INTERSPEECH 2009 BRIGHTON

Voice Activity Detection Using Partially Observable Markov Decision Process

Chiyoun Park, Namhoon Kim, Jeongmi Cho

Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd.
San #14-1, Nongseo-Dong, Kiheung-Gu, Yongin-Si, Kyungki-Do, 446-712, Korea
{chiyoun.park, namhoon.kim, jmcho007}@samsung.com

10.21437/Interspeech.2009-633

Abstract

Partially observable Markov decision process (POMDP) has been generally used to model agent decision processes such as dialogue management. In this paper, possibility of applying POMDP to a voice activity detector (VAD) has been explored. The proposed system first formulates hypotheses about the current noise environment and speech activity. Then, it decides and observes the features that are expected to be the most salient in the estimated situation. VAD decision is made based on the accumulated information. A comparative evaluation is presented to show that the proposed method outperforms other model-based algorithms regardless of noise types or signal-to-noise ratio.

Index Terms: voice activity detection, partially observable Markov decision process, feature selection

1. Introduction

Voice activity detector (VAD) is a crucial component of the front-ends in many speech related applications, such as automatic speech recognition systems or codecs. The more accurately the voice region is identified, the better performance can be expected from the systems. Disambiguating speech signal from noise, however, is not a simple task, especially in low signal-to-noise ratio (SNR) condition.

To avoid the interference of the background noise in detecting voice activity, one can either extract the characteristics that are less dependent on the noise type or energy such as long term spectral divergence[1] and periodic component to aperiodic component ratio[2], or estimate the magnitude of the noise signal so that it can be compensated in the feature extraction process. By extracting the features that are more robust to noise, the performance of the voice activity detector in noisy environment is expected to improve.

Recently, model-based methods have been proposed, which keep the information of the clean speech data and generate the speech and silence probability distribution model in noisy data according to the estimated characteristics of noise. For example, the algorithm presented by Sohn et al.[3] models the discrete Fourier transform coefficients of silence and speech signals by zero-mean Gaussian distributions with different variances, and modifies the variance of the probability density function according to the estimated signal-to-noise ratio. On the other hand, the switching Kalman filter method proposed by Fujimoto et al.[4] estimates the spectral distribution of noise and merges it to the model obtained from clean environment using log-add composition method.

Whereas the above-mentioned methods provide the ways to extract the features robust to noise, and to process the extracted information based on the characteristics of the estimated noise signal, there exists a restriction that the same type of feature vector needs to be applied for all frames regardless of the changing noise conditions.

However, the features that are effective in discriminating speech from noise may be different with respect to the noise type, SNR and the existence or type of the speech signal itself. For example, the energy of a frame may be a suitable measure for VAD in high SNR situation, but in low SNR it may not be as useful as spectral features such as linear prediction coefficients. In the case of babble noise, on the other hand, the spectral characteristics of the noise resemble those of the speech itself, thus reducing the effectiveness of the spectral features. In addition, when the energy of a frame is low, it can be concluded with confidence that the frame under consideration is a silence, but when it is not as low, the signal not only can be a part of speech, but also can be a type of noise; therefore, additional information, or feature, needs to be observed to reach a more clear decision in such cases.

To overcome such difficulties, in this paper, we propose a novel method that enables the VAD system to utilize different types of features according to the estimated state of the recording environment.

The next section gives a brief review on a partially observable Markov decision process (POMDP)[5], upon which the proposed method is implemented. Sections 2.2 and 2.3 describe the formulation of the POMDP-based VAD system. The evaluation result of the proposed system is reported in Section 3, and Section 4 summarizes the paper.

2. Method

2.1. Brief Review on POMDP

A partially observable Markov decision process[6] is generally used to model an agent decision process, in which it is assumed that the agent cannot observe the current states of the system directly; instead, the agent can make observations to estimate the current states and decide on the most appropriate action based on the belief, where the appropriateness is defined to be the one maximizing a certain reward function summed over a long time period. This framework has been applied to various fields of sequential decision processes, including robot navigation[7], autonomous planning[8] and dialogue modeling for spoken dialogue systems[9].

A POMDP model is defined as (S, A, O, P_S, P_O, R). The sets S, A, O represent states, actions and observations, respectively. $P_S(s'|s, a)$ is the state transition function from state s to state s' given that the action a is performed and $P_O(o|s', a)$ is the probability of observing o when transition to state s' is done by action a. The reward function $R(s, a)$ defines the immediate reward awarded when executing action a in state

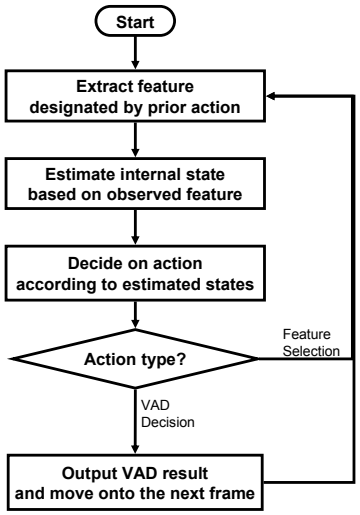


Figure 1: *Flowchart of the POMDP-based VAD*

s. At each step of the POMDP process, the action that is expected to yield the largest cumulative sum of immediate rewards, $\sum_{i=0}^{\infty} R(s_{i+1}|s_i, a_i)$, is said to be the optimal action policy.

In this paper, the POMDP framework is applied to the voice activity detection process, where continuous estimation of the hidden state—the existence of speech in a given frame—is required. By using this model, the system can be configured to decide the types of features to be observed so that only the features that are crucial in making a VAD decision in the current noise situation are considered.

The flowchart in Figure 1 briefly describes the system. The proposed algorithm first estimates the current state of the recording environment according to the observation. Based on the estimated belief of the state, the action decision process either decides on the type of the feature to be observed next, or outputs a VAD result. When the feature selection action is performed, a new feature designated by the action is extracted as an observation and the same procedure is repeated until VAD decision is made.

While it is possible to extract all the possible features at once and use them to make VAD decision, having unrelated features does not only increase the complexity of computation, but also degrade the overall performance of the VAD. To avoid the degradation, many methods have applied dimension reduction algorithms to compress the size the feature vector at the same time pruning out less salient features. However, because the efficiency of a feature depends on the noise type, SNR and speech activity, it is expected to be more rewarding to use different types of characteristics in different situations instead of having the same feature set across all kinds of noise environment.

2.2. POMDP Model for VAD

The voice activity detector is cast into a POMDP model as follows:

The state denotes the existence of the speaker activity and the surrounding environments. The proposed system factors the state into four independent states $S = (S_V, S_N, S_S, S_A)$, where each factored variable represents Voice existence, Noise type, SNR and Action history, respectively. An example of the state variables and their values is listed in Table 1(a).

The voice existence state represents whether the user is speaking or not, and has two possible values—speech and silence. However, the state variable can also be expanded to include other types of speech or noise activity which may affect the probability distribution of the observation variable. For example, when it is highly likely that the signal may contain breath sound, this state may be configured to have four possible values $S_V = \{\text{voiced, unvoiced, breath, silence}\}$, so that the difference between the breath signals and other noises and between the breath and unvoiced parts of speech can be separately modeled. The noise type state, S_N, reflects the type of noise in the background, such as car, babble or white noise. Unlike the noise type represented in the voice existence state, which has short time span such as breath or smack sounds, the types of noise this state represents are more consistent throughout time. The SNR state, S_S, stands for the loudness of the background noise. The action history state records the actions performed in the current frame, so that the same type of feature does not get extracted more than once. Otherwise, in some frames that are ambiguous, the system may decide to observe the same feature repeatedly, without ever making a decision. The action history can be represented as $S_A = \{0, 1\}^n$, where n is the number of features to be used, and 0 represents that the feature has not been observed, and 1 otherwise.

The action consists of two types of action sets: $A = A_V \cup A_F$. When the action in A_V is decided, the system outputs VAD decision according to the specified action—which is either speech-detection or silence-detection—and moves on to the next frame. Alternatively, the one in A_F selects the type of feature to be observed in the next step. As mentioned before, it should be restricted that the feature recorded in the action history should not be selected again. Any types of features that have distinct distribution between speech and noise can be applied for this system without any assumption of additivity or independence. In this paper, the energy and log-Mel spectra were used as the observed features.

The observation corresponds to the range of the extracted feature vector. In most cases, the range is $\mathcal{R}^k$ where k is the dimension of the feature vector. Because different features may have different range and because at most one feature is observed at each step, the observation can be defined as the following:

$$O = \{\text{N/A}\} \cup O_{f_1} \cup O_{f_2} \cup \cdots \cup O_{f_n} \tag{1}$$

where N/A represents the case when no feature is observed, and each O_{f_i} stands for the range of each feature $f_i \in A_F$.

2.3. Transition Functions and Rewards

The belief $b(s)$ of a state s is updated at each step according to the following equation.

$$b(s') \sim P_O(o|s', a) \sum_{s \in S} P_S(s'|s, a) b(s) \tag{2}$$

where o is the current observation and a is the decided action. The variables s and s' represent the current and updated states, respectively. The belief is normalized so that the sum of all the beliefs over the set of possible states becomes one.

It is assumed that each factored state, (S_V, S_N, S_S, S_A), is independent from one another, so that the state transition function $P_S(s'|s, a)$ can be factored as the following.

$$\begin{aligned} P_S(s'|s, a) = P_{S_V}(s'_V|s_V, a) \times P_{S_N}(s'_N|s_N, a) \\ \times P_{S_S}(s'_S|s_S, a) \times P_{S_A}(s'_A|s_A, a) \end{aligned} \tag{3}$$

(a) *State variables*

VOICE	NOISE TYPE	SNR	ACTION
Speech	Subway	Clean	(0,0)
Silence	Babble	20dB	(0,1)
	Car	10dB	(1,0)
	Exhibition	0dB	(1,1)

(b) *State transition*

	SILENCE	SPEECH
SILENCE	0.95	0.05
SPEECH	0.01	0.99

(c) *Rewards*

ACTION	STATE(VOICE)	REWARDS
Speech	Speech	10
	Silence	-100
Silence	Speech	-100
	Silence	10
Feature	Speech	-1
	Silence	-1

Table 1: *Parameters used in the POMDP-based VAD*

The transition function of the voice state P_{S_V} is handcrafted so that the voice state does not change until VAD decision action is made. When the VAD action is made, state transition occurs according to a certain transition rule T.

$$P_{S_V}(s'_V|s_V, a) = \begin{cases} 1 & \text{if } a \in A_F, s'_V = s_V \\ 0 & \text{if } a \in A_F, s'_V \neq s_V \\ T(s'_V|s_V) & \text{if } a \in A_V \end{cases} \quad (4)$$

The function T can be determined empirically by counting the number of transitions in the training data. An example of the transition rule is shown in Table 1(b). Similar rule applies to the transition of noise type and SNR states as well.

On the contrary, the action history state records the actions whenever the feature selection action occurs, and when VAD decision is made, it is reset to $\{0\}^n$.

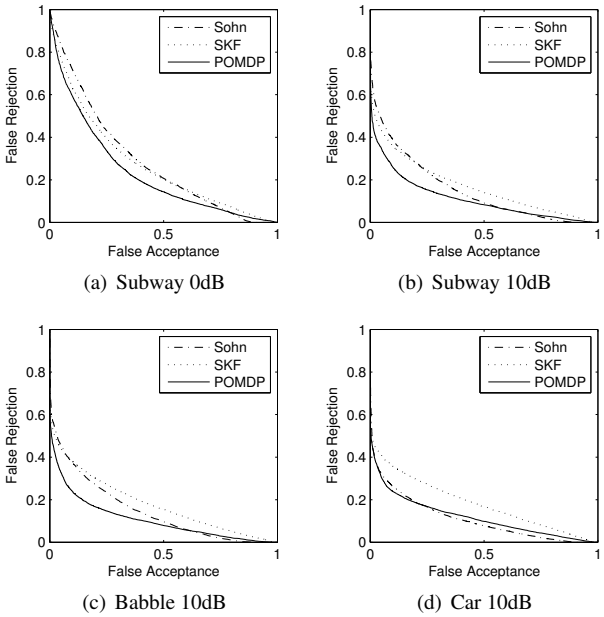
The observation function $P_O(o|s', a)$ is determined from the probability distribution of each feature designated by the prior action a in the noise and speech state s'. When the prior action is VAD decision, no observation can be made, and so only N/A should be observed.

The reward is the measure used to decide proper actions. The function can be created simply by comparing the voice existence state with the action output, as shown in Table 1(c). If the action determined the VAD output correctly, a positive reward is given, otherwise, a negative one. In addition, for each feature selection action, a small amount of negative reward is given, so that the number of steps before determining VAD decision for a frame can be optimized.

3. Experiment

3.1. Setup

Aurora-II database was used for the evaluation of the performance. This database is based on clean data sampled at 8kHz, and different types of noise signals have been artificially superimposed to get noisy data. Four types of noises—subway, car, babble and exhibition hall—were applied in the experiment, with varying SNR including 0dB, 10dB, 20dB and clean data. One hundred utterances from each type of noise and SNR were randomly chosen as training data, and another one hundred were used for evaluation.



(a) Subway 0dB (b) Subway 10dB

(c) Babble 10dB (d) Car 10dB

Figure 2: *Comparison of VAD performance*

NOISE	SNR	SOHN	SKF	POMDP
Subway	Clean	9.6	13.5	**8.7**
	20dB	16.9	21.4	**12.9**
	10dB	24.4	24.9	**18.5**
	0dB	33.9	32.2	**28.8**
Subway	10dB	24.4	24.9	**18.5**
Babble		24.1	26.3	**17.7**
Car		19.2	26.5	**19.0**
Exhibition		20.5	25.2	**17.3**

Table 2: *Comparison of equal error rates*

Two types of features were used: A_F ={energy, log-Mel spectra}. The features were extracted from each 10ms frames, with the pre-emphasis factor of 0.97 and 10ms Hamming window. 128-point DFT was performed to obtain Mel-frequency spectrum.

The proposed method was compared to other model-based methods: Sohn et al.[3] and Fujimoto et al.[4]. For Sohn's method, 128 point Fourier transform coefficients for every 10ms frame were applied, and the transition probabilities of $a_{01} = 0.2$ and $a_{10} = 0.1$ were used. The noise variance λ_N was estimated from the first ten frames of the signal. For the switching Kalman filter method of Fujimoto, 24th order log-Mel spectra calculated from every 10ms frame were used, and GMMs were trained with 32 mixture distributions.

Training of POMDP-based VAD consists of determining the model parameters and calculating the optimal policy. The parameters listed in Table 1 were used for the evaluation of the

POMDP-based VAD. As for the noise type and SNR states, they were assumed to be stationary throughout the utterance, that is,

$$P_{S_N}(s'_N|s_N, a) = \begin{cases} 1 & s'_N = s_N \\ 0 & s'_N \neq s_N \end{cases} \quad (5)$$

regardless of the action a. State transition functions and observation functions were obtained from the probability distribution of the training data. The observation set of each feature was quantized to have 100 discrete values instead of using a continuous parameter, in order to reduce complexity in calculating the optimal policy. For the purpose of drawing ROC curves, constant rewards were added and subtracted for VAD decision actions.

3.2. Results

Figure 2 shows the receiver operating characteristics (ROC) curves of the proposed VAD algorithm in comparison to other model-based methods, and Table 2 lists the equal error rate (EER) of the three methods in various noise conditions. The statistics show that the POMDP method performs better than other methods in noisy situation. Note that Sohn's method shows comparable performance in car noise, which is almost stationary throughout the utterance, but significant degradation is observed for babble and subway noise, which has large variance across time. The proposed method, however, does not exhibit such degradation in different noise types.

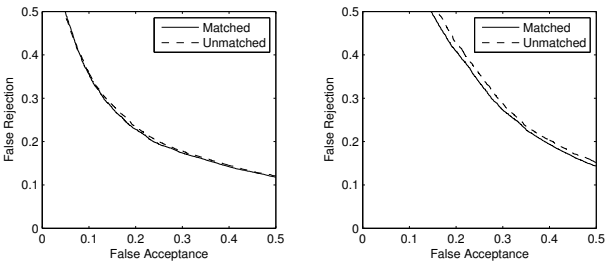
Most of the improvement is due to the belief update of the POMDP model, which tracks the type of noise and SNR, and the feature selection procedure that selects the most salient features in the given situation. From the experiment, it can be concluded that the model can estimate the type of noise and use the features that are the most effective in disambiguating speech from noise, based on the prior knowledge about possible types of noise.

It may be argued that all the expected noise types and SNRs should be included in the model, which results in the increased size of the model and high complexity of computation. However, the experiments with incomplete set of noise models prove otherwise. Figure 3(a) compares the results on Subway 5dB training set between the case when 5dB SNR is included as a separate SNR state value (matched) and the case when only 0dB, 10dB, 20dB are applied (unmatched). Similarly, Figure 3(b) compares the result on Subway 0dB noise data between the cases when the subway noise was included in and excluded from the model. It can be observed that while having a complete set of model does improve the performance of the VAD by a small amount, the effect is almost marginal; the improvement in performance in terms of EER is less than 0.5% in both cases. Therefore, it can be assumed that only the types that have significantly distinct characteristics need to be included in the model.

4. Conclusions

A possibility of applying POMDP model to a VAD system was explored. This system keeps track of multiple hypotheses about the speech activity and noise environments, and observes the types of features that are expected to be the most salient in the currently estimated situation, and outputs the VAD decision based on the accumulated information.

The proposed model needs the distribution of each feature in various noise types in different SNRs. The size of the model and the computational complexity increase as more types of noise are considered. However, it has been also shown by experiment that the performance of the VAD does not get affected seriously by applying an incomplete set of training data.



(a) 5dB (b) Subway

Figure 3: *Effect of incomplete noise state model*

Although the system presented in this paper utilizes two types of features, that is, energy and log-Mel spectra, the number of features does not need to be restricted. On the contrary, it is expected that the performance of the voice detection can be improved by including more characteristics that capture diverse aspects of speech. In the future, we are planning to apply different types of features, such as periodicity information or zero crossing rate, and measure the performance change and examine POMDP's action in various noise conditions.

5. References

[1] Ramirez, J., Segura, J. C., Benitez, C., Torre, A. and Rubio, A., "Efficient Voice Activity Detection Algorithms Using Long-term Speech Information", Speech Communication 42:271–287, 2004.

[2] Ishizuka, K. and Nakatani, T., "Study of Noise Robust Voice Activity Detection Based on Periodic Component to Aperiodic Component Ratio", Proceedings on SAPA 2006, 65–70, 2006.

[3] Sohn, J., Kim, N. S. and Sung W., "A Statistical Model-Based Voice Activity Detection", IEEE Signal Processing Letters 6(1):1–3, 1999.

[4] Fujimoto, M. and Ishizuka, K.,"Noise Robust Voice Activity Detection Based on Switching Kalman Filter", IEICE Transactions on Information and Systems E91-D(3):467–477, 2008.

[5] Sondik, E. J., "The Optimal Control of Partially Observable Markov Decision Processes Over the Infinite Horizon", Operations Research, 26:282–304, 1978.

[6] Kaelbling, L. P., Littman, M. L. and Cassandra, A. R., "Planning and Acting in Partially Observable Stochastic Domains", Artificial Intelligence 101:99–134, 1998.

[7] Simmons, R. and Koenig, S., "Probabilistic Navigation in Partially Observable Environments", Fourteenth International Joint Conference on Artificial Intelligence, 1080–1087, 1995.

[8] Eckles, J. E., "Optimum Maintenance With Incomplete Information", Operations Research 16:1058–1067, 1968.

[9] Williams, J. D. and Young, S., "Partially Observable Markov Decision Processes for Spoken Dialog Systems", Computer Speech and Language 21:393–422, 2007.