaussian pruning based on N -best approach

As the Gaussian pruning criterion, we employ the posterior probability conditioned on the observed signal derived by the following equation:

$$P_{t,j,k} = \frac{w_{j,k} \cdot \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,k}, \boldsymbol{\Sigma}_{O,t,j,k})}{\sum_{k'=1}^K w_{j,k'} \cdot \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,k'}, \boldsymbol{\Sigma}_{O,t,j,k'})} \quad (8)$$

By sorting $P_{t,j,k}$ in descending order, the sorted posterior probability $P_{t,j,k}^{Sort}$ is obtained. In the results of the descending order sorting, $P_{t,j,k}^{Sort}$ with a small k has a high value and $P_{t,j,k}^{Sort}$ with a large k has a low value. Thus, $N_{t,j}$ dominant Gaussian distributions can be obtained by selecting only the Gaussian distributions from the highest $P_{t,j,k}^{Sort}$. The number of selected Gaussian distributions $N_{t,j}$ is decided by the value that maximizes the likelihood of the pruned model $\hat{b}_j(\mathbf{O}_t)$ with normalized Gaussian weight $\hat{w}_{t,j,n}$ as follows:

$$N_{t,j} = \arg \max_{N_j} \left\{ \hat{b}_j(\mathbf{O}_t) \right\} \quad (9)$$

$$\hat{b}_j(\mathbf{O}_t) = \sum_n^{N_j} \hat{w}_{t,j,n} \cdot \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,n}, \boldsymbol{\Sigma}_{O,t,j,n}) \quad (10)$$

Here, a different Gaussian distribution is selected in each frame, because the value of $P_{t,j,k}$ changes depending on the characteristics of \mathbf{O}_t .

When the number of selected Gaussian distributions $N_{t,j}$ is decided, the likelihood of the pruned model $\hat{b}_j(\mathbf{O}_t)$ is used for the LRT as shown in Sec. 2.3.

4. Experiments

4.1. Experimental setup

The proposed method was evaluated by using the real recorded data of CENSREC-1-C [10]. The data were recorded in two real environments (a restaurant and a street) with two different sound pressure levels (avg. 60 dBA: high SNR and avg. 70 dBA: low SNR). The sampling rate was 8 kHz. There were ten speakers (five males and five females). The recorded speech consisted of four files per subject. A single file included 9–10 utterances of continuous 1–7 digit numbers with two-second intervals. The correct segment labels were tagged manually.

The feature parameters for the SKF-based VAD were 12th order log Mel-spectra that were extracted by using a Hamming window with a 20 msec frame length and a 10 msec frame shift length. We trained the silence and clean speech GMMs for SKF-based VAD by using the clean training data of CENSREC-1 [11]. Each GMM had 32 Gaussian distributions.

The VAD performance was evaluated on two types of criteria, i.e., the evaluations of utterance-level and the frame-level.

4.2. Experimental results of utterance-level evaluation

The utterance-level evaluation criteria are the utterance correct rate (Corr) and the utterance accuracy rate (Acc) as shown by

$$\text{Corr} = N_c / N_u \times 100 [\%] \quad (11)$$

$$\text{Acc} = (N_c - N_f) / N_u \times 100 [\%], \quad (12)$$

where N_u , N_c , and N_f denote the total number of speech utterances, the number of correctly detected utterances, and the number of incorrectly detected utterances, respectively. The threshold is set at the values that give the best Corr and Acc.

Table 1 compares the proposed and conventional methods. In the table, “Baseline,” “Sohn,” “AFE,” “G.729B,” “SKF,” and “Proposed” represent results obtained with the baseline VAD technique of CENSREC-1-C (energy-based VAD with adaptive threshold) [10], the statistical model-based VAD method proposed by Sohn [4], ETSI ES 202 050 (advanced front-end) [12], ITU-T G.729 Annex B [13], original SKF-based VAD [6], and the proposed Gaussian pruning with weight normalization, respectively. As seen in the table, the proposed method outperforms “SKF” and the other conventional methods. In particular, the results in the presence of restaurant noise show a significant improvement. Since a huge number of incorrectly detected utterances are inserted, the Accs of “AFE” and “G.729B” are very low. The other conventional methods also include serious detection errors. Although the conventional method includes serious detection errors, the proposed method perform well as regards Acc. This means that the proposed method not only detects the speech activity period correctly but also distinguishes between the characteristics of non-speech and speech, correctly.

Table 2 shows the average number of remaining Gaussian distributions after the Gaussian pruning. As seen in the table, the number of remaining Gaussian distributions is very small. These results prove that the effective likelihood calculation is

Table 1: VAD results obtained by utterance-level evaluation

Criterion	Corr (%)					Avg.
	Restaurant		Street			
	High	Low	High	Low		
Baseline	74.20	56.52	39.42	41.45	52.90	
Sohn	72.75	57.10	97.39	78.55	76.45	
AFE	43.77	46.67	79.13	71.88	60.36	
G.729B	51.88	46.67	43.48	42.61	46.16	
SKF	89.86	56.23	100.00	98.84	86.23	
Proposed	92.75	65.51	100.00	100.00	89.57	

Criterion	Acc (%)					Avg.
	Restaurant		Street			
	High	Low	High	Low		
Baseline	21.45	-43.48	-15.65	-33.91	-17.90	
Sohn	45.51	-6.38	94.49	57.39	47.75	
AFE	-73.62	-94.20	-245.51	-166.96	-145.07	
G.729B	-204.35	-199.42	-72.17	-117.39	-148.33	
SKF	69.57	6.96	98.26	95.07	67.47	
Proposed	73.62	15.07	98.84	95.94	70.87	

Table 2: The average number of remaining Gaussian distributions after Gaussian pruning. The values in parentheses are the standard deviations of the number of remaining distributions.

Noise	Restaurant		Street		Avg.
	High	Low	High	Low	
# of Gaussians of non-speech model	1.34 (1.08)	1.89 (1.18)	1.53 (1.03)	2.14 (1.03)	1.73 (1.08)
# of Gaussians of speech model	1.03 (0.30)	2.13 (0.52)	1.30 (0.20)	2.14 (0.23)	1.62 (0.34)

performed satisfactorily by a few dominant Gaussian distributions in terms of expressing the local characteristics of the observed signal.

4.3. Experimental results of frame-level evaluation

The frame-level evaluation criteria are the false rejection rate (FRR) and the false acceptance rate (FAR) as shown by

$$FRR = N_{FR}/N_s \times 100 [\%] \quad (13)$$

$$FAR = N_{FA}/N_{ns} \times 100 [\%], \quad (14)$$

where N_s , N_{ns} , N_{FR} , and N_{FA} are the total number of speech frames, the total number of non-speech frames, the number of speech frames detected as non-speech frames, and the number of non-speech frames detected as speech frames, respectively. FAR and FRR are controlled by a threshold, and have a trade-off relationship.

Table 3 shows results obtained with a frame-level evaluation. In the table, all the results with the exception of “AFE” and “G.729B” are obtained by adjusting the threshold to make the FRR and FAR approximately equal. These results are called the equal error rate. Since the parameters of “AFE” and “G.729B” are fixed, the results obtained with these methods cannot be adjusted to the equal error rate. As seen in the table, the proposed Gaussian pruning method outperforms “SKF” and the other conventional methods as well as the utterance-level evaluation described in Sec. 4.2.

5. Conclusion

This paper presented a technique of Gaussian pruning with weight normalization for statistical model-based VAD. The proposed method is designed to achieve an effective likelihood calculation that reflects the local characteristics of the observed signal. The evaluation results show that our proposed method

Table 3: VAD results obtained by frame-level evaluation

Criterion	FRR (%)					Avg.
	Restaurant		Street			
	High	Low	High	Low		
Baseline	16.90	28.00	34.40	37.10	29.10	
Sohn	29.70	38.10	31.00	30.40	32.30	
AFE	7.30	7.80	2.50	7.00	6.15	
G.729B	15.90	30.90	23.40	38.30	27.13	
SKF	14.10	25.30	5.00	6.70	12.78	
Proposed	11.10	22.20	4.50	5.80	10.90	

Criterion	FAR (%)					Avg.
	Restaurant		Street			
	High	Low	High	Low		
Baseline	18.40	25.00	33.00	35.20	27.90	
Sohn	28.90	40.20	31.70	31.20	33.00	
AFE	75.90	75.40	46.90	29.50	56.93	
G.729B	47.10	41.00	44.80	24.90	39.45	
SKF	14.00	24.10	12.00	9.00	14.78	
Proposed	9.30	24.70	4.60	7.00	11.40	

significantly improves VAD accuracy compared with our previous work and the conventional methods. In the future, we will investigate the integration of VAD and other speech processing techniques.

6. References

- [1] K. Li, M. N. S. Swamy, and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” *IEEE Trans. on SAP*, vol. 13, no. 5, pp. 965–974, Sept. 2005.
- [2] J. Ramírez, J. C. Segura, C. Benítez, A. d. I. Torre, and A. Rubio, “Efficient voice activity detection algorithm using long-term speech information,” *Speech Communication*, vol. 42, no. 3–4, pp. 271–287, Apr. 2004.
- [3] K. Ishizuka and T. Nakatani, “Study of noise robust voice activity detection based on periodic component to aperiodic component ratio,” in *Proc. of SAPA '06*, pp. 65–70, Sept. 2006.
- [4] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE SP Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [5] R. J. Weiss and T. Kristjansson, “DySANA: Dynamic speech and noise adaptation for voice activity detection,” in *Proc. of Interspeech '08*, pp. 127–130, Sept. 2008.
- [6] M. Fujimoto and K. Ishizuka, “Noise robust voice activity detection based on switching Kalman filter,” in *Proc. of Interspeech '07*, pp. 2933–2936, Aug. 2007.
- [7] V. Fischer and T. Rob, “Reduced Gaussian mixture models in a large vocabulary continuous speech recognizer,” in *Proc. of Eurospeech '99*, vol. 3, pp. 1099–1102, Sept. 1999.
- [8] K. Shinoda and K. Iso, “Efficient reduction of Gaussian components using MDL criterion for HMM-based speech recognition,” in *Proc. of ICASSP '02*, vol. 1, pp. 869–872, May 2002.
- [9] A. Ogawa and S. Takahashi, “Weighted distance measures for efficient reduction of Gaussian mixture components in HMM-based acoustic model,” in *Proc. of ICASSP '08*, pp. 4173–4776, Apr. 2008.
- [10] CENSREC-1-C Web site, <http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/>
- [11] CENSREC-1 Web site, <http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/AURORA-2J/>
- [12] ETSI ES 202 050 v.1.1.4, “Speech processing, Transmission and Quality aspects (STQ), Advanced Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithms”, Nov. 2006.
- [13] ITU-T Recommendation G.729 Annex B., “A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70”, Nov. 1996.