Partially observable Markov decision process (POMDP) has recently been applied to a voice activity detector (VAD), which makes it possible to incorporate knowledge about the recording environments in the decision process in order to achieve a more stable performance in noisy situations. In this paper, the model has been further explored to utilize prior knowledge about possible intermittent noise signals such as breath or click sounds, in addition to the stationary background noise types. The experimental result shows that application of sporadic noise models in a POMDP-based VAD reduces the equal error rate of the voice activity decision by about 7% relatively.

Index Terms—voice activity detection, acoustic signal detection, partially observable Markov decision process

1. INTRODUCTION
A partially observable Markov decision process (POMDP) is the framework that models an agent that performs a sequence of goal-oriented actions under uncertain conditions. This framework has been applied to various fields of sequential decision processes, including robot navigation, autonomous planning and dialogue modeling for spoken dialogue systems[1].

In Park et al.[2], the POMDP framework has been applied to the voice activity detection (VAD) process, where continuous estimation of the hidden state—the existence of speech in a given frame—is required. By applying this model, the system decides the types of features to be observed based on the estimated noise condition, so that only the features that are crucial in making a VAD decision under the hypothesized recording environment can be utilized. The experimental result shows that the prior knowledge about the possible noise condition in the recording environment can be easily applied in the VAD decision process, which improves the performance of the VAD at the same time increasing the stability of the system across different noise conditions.

The noise environment used in the system included subway, car, babble and exhibition hall. These types of noise, however, are almost stationary across time and do not fully describe possible occurrences of the sporadic noises in a real recording situation. For example, the speaker may take a deep breath before recording or make lip-smacking sounds between phrases. Such noise types are mostly short and do not depend on the recording condition as much.

In this paper, the use of a POMDP model in the VAD system is further explored to utilize knowledge about such occasional noises, as well as the stationary environmental noise. By providing more information about possible conditions, the system is expected to make a better hypothesis, which leads to an improved performance.

2. METHOD
2.1. Review on POMDP
A partially observable Markov decision process is generally used for modeling 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.

A POMDP model is defined as \((S, A, O, P_S, P_O, R)\), as shown in the influence diagram in Figure 1. The sets \(S, A, O\) represent states, actions and observations, respectively. \(S\) is the actual state of the system, which the agent cannot have direct access to. Instead, the agent can make observations \(O\), which partly depend on the state, so that it can make informed decision \(A\), which affects the process of the given system.
The system works in accordance to the Markov property, that is, the state changes according to the previous state $s$ and the action $a$ which the agent has decided to perform on the system. The state transition is represented by $P(s'|s,a)$. The observation that the agent can make is also affected by the agent’s previous action, which is represented by $P(o|s',a)$. The agent’s action is evaluated by a measure called reward $R$. The reward is defined by what action $a$ the agent did in the course of the state transition $s$ to $s'$. The agent must make a decision that maximizes the sum of the expected rewards, $\sum_{t=0}^{\infty} R(s_{t+1}|s_t, a_t)$.

### 2.2. POMDP-Based VAD

In Park et al. [2], the POMDP framework has been applied to the VAD process. Each state of the VAD process corresponds to a frame of the acoustic signal. The agent wants to know whether the speaker is talking or not in the given frame, but that information is not directly accessible. Instead, the agent can make observations which are known to be related to it. Such observations include signal energy, spectral properties, and so on. By making observations, the agent can decide whether the frame under consideration includes speech segment or not. If the observations collected up to the point are not clear enough, the agent may make further action to observe additional type of features, delaying the decision.

The system works in the following steps. At each frame, the proposed algorithm first extract feature designated by previously-decided action. Then, it estimates the current state of the recording environment according to the observation. Based on the estimated belief of the state, the agent either decides on the type of the feature to be observed next, or outputs a VAD result and moves on to process the next frame. 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.

Figure 2 shows the influence diagram of the POMDP model modified for VAD. Note that the POMDP state has been factored into four different states; each represents a different aspect of the recorded signal that needs to be estimated or recorded by the agent in order to make a more efficient VAD decision. $S_V$ is the existence of the voice in the given frame, $S_N$ is the type of the background noise, $S_S$ is the SNR of the noise, and $S_A$ records the history of the agent’s action in the current frame. The action is also divided into two classes. The agent can either decide whether the speaker is talking or not ($A_V$), or observe other types of features for a more confident decision ($A_F$). The reward is evaluated by how fast and accurately the VAD decision is made. Therefore, when the agent chooses to observe additional feature, the agent gets negative reward due to delayed decision. When the agent makes a VAD decision, the reward is given by comparing the actual speech existence ($S_V$) and the agent’s decision ($A_V$).

By applying the POMDP model to a VAD process as explained above, the VAD agent can estimate the current noise situation, and utilize the features that are expected to be the most effective in discriminating the speech signal from background noise in the estimated noise environment. For more detailed information about the POMDP-based VAD system, refer to Park et al. [2].

### 2.3. Extending Speech State

The experiment in Park et al. shows that the use of POMDP in a VAD system not only improves the performance of the VAD, but also makes the result more stable across different characteristics of noise. However, the noise types considered in Park et al. depend on the recording environments and are rather stationary throughout the signal. In this paper, therefore, another type of noise is considered that depends less on the recording situation, and has shorter time span.

For example, look at the signal in Figure 3. The breath region is the sound of the speaker inhaling deeply preparing for making an utterance, and the portion labeled ‘click’ is the sound of the user clicking on a button after the utterance. Such noises may not be observed often in a well-controlled recording sessions, but they tend to happen occasionally in actual situations with portable recording devices. Such noises may become nuisances in speech recognition since the breathing sound resembles aspiration signal, and the button click may be mistaken for a stop burst.

In this paper, we will apply such noise models to a
POMDP-based VAD. To be able to distinguish such spurious noises, the speech state \( S_V \) needs to be extended. That is, instead of using only SPEECH and SILENCE, additional states BREATH, CLICK and FRICATION are included. The state FRICATION is not a noise model, but it represents fricatives and aspiration sounds. This is added to avoid friction \( f, z \) from being mistaken for BREATH. Note that \( S_V \) is independent from environmental noise \( S_N \). While the noise types in \( S_N \) occur simultaneously with either speech or breathing signal, the BREATH or CLICK sound does not coincide with speech, and spans for a short period of time only.

The transition matrix of \( S_V \) is shown in Table 1. The CLICK has very high probability to make transition to other states, which means that the click sound tends to have a very short time span. On the other hand, BREATH has rather high probability of remaining in the current state, which makes the noise longer. It is assumed that the transition probability is independent from the type of environmental noise \( S_N \).

The observation depends not only on \( S_V \), but also on the noise environment and the action selected by the agent (that is, the type of extracted feature.) Figure 4 shows the observation probability distribution of the signal energy in the inhouse environment. Note that the breath signal has medium-level energy, which is greater than silence but is less than the speech signal. On the other hand, the distribution of the click sound is mostly concentrated in the high energy region.

Although the states have been subdivided into five different classes, the VAD does not have to decide which of the five classes the frame belongs to. It only needs to determine whether the signal in question is a speech or not. Therefore, the VAD action is not modified; instead, the reward function is adjusted accordingly. Table 2 shows the changed reward.

### 3. EXPERIMENT

#### 3.1. Setup

The performance of the VAD system was evaluated by the utterances recorded in the actual noise environment, using a portable recording device with a push-to-talk button, so that intermittent noises such as inhaling and button clicks can be observed more than in the controlled recording situations. Each utterance consists of a short word or a phrase with less than ten syllables which spans for 950ms on average. The recordings were done by ten people in nine different locations, including restaurant, house and subway. Because the recording situations were not controlled, the SNRs vary not only across different locations, but also across individual utterances. For example, the SNRs in house conditions vary across 15–25dB, and the SNRs in bus and mart ranges within 5–15dB. Twenty utterances from each recording environment were used for the performance evaluation.

Because the distinction in SNR is not meaningful, the state \( S_N \) was disabled for this experiment. Instead, \( S_N \) consists of nine recording situations. It is assumed that there is no transition in \( S_N \) during an utterance.

Two types of observations are made: one being the signal energy, and the other being the log-Mel spectra.

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 \( \lambda_N \) 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.

To be able to verify the effect of using spurious noise models, the POMDP-based VAD with only \( S_V = \{ \text{silence, speech} \} \) was applied as well.

#### 3.2. Performance

Figure 5 shows the receiver operating characteristics (ROC) curves of the proposed VAD algorithm in comparison to other
methods, and Figure 6 compares the equal error rates (EER) of the four methods in different noise conditions. Note that both the POMDP models outperform the other two methods in all of the noise environments, and the application of the intermittent noise model improves the performance. The EER changed from 12.4% to 11.5%, achieving reduction of 7% on average relative to the POMDP without noise model.

3.3. Belief State

Figure 7 shows the change in the belief of $S_V$ during the processing of an utterance. The last state in each frame was plotted, since a frame may be processed multiple times if the agent chooses to observe additional features instead of making a VAD decision. It can be observed that the breath state is correctly assumed during the inhaling sound signal, and click state gets more belief whenever the energy increases abruptly, but dies down quickly if the energy spans for more than three frames. The click state does not get the largest belief during the button click sound, but it is high enough to make non-speech decision.

4. CONCLUSION

The use of POMDP model to VAD systems has been explored. In this paper, we have shown that prior knowledge about possible sporadic noise signals, in addition to stationary noise from recording environment, can be easily incorporated to VAD processes by using POMDP models. The experiments show that the agent can make a more detailed hypothesis by incorporating the model of occasional noise, thus achieving a better VAD conclusion. The spurious noise models reduce the equal error rate of the VAD result by as much as 7%.

Although the system presented in this paper utilizes two types of features, 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. Furthermore, the features that can be used in this algorithm do not have to be speech-specific, that is, other modalities such as gesture or image-based information can be applied without modifying the algorithm. This means that the presented algorithm can be extended for multimodal input based VAD as well.

5. REFERENCES