Suppressing Unnecessary Clarification Requests for Unknown Word Acquisition in Spoken Dialogue Using Syllable-Based ASR Confidence

Takumi Furuta, Ryu Takeda and Kazunori Komatani
SANKEN, The University of Osaka
8-1 Mihogaoka, Ibaraki, Osaka, Japan
{takumi-furuta@ei.,rtakeda@,komatani@}sanken.osaka-u.ac.jp

Abstract

Clarification requests are a promising way for spoken dialogue systems to acquire unknown words from users, but asking too often can burden users. Because unknown words are not in the system's vocabulary, utterances must first be represented as syllable sequences, which are then segmented into words. In this setting, syllable-based automatic speech recognition (S-ASR) errors can cause utterances containing only known words to appear to contain unknown words, leading to unnecessary clarification requests. To address this issue within a stream-based active learning framework, we extend the reinforcement learning policy state with recognition reliability features. Specifically, we incorporate two confidence measures derived from S-ASR to make clarification request selection sensitive to S-ASR errors. We further incorporate segmentation confidence over N-best hypotheses to reduce the impact of minor S-ASR errors. Experiments using prerecorded speech data showed that the number of clarification requests on utterances affected by S-ASR errors was reduced by 1.34. The area under the learning curve for word segmentation also numerically increased by 0.07.

1 Introduction

Clarification requests are a promising way for spoken dialogue systems to acquire unknown words. Such words include personal nicknames, community-specific abbreviations, and newly coined terms, which are unpredictable and user-specific by nature and cannot be covered even by large language models. For example, a home assistant may fail to recognize a family member's nickname, or a care robot may misunderstand a patient's personal term for a medication. Moreover, unknown words are often misrecognized as acoustically similar known words in automatic speech recognition (ASR) (Kumar et al., 2024). Such errors then pass silently to downstream language understanding, leaving no opportunity for correction. Through clarification requests, information about unknown words can be obtained from the user for future interactions. In this work, we focus on clarification request selection for acquiring pronunciation-related information about unknown words under speech input.

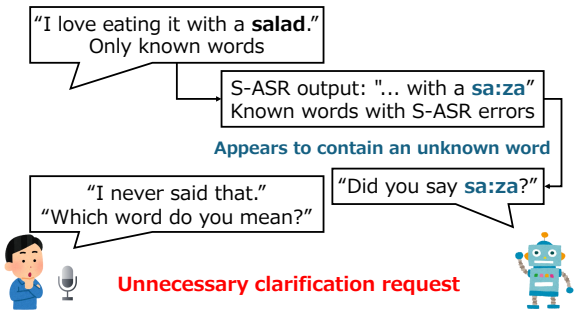


Figure 1: Example of an unnecessary clarification request caused by S-ASR errors in an utterance containing only known words.

However, under speech input, syllable-based ASR (S-ASR) errors can trigger clarification requests even for utterances that contain no unknown words. To issue clarification requests selectively, identifying whether each utterance contains unknown words is necessary. Since the written form of unknown words is unavailable from speech alone, utterances must first be represented as syllable sequences by S-ASR and then segmented into words. As illustrated in Figure 1, S-ASR errors can transform an utterance containing only known words into a syllable sequence that cannot be explained by known words. A clarification request may then be issued, forcing the user to infer what was misunderstood and providing no useful information about unknown words. Furthermore, users are not accustomed to being interrupted by such requests (Hancock et al., 2019). Repeatedly asking the same type of question may also degrade user impression (Komatani and Nakano, 2020). There-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 51–61
August 2–5, 2026.

fore, minimizing the number of such erroneous clarification requests is critical. More generally, requests should be issued only when they are likely to be informative.

Stream-based active learning, originally proposed by Atlas et al. (1989), provides a framework for deciding when to issue clarification requests while keeping their number as small as possible. Fang et al. (2017) showed that the decisions in this framework can be optimized via reinforcement learning (RL). Previous work from our group (Waki et al., 2025) formulated clarification request selection for unknown word acquisition as a stream-based active learning problem and optimized the selection policy using RL. However, their formulation assumed text input, where S-ASR errors do not arise.

To suppress unnecessary clarification requests under speech input, we propose making clarification request selection responsive to S-ASR errors. Specifically, we incorporate S-ASR confidence into the RL policy state and further incorporate segmentation confidence over N-best alternative hypotheses to reduce the impact of minor S-ASR errors. Experiments using pre-recorded speech data showed that these features reduced unnecessary clarification requests while maintaining unknown word acquisition efficiency.

The contributions of this work are twofold:

- We address unknown word acquisition through spoken dialogue under speech input, where S-ASR errors must be accounted for.
- We propose incorporating S-ASR confidence features and segmentation confidence over N-best hypotheses into the RL policy state. This reduces unnecessary clarification requests caused by S-ASR errors, addressing one source of potential user burden during unknown word acquisition.

2 Related Work

Handling unknown words is a general challenge, but prior work has not treated ASR errors as a central challenge in acquiring them through spoken dialogue. Explicit clarification requests have long been used to directly ask users for information about unknown words (Purver, 2002), while implicit confirmations have been studied to reduce the burden of repetitive explicit questions (Ono et al., 2017). However, this line of work assumes that the target term can already be identified from the user's utterance and therefore does not address unknown word acquisition when ASR errors are present.

Research on out-of-vocabulary (OOV) words in ASR typically assumes that candidate target forms can be enumerated in advance. This assumption underlies approaches that bias recognition toward pre-specified terms (Williams et al., 2018). Other approaches use smaller units to improve recognition of rare or unseen forms (Parada et al., 2011). In contrast, our target words are user-specific or newly coined terms. Their forms are unavailable at training time, which motivates acquiring them through clarification requests in dialogue.

Confidence measures derived from ASR have been used to detect ASR errors (Jiang, 2005), and prior work has applied them to several downstream purposes. In ASR post-processing, posterior-based scores have been used to identify where an OOV word likely occurs in the output (Kombrink et al., 2009), and end-to-end work has exploited Connectionist Temporal Classification (CTC) and attention alignments for the same purpose (Egorova et al., 2021). In dialogue systems, Skantze (2007) showed that ASR confidence scores are effective for grounding decisions. Prior work has not incorporated ASR confidence into the policy states used for selecting clarification requests for unknown words.

Dialogue policy learning has also been studied under uncertainty caused by ASR errors. Henderson et al. (2014) used ASR uncertainty for dialogue state tracking. Williams and Young (2007) modeled spoken dialogue systems as Partially Observable Markov Decision Processes, treating ASR results as uncertain observations for dialogue management. These studies use ASR-related uncertainty to infer dialogue states or select actions for ongoing dialogue. In contrast, our policy uses S-ASR confidence to decide whether an utterance provides useful evidence for learning an unknown word.

Stream-based active learning (Atlas et al., 1989; Settles, 2009), where the learner receives instances sequentially and decides at each step whether to query the oracle, has typically assumed that the input itself is reliable. Prior work on active learning under noisy input has mainly operated in a pool-based setting (Hadian and Sameti, 2014), where density-based strategies can globally identify noisy instances.

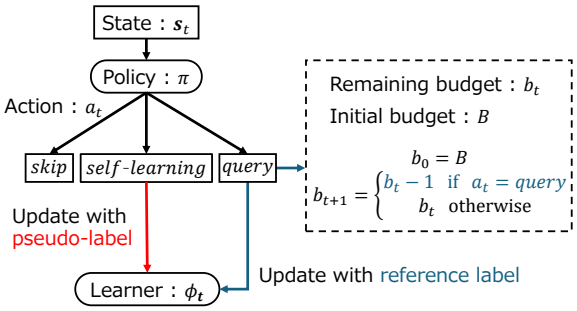


Figure 2: Process overview of stream-based active learning for unknown word acquisition.

3 Preliminary

Following Waki et al. (2025), we formulate clarification request selection as a stream-based active learning problem, originally introduced by Atlas et al. (1989). The goal is to learn a policy that determines when to issue clarification requests so as to maximize the learner's performance within a limited budget B. At each step $t \in \mathbb{N}$, a state $\mathbf{s}_t$ is observed and an action $a_t \in \{query, self\text{-}learning, skip\}$ is selected according to the policy, as illustrated in Figure 2. *Query* issues a clarification request, consuming one unit of budget B, and obtains a reference label from the oracle (i.e., the user) to update the learner. The remaining budget b_t decreases by one per *query*, and the episode terminates when $b_t = 0$. The remaining actions, *self-learning* and *skip*, consume no budget. *Self-learning*, following Waki et al. (2025), updates the learner using a pseudo-label (Lee, 2013), while *skip* performs no update.

This decision process corresponds to a Markov Decision Process (MDP) (Fang et al., 2017), in which the policy $\pi(a_t|\mathbf{s}_t)$ is optimized using reinforcement learning. Since selecting *query* updates the learner and changes the utility of future actions, a short-sighted policy is insufficient. The immediate reward r_{t+1} is defined as the improvement in learner performance before and after each action. In our experiments, the concrete definition of this reward is given in Section 5.2. The objective is to learn a policy $\pi(a_t|\mathbf{s}_t)$ that maximizes the expected discounted cumulative reward $\mathbb{E}\left[\sum_t \gamma^t r_{t+1}\right]$, where $\gamma \in [0, 1]$ is a discount factor. The Q-value function $Q^\pi(\mathbf{s}_t, a_t)$ is approximated using a Deep Q-Network (DQN) (Mnih et al., 2015), which takes $\mathbf{s}_t$ as input and outputs Q-values for each action. Waki et al. (2025) defined the base state as consisting of the learner uncertainty $\mathbf{u}_t$ (Lewis, 1995) and the normalized remaining budget $\rho_t = b_t/B$. The base state is written as

$$\mathbf{s}_t^{(\text{base})} = (\mathbf{u}_t, \rho_t). \quad (1)$$

4 Proposed Method

To handle clarification request selection under speech input, this section formulates the problem setting and introduces the proposed state representation. We first describe the speech-input setting and then present the state features used in the proposed method.

4.1 Problem Setting under Speech Input

In this work, the input is a speech signal $\mathbf{x}_t$. Under speech input, the learner does not observe a reliable text sequence directly. Instead, it receives a 1-best hypothesis $\hat{\mathbf{y}}_t = (\hat{y}_{t,1}, \ldots, \hat{y}_{t,|\hat{\mathbf{y}}_t|})$ from S-ASR, where $|\cdot|$ denotes the length of a syllable sequence. This hypothesis is segmented into a boundary sequence by the word segmentation model with parameters ϕ_t. A posterior distribution over word boundary sequences is written as $P(\mathbf{c}_t|\hat{\mathbf{y}}_t, \phi_t)$, where word boundary sequences specify where word boundaries should be placed in $\hat{\mathbf{y}}_t$. Here, $\mathbf{c}_t \in \{0, 1\}^{|\hat{\mathbf{y}}_t|-1}$ is a binary boundary sequence, with $c_{t,j} = 1$ indicating a boundary between $\hat{y}_{t,j}$ and $\hat{y}_{t,j+1}$ for $j = 1, \ldots, |\hat{\mathbf{y}}_t| - 1$. The predicted segmentation is given by $\hat{\mathbf{c}}_t = \text{argmax}_{\mathbf{c}_t} P(\mathbf{c}_t|\hat{\mathbf{y}}_t, \phi_t)$. An unknown word is defined as a word not yet present in the vocabulary of the word segmentation model at step t.

In this stream-based active learning setting, the target information to be acquired is the correct word boundaries for syllable sequences containing unknown words. Each action therefore determines the training data $\mathcal{D}_t$ used to update the word segmentation model. When *query* is selected, the reference syllable sequence $\mathbf{y}_t^*$ and boundary sequence $\mathbf{c}_t^*$ are obtained from the oracle, which may differ from $\hat{\mathbf{y}}_t$ and $\hat{\mathbf{c}}_t$ due to S-ASR errors. For *self-learning*, the pseudo-label $(\hat{\mathbf{y}}_t, \hat{\mathbf{c}}_t)$ is used in place of the oracle label. *Skip* performs no update. These three cases are summarized as follows:

$$\mathcal{D}_t = \begin{cases} \{(\mathbf{y}_t^*, \mathbf{c}_t^*)\} & \text{if } a_t = query \\ \{(\hat{\mathbf{y}}_t, \hat{\mathbf{c}}_t)\} & \text{if } a_t = self\text{-}learning \\ \emptyset & \text{if } a_t = skip \end{cases}. \quad (2)$$

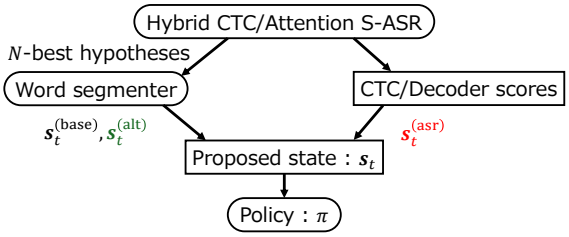


Figure 3: Overview of the proposed state representation constructed from S-ASR confidence and segmentation confidence of alternative hypotheses.

4.2 Proposed State: Syllable-Based ASR Confidence

We extend the base state with S-ASR confidence features, denoted by $\mathbf{s}_t^{(\mathrm{asr})}$, to capture recognition reliability. Figure 3 gives an overview of the proposed state representation. While $\mathbf{s}_t^{(\mathrm{base})}$ reflects segmentation uncertainty given the 1-best hypothesis, $\mathbf{s}_t^{(\mathrm{asr})}$ reflects the reliability of the recognized hypotheses. The resulting state is defined as

$$\mathbf{s}_t = (\mathbf{s}_t^{(\mathrm{base})},\ \mathbf{s}_t^{(\mathrm{asr})}). \tag{3}$$

Figure 4 illustrates how these features are computed from an N-best hypothesis set. We use its 1-best hypothesis for posterior features and the full $N^{(\mathrm{asr})}$-best set for score entropy. Following the Hybrid CTC/Attention recognizer of Watanabe et al. (2017), let $P^{(\mathrm{ctc})}(\mathbf{y}_t|\mathbf{x}_t)$ and $P^{(\mathrm{dec})}(\mathbf{y}_t|\mathbf{x}_t)$ denote the output probabilities of the CTC branch and the attention decoder, respectively. Using their combined score,

$$\begin{aligned} S(\mathbf{y}_t|\mathbf{x}_t) = \lambda \log P^{(\mathrm{ctc})}(\mathbf{y}_t|\mathbf{x}_t) \\ + (1-\lambda)\log P^{(\mathrm{dec})}(\mathbf{y}_t|\mathbf{x}_t), \end{aligned} \tag{4}$$

we generate an N-best hypothesis set $\{\mathbf{y}_t^{(n)}\}_{n=1}^{N}$, where $\mathbf{y}_t^{(1)}$ is the 1-best hypothesis. We use both the CTC branch and the attention decoder, since they provide complementary signals of recognition reliability.

From the 1-best hypothesis, we derive 1-best posterior features for each branch. For branch $m \in \{\mathrm{ctc}, \mathrm{dec}\}$, let $P_{t,j}^{(m,1)}$ denote the posterior probability of the j-th syllable in the 1-best hypothesis, and let $\bar{P}_t^{(m)}$ denote its arithmetic mean over syllables. We compute the geometric mean $\mu_t^{(m)}$ as an overall confidence feature and the standard deviation $\sigma_t^{(m)}$ as a feature of syllable-level confidence variation.

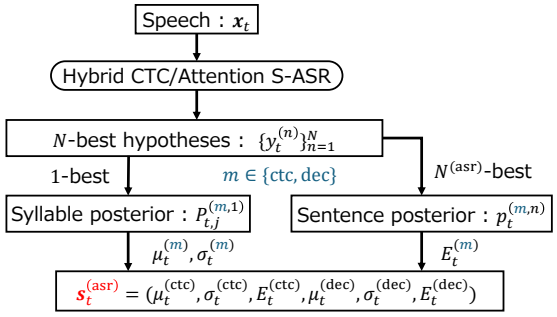


Figure 4: Computation of $\mathbf{s}_t^{(\mathrm{asr})}$ from an N-best hypothesis set. The 1-best hypothesis is used to compute posterior features, and the $N^{(\mathrm{asr})}$-best set is used to compute entropy for each branch.

From the same N-best set, we derive an N-best score entropy feature for each branch. For each branch m, let $p_t^{(m,n)}$ denote the softmax normalized score of the n-th hypothesis in the N-best set. We define $E_t^{(m)}$ as the entropy of this distribution, which represents how concentrated or dispersed the scores from each branch are over competing hypotheses.

The resulting S-ASR confidence state consists of three features from each of the two branches. Specifically, it includes $\mu_t^{(m)}$, $\sigma_t^{(m)}$, and $E_t^{(m)}$ for both $m \in \{\mathrm{ctc}, \mathrm{dec}\}$. These features and the resulting S-ASR confidence state are defined as follows:

$$\mu_t^{(m)} = \exp\left(\frac{1}{|\mathbf{y}_t^{(1)}|}\sum_{j=1}^{|\mathbf{y}_t^{(1)}|}\log P_{t,j}^{(m,1)}\right) \tag{5}$$

$$\sigma_t^{(m)} = \sqrt{\frac{1}{|\mathbf{y}_t^{(1)}|}\sum_{j=1}^{|\mathbf{y}_t^{(1)}|}\left(P_{t,j}^{(m,1)} - \bar{P}_t^{(m)}\right)^2} \tag{6}$$

$$E_t^{(m)} = -\sum_{n=1}^{N^{(\mathrm{asr})}} p_t^{(m,n)} \log p_t^{(m,n)} \tag{7}$$

$$\begin{aligned} \mathbf{s}_t^{(\mathrm{asr})} = (&\mu_t^{(\mathrm{ctc})},\ \sigma_t^{(\mathrm{ctc})},\ E_t^{(\mathrm{ctc})}, \\ &\mu_t^{(\mathrm{dec})},\ \sigma_t^{(\mathrm{dec})},\ E_t^{(\mathrm{dec})}). \end{aligned} \tag{8}$$

4.3 Proposed State: Segmentation Confidence of Alternative Hypotheses

We further incorporate segmentation confidence of alternative hypotheses into the policy state to account for minor S-ASR errors. Unknown words often co-occur with minor S-ASR errors, so clarification may still be needed even when the 1-best hypothesis is slightly corrupted. However, such cases are not always captured by $\mathbf{s}_t^{(\mathrm{asr})}$ alone.

Useful information for clarification can remain in alternative hypotheses beyond the 1-best. When the correct sequence is included among these alternatives, segmentation confidence may be higher for an alternative hypothesis than for the 1-best hypothesis. This contrast can suggest that the uncertainty of the 1-best hypothesis may be caused by an S-ASR error rather than by a genuine unknown word.

To capture this information, we define an additional feature $\mathbf{s}_t^{(\mathrm{alt})}$ from word segmentation results for alternative hypotheses. For each alternative hypothesis $\mathbf{y}_t^{(n)}$ ($n \geq 2$), the word segmentation model estimates a posterior distribution over boundary sequences. From this distribution, we extract the top-$K^{(\mathrm{alt})}$ boundary probabilities $\kappa_t^{(n,k)} = P\left(\mathbf{c}_t^{(n,k)} | \mathbf{y}_t^{(n)},\ \phi_t\right)$, where $k = 1, \ldots, K^{(\mathrm{alt})}$ indexes the boundary candidate. These probabilities are used as features in $\mathbf{s}_t^{(\mathrm{alt})}$. Only alternative hypotheses ($n \geq 2$) are used, since the uncertainty of the 1-best segmentation is already reflected in $\mathbf{s}_t^{(\mathrm{base})}$. The final proposed state is therefore defined as

$$\mathbf{s}_t = \left(\mathbf{s}_t^{(\mathrm{base})}, \mathbf{s}_t^{(\mathrm{asr})}, \mathbf{s}_t^{(\mathrm{alt})}\right).$$

5 Experiments

This section describes the experimental design used to evaluate the proposed method. The evaluation examines whether the method can reduce clarification requests caused by S-ASR errors while maintaining unknown word acquisition performance. We first describe the dataset and setup, and then present the compared conditions and evaluation metrics.

5.1 Dataset

The dataset contains Japanese spoken dialogues designed for unknown word acquisition scenarios. It consists of 900 user utterances from 47 speakers. Each dialogue begins with a user utterance containing one target unknown word and is followed by several turns in which the system confirms that word with the user. Foreign food names were used as target unknown words and were treated as newly coined terms for the system. The dataset contains 14 such words, including Eisbein, a German pork dish, and Dal Bhat, a Nepalese dish.

The dataset was divided into Stream and Eval sets so that policy interaction and segmentation evaluation were separated. Table 1 summarizes the four splits. The Stream sets were used as sequential inputs for policy training and evaluation. The Eval sets consisted of human reference transcriptions containing the same target unknown words as the Stream sets. Within Stream, Stream Train was used during DQN training and Stream Test was used during policy evaluation. Within Eval, Eval Train was used to compute the reward during DQN training, and Eval Test was used to evaluate the learned policy through word segmentation performance. No speaker appeared in both the training and test splits.

Table 1: Dataset splits used in the experiments. Character error rate (CER) was computed on 1-best S-ASR hypotheses and is reported only for the Stream splits.

Type	Split	No. of utterances	CER (%)
Stream	Train	336	7.22
Stream	Test	261	5.16
Eval	Train	148	–
Eval	Test	155	–

5.2 Setup

We evaluated policies in a controlled simulated environment using pre-recorded speech data. At each step, the policy selected an action from the current state, and the word segmentation model was updated accordingly. Because the input stream was fixed, the policy did not affect subsequent inputs, but its actions affected future states through updates to the word segmentation model. In each episode, the policy processed one finite input stream, and the episode ended when either all utterances in the stream had been processed or the remaining budget reached zero. The budget was initialized to $B = 15$ per episode, which was sufficient to cover all 14 unknown words. This controlled setting also isolated clarification request selection from variation in request phrasing and user responses. In our formulation, the policy decided whether to issue a clarification request, but not how it was phrased. By fixing these factors, the setup allowed us to evaluate the selection policy itself more directly.

The evaluation trials summarized variability over both policy initialization and Stream order. For each condition, we trained 20 policies with different random seeds. Each trained policy was evaluated over 20 test episodes using the same Stream Test split with different utterance orders. Thus, the 400 evaluation trials were not independent draws of

different test data. For significance testing, we first averaged each metric over the 20 episode orders for each trained policy. We then conducted two-sided paired permutation tests on the resulting 20 seed-level values. Confidence intervals were computed by paired bootstrap over the same seed-level differences.

For S-ASR, we used a Japanese Hybrid CTC/Attention model[1] previously trained by Takeda and Komatani (2025) and implemented in ESPnet (Watanabe et al., 2018). We set the beam width to 20 and the CTC weight λ to 0.2, without a language model. These parameters were selected because they achieved the highest unknown-word recognition performance on the audio corresponding to the Eval Train transcriptions among the tested values. For word segmentation, we used the Nested Pitman-Yor Language Model (Mochihashi et al., 2009), pre-trained on the Corpus of Spontaneous Japanese (Maekawa et al., 2000).

The policy was optimized via reinforcement learning using a performance-based reward. The reward r_{t+1} was defined as the improvement in micro-F1 of word segmentation before and after each action, measured on the evaluation training set, calculated as

$$r_{t+1} = \text{micro-F1}(\phi_{t+1}) - \text{micro-F1}(\phi_t). \quad (9)$$

Here, micro-F1 was computed over word boundary positions across all utterances in the evaluation training set. We used micro-F1 for the reward because it provided a robust signal for small changes in segmentation performance.

The policy network and training procedure were shared across all RL-based conditions. We used a DQN with three extensions. Double DQN (van Hasselt et al., 2016) decouples action selection from evaluation to reduce overestimation of Q-values, and the target network was updated every 200 steps. Noisy Networks for Exploration (Fortunato et al., 2018) adds factorized Gaussian noise with scale $\sigma_0 = 0.4$ to encourage exploration, and ϵ-greedy exploration was disabled. We employed Prioritized Experience Replay (Schaul et al., 2016) with rank-based prioritization to improve sample efficiency. Regarding the hyperparameters, we set $\alpha = 0.6$ and the initial $\beta = 0.4$, annealed β to 1.0 over 10,000 steps, used a replay buffer capacity of 100,000 transitions, and started learning after 1,000 steps. The policy network had one hidden layer with 64 units followed by a Rectified Linear Unit activation. Training used discount factor $\gamma = 0.99$, the Adam optimizer, and batch size 256. The learning rate was linearly decayed from 10^{-3} to 10^{-4} over episodes 100 to 200 and then to 10^{-5} over episodes 300 to 400.

[1] S-ASR model page on Hugging Face.

5.3 Compared Conditions

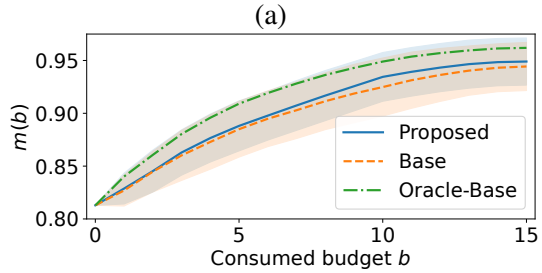
The compared conditions were designed to isolate the effects of S-ASR confidence and segmentation confidence of alternative hypotheses in the policy state. The main comparison included Base, Proposed, and Oracle-Base. All three conditions used the same RL algorithm and differed only in the information provided to the policy state.

Base, following Waki et al. (2025), uses $\mathbf{s}_t^{(\text{base})} = (\mathbf{u}_t, \rho_t)$ as the state, where $\mathbf{u}_t = (\kappa_t^{(1,1)}, \kappa_t^{(1,2)}, \kappa_t^{(1,3)})$ represents the top-3 boundary probabilities of the 1-best hypothesis. Proposed extends Base with both $\mathbf{s}_t^{(\text{asr})}$ and $\mathbf{s}_t^{(\text{alt})}$, where $\mathbf{s}_t^{(\text{alt})} = (\kappa_t^{(2,1)}, \kappa_t^{(2,2)}, \kappa_t^{(2,3)})$ was defined as the top-3 boundary probabilities of the second-best hypothesis. For $\mathbf{s}_t^{(\text{asr})}$, we set the number of S-ASR hypotheses $N^{(\text{asr})}$ to 5 and computed entropy over these hypotheses. Oracle-Base uses $\mathbf{s}_t^{(\text{base})}$ but replaces $\hat{\mathbf{y}}_t$ with the reference syllable sequence $\mathbf{y}_t^*$ when computing $\mathbf{u}_t$, providing an upper bound on performance achievable without S-ASR errors.

The ablation conditions were designed to isolate the contributions of S-ASR confidence and segmentation confidence of alternative hypotheses. Three variants extend Base with $\mathbf{s}_t^{(\text{asr})}$ alone: Base+ASR(CTC) uses only the CTC branch, Base+ASR(Dec) uses only the attention-based decoder, and Base+ASR(CTC+Dec) combines both branches. Base+Seg adds $\mathbf{s}_t^{(\text{alt})}$ to $\mathbf{s}_t^{(\text{base})}$ alone, isolating its effect without S-ASR confidence.

In addition to the RL-based conditions, we included a non-RL rule-based baseline, Rule-Seg, that selects only between *query* and *skip*. Let $\hat{\mathbf{c}}_t$ denote the 1-best segmentation of the 1-best S-ASR hypothesis $\hat{\mathbf{y}}_t$. This baseline uses the posterior probability $P(\hat{\mathbf{c}}_t|\hat{\mathbf{y}}_t, \phi_t)$ as a single confidence score. The threshold was selected by grid search on the training split to maximize the AUC metric defined in Section 5.4. At test time, Rule-Seg selects *query* when this score falls below the selected threshold, and otherwise selects *skip*.

(a)



(b)

Method	Err. req. (↓)	AUC (↑)
Base	8.31 ± 1.34	13.47 ± 0.18
Proposed	$\mathbf{6.97 \pm 1.14}$	$\mathbf{13.54 \pm 0.11}$
Oracle-Base	–	13.78 ± 0.12

Figure 5: Main results on the test set. (a) Exact boundary match rate $m(b)$ over the consumed budget. (b) The number of clarification requests on utterances affected by S-ASR errors and AUC. Values are mean $\pm$ std. over 400 trials; shaded regions in (a) show std.

5.4 Evaluation Metrics

We evaluated the proposed method using two complementary metrics. The first metric was the number of clarification requests on utterances affected by S-ASR errors. The second metric was the area under the learning curve (AUC) of word segmentation over the clarification budget, used as a measure of acquisition efficiency. Although the policy was trained with a reward based on micro-F1, this learning curve is based on the exact boundary match rate. This rate provides a stricter measure of segmentation quality.

The first metric counts clarification requests issued for utterances affected by S-ASR errors. For a test episode of length T, this count is computed over Stream Test utterances as

$$\sum_{t=1}^{T} \mathbb{1}[a_t = \textit{query} \wedge \hat{\mathbf{y}}_t \neq \mathbf{y}_t^*].$$

A lower count is generally preferable, although some such queries may remain necessary because S-ASR errors often occur with unknown words. This metric is interpreted as a proxy for unnecessary clarification requests rather than a direct measure of user burden, because no interactive user evaluation was conducted. To contextualize this count, we also report the total number of clarification requests, which indicates how aggressively each policy queries overall.

The second metric, AUC, measures how word segmentation performance improves as the clarification budget is consumed. Let $\mathcal{E}_{\text{test}}$ denote Eval Test, and let $\phi^{(b)}$ denote the word segmentation model after consuming b units of budget. For the i-th utterance $(\mathbf{y}_i^*, \mathbf{c}_i^*) \in \mathcal{E}_{\text{test}}$, we apply the word segmentation model to the reference syllable sequence $\mathbf{y}_i^*$ and predict the boundary sequence by $\hat{\mathbf{c}}_i^{(b)} = \text{argmax}_{\mathbf{c}} P(\mathbf{c}|\mathbf{y}_i^*, \phi^{(b)})$.

Let b_{end} denote the number of budget units actually consumed in the episode. For $b \leq b_{\text{end}}$, the exact boundary match rate is defined as

$$m(b) = \frac{1}{|\mathcal{E}_{\text{test}}|} \sum_{(\mathbf{y}_i^*, \mathbf{c}_i^*) \in \mathcal{E}_{\text{test}}} \mathbb{1}\left[\hat{\mathbf{c}}_i^{(b)} = \mathbf{c}_i^*\right], \quad (10)$$

where the summation is over all utterances in Eval Test. If the policy stops issuing query actions before exhausting the full budget, the word segmentation model is no longer updated, and we set $m(b) = m(b_{\text{end}})$ for $b > b_{\text{end}}$. We computed AUC over the full budget range by the trapezoidal rule.

6 Results

To examine whether the proposed method works as intended, this section reports the results of the main comparison and the component analysis. We first compare the main conditions and then analyze the contribution of each proposed component. Unless otherwise noted, values in the figures and tables are reported as mean $\pm$ std. over 400 evaluation trials to summarize variability over both policy initialization and stream order.

6.1 Comparison with Baseline

Figure 5(b) shows that Proposed reduced the number of clarification requests on utterances affected by S-ASR errors relative to Base. The mean number decreased from 8.31 ± 1.34 for Base to 6.97 ± 1.14 for Proposed. This reduction was statistically significant, with a 95% confidence interval of $[0.62, 2.11]$ and a p-value of 0.002.

Figure 5 also shows that Proposed achieved a numerically higher segmentation performance than Base. Figure 5(b) shows that the mean AUC increased from 13.47 ± 0.18 for Base to 13.54 ± 0.11 for Proposed. Figure 5(a) shows the same pattern over the budget range, as the mean curve of Proposed first rises above that of Base at around $b = 3$

Table 2: Ablation results on the test set (mean ± std. over 400 trials). Err. req. denotes clarification requests on utterances affected by S-ASR errors. Total req. denotes the total number of clarification requests.

Method	Err. req. (↓)	Total req.	AUC (↑)
Base	8.31 ± 1.34	13.38 ± 1.66	13.47 ± 0.18
Base+ASR(CTC)	7.04 ± 2.03	12.50 ± 3.49	13.46 ± 0.31
Base+ASR(Dec)	6.82 ± 2.51	12.06 ± 4.16	13.45 ± 0.37
Base+ASR(CTC+Dec)	7.84 ± 1.54	13.34 ± 1.98	13.51 ± 0.22
Base+Seg	8.58 ± 1.02	13.82 ± 1.14	13.50 ± 0.08
Rule-Seg	10.60 ± 1.40	15.00 ± 0.00	13.58 ± 0.27
Proposed	6.97 ± 1.14	12.75 ± 1.78	13.54 ± 0.11

and remains above it through $b = 15$. However, this AUC difference was not statistically significant. The 95% confidence interval ranged from -0.02 to 0.16, and the corresponding p-value was 0.154.

These results indicate that Proposed reduced clarification requests on utterances affected by S-ASR errors without a clear decrease in acquisition performance. Oracle-Base achieved a higher mean AUC than both non-oracle conditions, 13.78 ± 0.12, and its curve in Figure 5(a) remained above the others across the budget range. This remaining gap indicates that S-ASR errors still limit performance under speech input.

6.2 Contribution of Each Component

Table 2 shows that S-ASR confidence was more directly related to reducing clarification requests on utterances affected by S-ASR errors. All variants that incorporated S-ASR confidence reduced the mean number of such clarification requests relative to Base, whereas Base+Seg did not. The lowest mean was obtained by Base+ASR(Dec), which achieved 6.82 ± 2.51, whereas Proposed achieved 6.97 ± 1.14 with a substantially smaller standard deviation.

Base+Seg showed a different pattern from the variants that incorporated S-ASR confidence. Its mean AUC increased from 13.47 ± 0.18 for Base to 13.50 ± 0.08, but the mean number of clarification requests on utterances affected by S-ASR errors also increased from 8.31 ± 1.34 to 8.58 ± 1.02. Taken together, these results suggest that segmentation confidence of alternative hypotheses mainly supported acquisition performance, whereas S-ASR confidence was more directly related to reducing clarification requests on utterances affected by S-ASR errors.

Among the RL-based non-oracle methods, Proposed achieved the highest mean AUC at 13.54 ± 0.11. The non-RL threshold baseline showed a different trade-off from the RL-based methods. Its mean AUC was slightly higher than that of Proposed, but it issued more clarification requests on utterances affected by S-ASR errors than both Proposed and Base.

Overall, Proposed provided the best balance between the two metrics in the RL-based non-oracle setting. It issued fewer clarification requests on utterances affected by S-ASR errors than Base, while maintaining the highest mean AUC among the RL-based non-oracle methods.

7 Discussion

7.1 Stability and Comparison with Rule-Seg

Although Base+ASR(Dec) achieved the lowest mean number of clarification requests on utterances affected by S-ASR errors, its standard deviations were substantially larger for Err. req., Total req., and AUC. This indicates that using decoder-based ASR confidence alone sometimes led to overly conservative or unstable querying behavior depending on the policy initialization and stream order. In contrast, Proposed achieved a similarly low Err. req. with smaller standard deviations and no clear loss in acquisition performance. These results suggest that combining S-ASR confidence with segmentation confidence over alternative hypotheses made the policy less sensitive to variation across training seeds and input orders.

The Rule-Seg baseline shows that this simple rule-based strategy was not sufficient to suppress unnecessary clarification requests. It issued 15 clarification requests in every episode, including 10.60 ± 1.40 requests on utterances affected by S-ASR errors, exceeding both Proposed and Base. Although its mean AUC was slightly higher than that of Proposed, this result indicates that Rule-Seg improved acquisition performance by querying aggressively, but did not control unnecessary clarification.

Table 3: *Self-learning* behavior on the test set. S-ASR err. sel. rate denotes the proportion of utterances with S-ASR errors selected for *self-learning*, and No. of *self-learning* denotes the total number of *self-learning* actions per episode.

Method	S-ASR err. sel. rate (↓)	No. of *self-learning*
Base	0.316 ± 0.052	90.1 ± 60.1
Proposed	0.268 ± 0.081	69.5 ± 48.5

7.2 Effect on Self-learning

The proposed method also reduced the proportion of utterances containing S-ASR errors that were selected for *self-learning*, as shown in Table 3. This rate decreased from 0.316 ± 0.052 for Base to 0.268 ± 0.081 for Proposed, and the total number of *self-learning* actions decreased. These results suggest that the S-ASR confidence features helped the policy identify utterances likely to contain S-ASR errors and select *skip* rather than *self-learning*. This may have reduced updates based on noisy pseudo-labels, suggesting that the benefit of the proposed method was not limited to *query* selection.

7.3 Per-word Analysis

Table 4 relates word-level clarification behavior to S-ASR reliability. Clarification rate denotes the proportion of evaluation episodes in which the policy selected *query* at least once for an utterance containing the target word. Because all words in the table are target unknown words, a higher rate indicates that oracle supervision for that word was obtained in more episodes. However, this rate should be interpreted with the S-ASR match rate, since each *query* consumes the limited budget and low match rates indicate less reliable S-ASR evidence.

Words with more reliable S-ASR outputs tended to have higher clarification rates under Proposed. For Krofler, whose 1-best S-ASR match rate was 0.97, the clarification rate increased from 0.85 to 0.97, and for Eisbein, whose 1-best S-ASR match rate was 0.81, it increased from 0.71 to 0.79. These cases are consistent with the policy selecting *query* when the S-ASR output sequence is likely to support useful supervision for word boundaries. In contrast, Dal Bhat and Choucroute had lower 1-best S-ASR match rates of 0.63 and 0.47, respectively. The clarification rate decreased from 0.59 to 0.26 for Dal Bhat and from 0.61 to 0.44 for Choucroute. These decreases suggest that the policy became more conservative for words whose S-ASR outputs were less reliable, although this may also reduce oracle supervision for such words.

Table 4: S-ASR output match rates and clarification rates for selected unknown words on the test set. Match rate denotes the proportion of utterances containing the word in which the 1-best S-ASR hypothesis matched the reference transcription. Clarification rate denotes the proportion of episodes in which the policy selected *query* at least once for the word.

Word	Match rate	Base	Proposed
Maritozzo	0.03	1.00 ± 0.01	0.98 ± 0.07
Semifreddo	0.23	0.41 ± 0.24	0.55 ± 0.38
Choucroute	0.47	0.61 ± 0.24	0.44 ± 0.32
Dal Bhat	0.63	0.59 ± 0.28	0.26 ± 0.27
Snert	0.73	0.86 ± 0.22	0.94 ± 0.19
Eisbein	0.81	0.71 ± 0.20	0.79 ± 0.28
Krofler	0.97	0.85 ± 0.10	0.97 ± 0.04

However, this tendency did not hold uniformly. Semifreddo had a low 1-best S-ASR match rate of 0.23, yet its clarification rate increased from 0.41 to 0.55. We speculate that its correct syllable sequence appeared more frequently in non-1-best hypotheses. In such cases, segmentation confidence of alternative hypotheses may help preserve clarification even when the 1-best hypothesis is unreliable.

Taken together, this per-word analysis suggests that the proposed method did not simply reduce clarification uniformly. It tended to maintain higher clarification rates for some words with reliable S-ASR outputs, while becoming more conservative for some words with less reliable outputs.

8 Conclusion

We proposed a method for suppressing unnecessary clarification requests in unknown word acquisition. The method incorporates S-ASR confidence into clarification request selection in a stream-based active learning framework. It further uses segmentation confidence over alternative hypotheses to account for cases where useful information remains beyond the 1-best S-ASR hypothesis.

Experiments using pre-recorded speech data showed that S-ASR confidence reduced clarification requests on utterances affected by S-ASR errors. The total number of clarification requests also decreased, and AUC was numerically but not significantly higher than that of Base. Combining S-ASR confidence with alternative-hypothesis segmentation confidence provided the best balance between the two metrics.

Limitations

To evaluate clarification request selection itself under speech input, we adopted a simplified formulation that isolates this decision from several other factors in dialogue. The results should therefore be interpreted with the following limitations in mind.

First, our evaluation was conducted in a simulated dialogue environment using pre-recorded speech. In real dialogue settings, the way clarification requests are phrased may affect user responses, and the information obtained may vary accordingly. In addition, when *query* is selected, the oracle provides both the correct syllable sequence and word boundaries. This supervision is stronger than what would typically be available in real dialogue. These aspects of clarification and dialogue are therefore not explicitly modeled in the present formulation.

Second, each episode is based on a short finite input stream, so the stream may end before the budget is exhausted. As a result, the setting may not fully capture the difficulty of allocating clarification requests over longer interactions. Issuing clarification requests aggressively may also be less strongly penalized than in longer dialogues.

Third, the dataset is limited in size and variety of unknown words, which may limit generalization to unseen S-ASR error patterns. In particular, clarification requests were sometimes directed at utterances containing only recognition errors rather than unknown words. This occurred especially in short utterances containing weak sounds, such as long vowels, geminate consonants, and nasal sounds.

Finally, the acquired unknown words are not reflected back into the S-ASR model. A policy that accounts for changes in S-ASR performance as unknown words are acquired would be needed to fully exploit the long-term benefit of unknown word acquisition.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and valuable feedback.

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K28147 and JP22H00536, and JST Moonshot R&D Grant Number JPMJMS2011, Japan.

References

Les Atlas, David Cohn, and Richard Ladner. 1989. Training connectionist networks with queries and selective sampling. In *Proc. NeurIPS*, pages 566–573.

Ekaterina Egorova, Hari Krishna Vydana, Lukáš Burget, and Jan Černocký. 2021. Out-of-vocabulary words detection with attention and CTC alignments in an end-to-end ASR system. In *Proc. INTERSPEECH*, pages 2901–2905.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proc. EMNLP*, pages 595–605.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. 2018. Noisy networks for exploration. In *Proc. ICLR*.

Hossein Hadian and Hossein Sameti. 2014. Active learning in noisy conditions for spoken language understanding. In *Proc. COLING*, pages 1081–1090.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proc. ACL*, pages 3667–3684.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proc. SIGDIAL*, pages 292–299.

Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Communication*, 45(4):455–470.

Kazunori Komatani and Mikio Nakano. 2020. User impressions of questions to acquire lexical knowledge. In *Proc. SIGDIAL*, pages 147–156.

Stefan Kombrink, Lukás Burget, Pavel Matejka, Martin Karafiát, and Hynek Hermansky. 2009. Posterior-based out of vocabulary word detection in telephone speech. In *Proc. INTERSPEECH*, pages 80–83.

Rishabh Kumar, Sabyasachi Ghosh, and Ganesh Ramakrishnan. 2024. Beyond common words: Enhancing ASR cross-lingual proper noun recognition using large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 6821–6828.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proc. ICML WREPL*.

David D. Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. *ACM SIGIR Forum*, 29(2):13–19.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proc. LREC*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc. ACL-IJCNLP*, pages 100–108.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proc. SIGDIAL*, pages 50–59.

Carolina Parada, Mark Dredze, Abhinav Sethy, and Ariya Rastrow. 2011. Learning sub-word units for open vocabulary speech recognition. In *Proc. ACL-HLT*, pages 712–721.

Matthew Purver. 2002. Processing unknown words in a dialogue system. In *Proc. SIGDIAL*, pages 174–183.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In *Proc. ICLR*.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison.

Gabriel Skantze. 2007. Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds. In *Proc. SIGDIAL*, pages 206–210.

Ryu Takeda and Kazunori Komatani. 2025. Reducing orthographic dependency on paired data by probabilistic integration via syllabogram for Japanese dialogue speech recognition. In *Proc. APSIPA ASC*, pages 549–554.

Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double Q-learning. In *Proc. AAAI*, pages 2094–2100.

Issei Waki, Ryu Takeda, and Kazunori Komatani. 2025. Learning to ask efficiently in dialogue: Reinforcement learning extensions for stream-based active learning. In *Proc. SIGDIAL*, pages 431–440.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proc. INTERSPEECH*, pages 2207–2211.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE JSTSP*, 11(8):1240–1253.

Ian Williams, Anjuli Kannan, Petar S. Aleksic, David Rybach, and Tara N. Sainath. 2018. Contextual speech recognition in end-to-end neural network systems using beam search. In *Proc. INTERSPEECH*, pages 2227–2231.

Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.