Voice activity detection based on conditional random fields using multiple features

Akira Saito, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology

Abstract

This paper proposes a Voice Activity Detection (VAD) algorithm based on Conditional Random Fields (CRF) using multiple features. VAD is a technique used to distinguish between speech and non-speech in noisy environments and is an important component in many real-world speech applications. The posterior probability of output labels in the proposed method is directly modeled by the weighted sum of the feature functions. Effective features are automatically selected by estimating appropriate weight parameters to improve the accuracy of VAD. Experimental results on the CENSREC-1-C database revealed that the proposed approach can decrease error rates by using CRF.

Index Terms: voice activity detection, conditional random fields

1. Introduction

Various systems with speech interface have recently been developed as real world applications of speech technologies. One of the most serious problems in these systems is how to obtain high recognition rates in noisy environments. Noise-robust voice activity detection (VAD), which is the task of separating conversational speech and non-speech, is one of the most important components of ASR. For that reason, various types of VAD algorithms have been proposed. In statistical VAD approaches [1], Gaussian Mixture Model (GMM)-based VAD [2, 3] is a popular technique. This method constructs two GMMs (i.e., speech and non-speech GMMs) from the training data, and voice activity is detected by calculating the probability of test data for each GMM. However, we need to select the feature vectors for the GMMs.

There are numerous features that are useful for VAD (e.g., amplitude, zero-crossing, MFCC and etc.) [4], and some of these may be complementary and their degree of usefulness may be different. Therefore, it is important how these various features are combined and how useful information is extracted from them. VAD is essentially a binary classification problem. Therefore, discriminative machine learning techniques are effective (e.g., Support Vector Machine (SVM)) for solving it. Although generative models (e.g., GMM) represent the data variations of all feature dimensions in probability space, discriminative approaches can select effective features automatically to correctly classify training data. We proposed a VAD system based on conditional random fields (CRF) [5]. CRF can accommodate many statically correlated features of inputs, and they are trained discriminatively. CRF also has an advantage of flexibility to include a wide variety of arbitrary, non-independent features of inputs. We can use multiple features as united frames in the proposed method, and model the temporal features of speech/non-speech labels.

The following section describes a VAD algorithm based on CRF. Section 3 describes the features for VAD. The experimental results are presented in Section 4. The conclusion is given and future work is described in the final section.

2. Voice activity detection based on Conditional random fields

2.1. Conditional random fields

CRF involves a probabilistic framework that is widely used for labeling and segmenting sequential data and it also performs well in natural language processing. CRF is learned to maximize conditional probability $P(\boldsymbol{y} \mid \boldsymbol{x})$ written as :

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(x_t, y_t)\right] \quad (1)$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(x_t, y_t)\right] \quad (2)$$

where $Z(\boldsymbol{x})$ is a normalization factor over all candidate paths, λ_k is a weight parameter for each feature function to be estimated from the training data, f_k is a feature function, and K is the number of feature functions. A feature function f_k represents the relation between inputs $\boldsymbol{x}$ and outputs $\boldsymbol{y}$, and is defined as :

$$f_k(x, y) = \begin{cases} 1 & ((x, y) \text{ satisfies a particular condition}) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

Weights are typically estimated by maximizing the conditional log-likelihood of training data. The learned weight λ_k for each feature function f_k should intuitively be positive for features that are correlated with the target label, negative for features that are anti-correlated with the label, and near zero for relatively uninformative features. CRF directly models the conditional distribution $P(\boldsymbol{y} \mid \boldsymbol{x})$ which is used in final objectives (e.g., recognition and detection), even though many statistical approaches are based on a generative model $P(\boldsymbol{x} \mid \boldsymbol{y})$ of observation.

2.2. Proposed method for VAD

The CRF in the proposed method defines the conditional probability of a state sequence $\boldsymbol{y}$ given an input sequence $\boldsymbol{x}$ that is :

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left[\sum_{t=1}^{T}\left\{\sum_{k=1}^{K} \lambda_k^{(a)} f_k^{(a)}(y_{t-1}, y_t) + \sum_{l=1}^{L}\sum_{d=1}^{D} \lambda_{l,d}^{(b)} f_{l,d}^{(b)}(x_{t,d}, y_t)\right\}\right] \quad (4)$$

where $y_t \in \{0, 1\}$ is a speech/non-speech label and $\boldsymbol{x}_t = [x_t^{(1)}, \ldots, x_t^{(D)}]$ is an observed feature vector at time t of dimension D. Two kinds of feature functions are defined for VAD. The first is a transition feature function $f_k^{(a)}(y_{t-1}, y_t)$ which represents the correlation between two successive speech/non-speech labels. This transition feature function is defined as :

$$f_k^{(a)}(y_{t-1}, y_t) = \begin{cases} 1 & ((y_{t-1}, y_t) = (y_k, y_k')) \\ 0 & (\text{otherwise}) \end{cases} \quad (5)$$

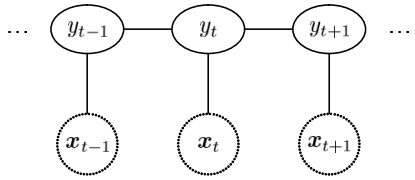

Figure 1: *Graphical model representation of CRF for VAD*

where y_k and y'_k are specific labels that are dependent on k, and weight $\lambda_k^{(a)}$ corresponds to a state transition probability of a hidden Markov Model (HMM). The other function is an observation feature function $f_{l,d}^{(b)}(x_{t,d}, y_t)$, which is defined as :

$$f_{l,d}^{(b)}(x_{t,d}, y_t) = \begin{cases} x_{t,d} & (y_t = y'_l) \\ 0 & (\text{otherwise}) \end{cases} \tag{6}$$

where y'_l is a label that depends on l. Although equation (6) is defined as the first-order function of features, an arbitrary function can be used. The proposed CRF has a similar structure of an HMM. The weight parameters, $\lambda_k^{(a)}$ and $\lambda_{l,d}^{(b)}$, correspond to the transition probability and the parameter of output probability respectively. However, although a training HMM is typically implemented using the maximum likelihood (ML) criterion, the parameters of CRF are discriminatively trained because CRF directly models the conditional distribution of a speech/non-speech label sequence and the posterior probability of the correct label sequences is maximized. The proposed CRF has an advantage in that the temporal correlation of speech/non-speech labels can be modeled between successive frames, comparing it with the VAD method based on SVM. Figure 1 has a graphical model representation of a CRF for VAD.

The parameters of the CRF are trained to maximize the conditional log-likelihood $\mathcal{L}(\boldsymbol{\Lambda})$ as :

$$\begin{aligned} \hat{\boldsymbol{\Lambda}} &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\Lambda}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \sum_{n=1}^{N} \log P(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}) \end{aligned} \tag{7}$$

where $(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$, $n = 1, \ldots, N$ represents the training data. The following representation is used for simplicity.

$$P(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}) = \frac{1}{Z(\boldsymbol{x}^{(n)})} \exp\{F(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\} \tag{8}$$

Then, equation (7) can be rewritten as follows.

$$\begin{aligned} \hat{\boldsymbol{\Lambda}} &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \sum_{n=1}^{N} \log P(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \sum_{n=1}^{N} \log \frac{1}{Z(\boldsymbol{x}^{(n)})} \exp\{F(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\} \\ &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \sum_{n=1}^{N} \left[F(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}) - \log Z(\boldsymbol{x}^{(n)})\right] \end{aligned} \tag{9}$$

It can be seen from equation (9) that the model parameters that separate the function value of the correct path from that of the other paths can be estimated because the normalization factor is the sum over all possible paths of F. Although the Newton method is often used to estimate CRF parameters, we simply used a gradient method that only used the first derivative of $\mathcal{L}(\boldsymbol{\Lambda})$. The first derivative of each parameter is given in the next equation.

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\Lambda})}{\partial \lambda_k^{(a)}} &= -\sum_{n=1}^{N} \sum_{\boldsymbol{y}^{(n)}} P(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}) \sum_{t=1}^{T} f_k^{(a)}(y_{t-1}^{(n)}, y_t^{(n)}) \\ &\quad + \sum_{n=1}^{N} \sum_{t=1}^{T} f_k^{(a)}(y_{t-1}^{(n)}, y_t^{(n)}) \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\Lambda})}{\partial \lambda_{l,d}^{(b)}} &= -\sum_{n=1}^{N} \sum_{\boldsymbol{y}^{(n)}} P(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}) \sum_{t=1}^{T} f_{l,d}^{(b)}(x_{t,d}^{(n)}, y_t^{(n)}) \\ &\quad + \sum_{n=1}^{N} \sum_{t=1}^{T} f_{l,d}^{(b)}(x_{t,d}^{(n)}, y_t^{(n)}) \end{aligned} \tag{11}$$

Although equations (10) and (11) involve expectations over all possible label sequences, they can be efficiently calculated with a procedure similar to the Forward-Backward algorithm used in training of HMM (Baum-Welch algorithm). Then, the model parameters that maximize the log-likelihood can be estimated by repeating the gradient method and calculating the expectation. In detection, the best label sequence $\hat{\boldsymbol{y}}$ given an input sequence $\boldsymbol{x}$ can be written with equation (12).

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} P(\boldsymbol{y} \mid \boldsymbol{x}) \tag{12}$$

The solution to equation (12) is equivalent to the decision of passing of the trellis, and it can be obtained with a similar procedure to that of the Viterbi algorithm.

3. Features for VAD

A VAD system determines the speech section by using features that can be extracted from the input signal. The selection of the features is important because the accuracy of VAD is decided by what features are used as input sequences. We used five features in the experiment.

We extracted f0, zero-crossing, and the amplitude of each frame, and used the differences between these features of speech detection and non-speech detection for VAD. F0 is hardly ever observed in non-speech detection because it cannot be observed from unvoiced sound. Zero-crossing is the number at which the input signal crosses 0 during a given period, and the number of zero-crossings tends to increase in speech detection. Amplitude is one of the features often used for VAD. It is a very effective feature when the distance between the microphone and the speaker is short and SNR is high. It is defined as the average signal observed by using one frame.

3.1. GMM log likelihood

The log-likelihood of speech/non-speech GMMs is used as the feature for VAD. The log-likelihood of an M-mixture GMM is defined as :

$$\begin{aligned} o_t^{(gmm)} &= \log P(\boldsymbol{o}_t \mid \lambda) \\ &= \log \left[\sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \end{aligned} \tag{13}$$

where λ denotes a set of model parameters, w_i is the weight of mixture i, and $\mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the Gaussian basis function with mean vector $\boldsymbol{\mu_i}$ and covariance matrix $\boldsymbol{\Sigma}_i$.

Table 1: *Result of VAD*

	Rest. Hi		Rest. Lo		St. Hi		St. Lo	
method	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR
Baseline	24.59	6.69	38.15	23.70	7.84	60.09	29.97	44.72
GMM	18.58	25.72	27.61	48.97	20.21	14.87	25.38	27.50
Weighted GMM	31.39	12.00	48.57	24.57	28.17	4.41	38.80	9.42
$\langle C, G1, G2 \rangle$	10.60	**5.37**	33.38	28.36	13.44	1.61	21.23	**2.68**
$\langle C, G1, G2, Z \rangle$	**6.49**	8.31	36.48	23.04	**9.86**	2.09	17.27	4.01
$\langle C, G1, G2, F_0 \rangle$	13.77	6.52	33.78	28.56	13.91	1.76	19.76	3.05
$\langle C, G1, G2, A \rangle$	16.90	7.78	**28.87**	27.07	17.99	**1.42**	22.35	2.97
$\langle C, G1, G2, P \rangle$	13.36	10.50	38.21	**15.12**	11.06	4.16	**10.81**	11.41
$\langle C, G1, G2, Z, F_0 \rangle$	**8.92**	9.30	36.63	22.99	**10.69**	2.22	16.60	3.92
$\langle C, G1, G2, Z, A \rangle$	16.91	7.72	**28.91**	27.07	17.91	**1.42**	22.27	**2.96**
$\langle C, G1, G2, Z, P \rangle$	13.36	10.50	38.26	**15.12**	11.06	4.16	10.94	11.41
$\langle C, G1, G2, F_0, A \rangle$	17.26	**7.71**	29.77	26.92	17.89	1.43	21.98	2.97
$\langle C, G1, G2, F_0, P \rangle$	13.43	10.50	38.27	15.30	11.10	4.15	10.69	11.41
$\langle C, G1, G2, A, P \rangle$	14.72	10.71	39.23	17.55	12.24	3.21	**10.19**	9.61
$\langle C, G1, G2, Z, F_0, A \rangle$	17.27	**7.67**	**29.83**	26.92	17.89	**1.43**	21.98	**2.98**
$\langle C, G1, G2, Z, F_0, P \rangle$	**13.42**	10.50	38.27	**15.30**	**11.07**	4.14	10.69	11.41
$\langle C, G1, G2, Z, A, P \rangle$	14.72	10.71	39.26	17.52	12.38	3.21	**10.18**	9.72
$\langle C, G1, G2, F_0, A, P \rangle$	15.01	10.74	39.69	17.57	12.43	3.24	**10.18**	10.06
$\langle C, G1, G2, Z, F_0, A, P \rangle$	15.01	10.75	39.26	17.57	12.43	3.21	**10.18**	10.06

3.2. Log posterior probability of GMM

The posterior probability of GMM is the probability belonging to the mixture element m when the feature vector $\boldsymbol{o}_t$ is given. The posterior probability of the m-th mixture component is defined as :

$$o_t^{(pb,m)} = \frac{w_m \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \tag{14}$$

4. Experiments

4.1. Experimental condition

To evaluate the performance of the proposed method, a VAD experiment on the CENSREC-1-C database [6] was performed. CENSREC-1-C is the platform for evaluating noisy speech recognition, and contains data recorded in a real environment and data created artificially. Speech signals were sampled at a rate of 8 kHz and windowed with a 10-ms frame rate using a 25-ms Hamming window. Data recorded in two real environments were used, which were from a restaurant (Rest.) and a street (St.) with two different sound pressure levels (Hi and Lo). One set of speech and non-speech GMMs with 128 mixtures was trained using 64 files (including two environments $\times$ two SNRs) excluded from the evaluation data. The feature vectors for GMMs consisted of 12th order mel-cepstral coefficients including the zeroth coefficients and their delta and delta-delta coefficients. The features of CRF in the proposed method were defined as the log-likelihood of the GMM of speech ($G1$) and non-speech ($G2$), zero-crossing (Z), f0 (F_0), amplitude (A), the log posterior probability of GMMs (P), and the class index (C). One CRF was constructed for all four noise conditions. In other words, we assumed that the environment for the input signals was unknown and the CRF was modeled independently of the environments. The training data consisted of 64 sentences uttered by four speakers, and the evaluation data consisted of 16 sentences uttered by one speaker. The leave-one-out method was used for the data from the five speakers. The VAD system was evaluated in each frame where one frame was 10 ms, and false rejection rate (FRR) and false acceptance rate (FAR) were used.

$$\mathrm{FRR} = N_{FR}/N_s \times 100 \quad [\%] \tag{15}$$

$$\mathrm{FAR} = N_{FA}/N_{ns} \times 100 \quad [\%] \tag{16}$$

where N_s and N_{ns} denote the number of speech and non-speech frames in correct labels, and N_{FR} and N_{FA} correspond to the number of speech frames mis-detected as non-speech and vice versa.

4.2. Results

Table 1 lists the results for the restaurant and street with high SNR and low SNR, and Figures 2–5 plot the FRR and FAR under all conditions. "Baseline" in Table 1 is energy-based VAD with adaptive thresholding [6]. The threshold value was determined to minimize the value of FAR plus FRR using all environment data. "GMM" is the technique using the comparison of the log-likelihood of speech/non-speech, and "Weighted GMM" is CRF-based VAD without the transition feature function. The method denoted by $\langle \cdot \rangle$ is the CRF-based VAD we propose with the features listed in angles. For example, $\langle C, G1, G2 \rangle$ represents the CRF using the log-likelihood of speech and non-speech GMMs as features.

It can be seen from the results that the proposed method performed the best under all noise conditions. Comparing "Weighted GMM" with $\langle C, G1, G2 \rangle$, we can see the latter outperformed the former. This is because $\langle C, G1, G2 \rangle$ had the transition feature function and the temporal correlations between the speech/non-speech labels were effectively used for VAD. Although the best combination of features was different in each environment, zero-crossing seemed effective under all conditions.

Focusing on Figures 2–5, "Baseline" with varying thresholds are represented, and the proposed method is represented by the squares in the four figures. In the street environments (Figures 4 and 5), "GMM" and "Weighted GMM" obtained lower error rates than "Baseline". Furthermore, CRF-based VAD outperformed the GMM-based methods because transition and multiple observation features were used. In the restaurant envi-

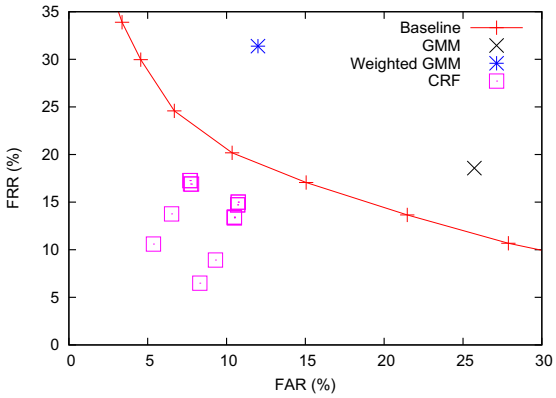

Figure 2: *VAD in restaurant with high SNR*

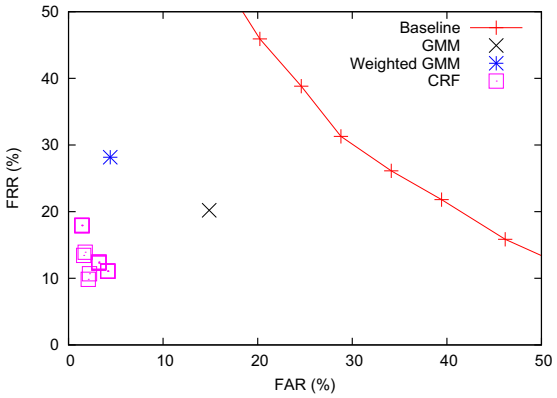

Figure 4: *VAD in street with high SNR*

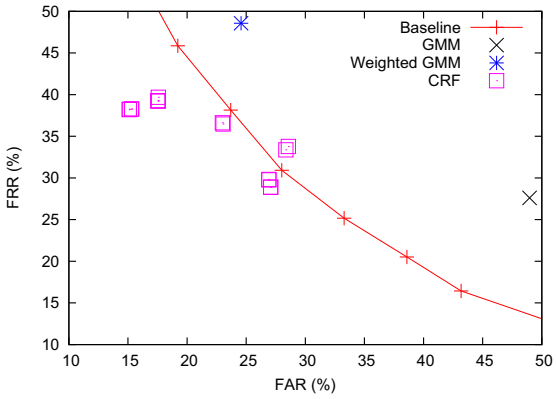

Figure 3: *VAD in restaurant with low SNR*

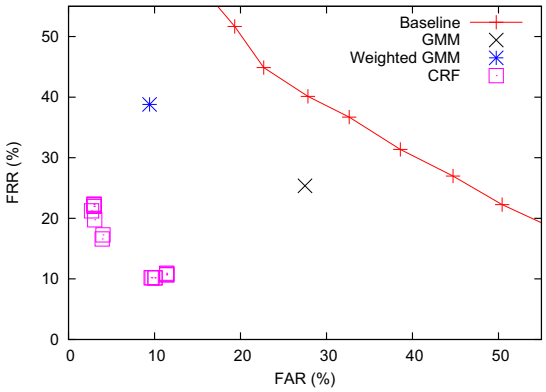

Figure 5: *VAD in street with low SNR*

ronments (Figures 2 and 3), GMM-based methods were worse than "Baseline". This may be because background noise in a restaurant includes speech noise from other people. Although the CRF-based methods used the GMM likelihood, they obtained similar or lower error rates by utilizing other features. This means that CRF estimated appropriate weights for transition and observation feature functions, and successfully selected effective features for VAD automatically.

5. Conclusion

We proposed a VAD technique based on CRF. Using two different feature functions, i.e., the transition-feature and observation-feature functions, appropriate weights could be estimated for multiple features in each speech/non-speech state. The proposed method obtained better results than the conventional methods in the experiments using only the log-likelihood of GMMs. In future work, it is necessary to evaluate the proposed approach in various environments. Investigations into useful features and an optimal model structure for VAD also needs to be done in future work.

6. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

7. References

[1] 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.

[2] N. Binder, K. Markov, R. Gruhn, and S. Nakamura, "Speech Non-Speech Separation with GMMs," in Proc. Acoustic Society of Japan Fall Meeting, Vol. 1, pp. 141–142, Oct. 2001.

[3] M. Fujimoto and T. Nakatani, "A study on Gaussian selection and probability weighting for statistical model-based voice activity detection," in Proc. Acoustic Society of Japan fall meeting, 1-1-14, pp. 43–46, Sep. 2009.

[4] Y. Kida and T. Kawahara, "Voice activity detection based on optimally weighted combination of multiple feature," in Proc. Interspeech, Sep. 2005.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. of ICML, pp. 282–289, 2001.

[6] N. Kitaoka *et al.*, "Development of VAD Evaluation Framework CENSREC-1-C and Investigation of Relationship Between VAD and Speech Recognition Performance," Proc. IEEE ASRU2007, pp. 607–612, Dec. 2007.