Silent HMMs: Generalized Representation of Hidden Semi-Markov Models and Hierarchical HMMs

Kei Wakabayashi
1-2 Kasuga, Tsukuba, Ibaraki, Japan
University of Tsukuba
kwakaba@slis.tsukuba.ac.jp

Abstract

Modeling sequence data using probabilistic finite state machines (PFSMs) is a technique that analyzes the underlying dynamics in sequences of symbols. Hidden semi-Markov models (HSMMs) and hierarchical hidden Markov models (HHMMs) are PFSMs that have been successfully applied to a wide variety of applications by extending HMMs to make the extracted patterns easier to interpret. However, these models are independently developed with their own training algorithm, so that we cannot combine multiple kinds of structures to build a PFSM for a specific application. In this paper, we prove that silent hidden Markov models (silent HMMs) are flexible models that have more expressive power than HSMMs and HHMMs. Silent HMMs are HMMs that contain silent states, which do not emit any observations. We show that we can obtain silent HMM equivalent to given HSMMs and HHMMs. We believe that these results form a firm foundation to use silent HMMs as a unified representation for PFSM modeling.

1 Introduction

Probabilistic finite state machines (PFSMs) are widely used for modeling non-deterministic behaviors in languages (Wang and Manning, 2012). One of the powerful applications of PFSMs is automatic (unsupervised) induction of language patterns (Stratos et al., 2016). The automatic induction of finite state models can potentially impact the direction that research takes on finite state machines, which have been applied to natural language processing such as morphological modeling (Ehsani et al., 2017), word transduction between different languages (Sharma and Singh, 2017), dialog action (Torres, 2013), etc.

Hidden Markov models (HMMs) are the simplest and most well-known probabilistic finite state machines. However, the unsupervised training of HMMs usually does not produce good finite state machines like the ones crafted by human experts because of the complexity of reconstructing language patterns from a finite number of observations. Human experts can build finite state machines that are comprehensible because they have intuition about the latent structure of languages.

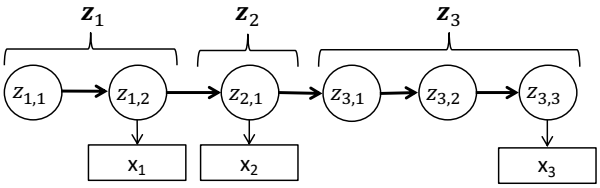

Figure 1: A hidden state sequence of silent HMM. $\mathbf{z}_t$ corresponds to multiple hidden states that constitute a silent Markov chain.

This discussion suggests that we need to incorporate prior knowledge into the model structure of HMMs, which is a basic idea that pervades the recent methods of automatic induction of language patterns (Stratos et al., 2016; Jin et al., 2018). Several kinds of PFSMs, such as hidden semi-Markov models (HSMMs) (Moore and Savic, 2004; Yu, 2010) and hierarchical hidden Markov models (Fine et al., 1998; Wakabayashi and Miura, 2012), reflect several different additional structural assumptions. Each model comes with a specialized training algorithm that has to be implemented separately. This requirement prevents us from trying several models; more importantly, we cannot easily combine multiple assumptions that are implemented in different PFSMs. To move the research of automatic finite state machine induction forward, we need to develop a more flexible way to incorporate our prior knowledge into the PFSMs.

Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing, pages 98–107
Dresden, Germany, September 23-25, 2019.

In this paper, we propose silent hidden Markov models (silent HMMs) as a generalized representation of other PFSMs that at least can express the structure that is assumed in HSMMs and HHMMs. A silent HMM is an HMM that contains silent states, which do not emit any observations. We prove that the expressive power of silent HMMs is better than HSMMs and HHMMs, and we propose a method that obtains a silent HMM that is equivalent to an HSMM and an HHMM. This result indicates that we can combine and/or customize the structural assumptions of HSMMs and HHMMs in the unified framework of silent HMMs, potentially leading us to more precise and flexible automatic induction of finite state machines.

The rest of the paper is organized as follows. In Section 2, we dfine silent HMMs. In Sections 3 and 4, we detail the HSMMs and HHMMs respectively and prove the expressivity of silent HMMs is better than these models. In Section 5, we discuss an inference algorithm of silent HMMs. In Section 6, we conclude the discussion and mention future work.

2 Silent HMMs

The concept of the silent states, also known as "null emission" in HMMs, has been used in speech recognition (Bahl et al., 1983; Rabiner, 1989) and DNA modeling in bioinformatics (Krogh et al., 1994; Eddy, 1998) to express optional sounds or letters in sequences that are implicitly dropped from observations. Recently, Wakabayashi (2018) applied a silent HMM to natural language sentences to extract phrase structures in an unsupervised manner. However, surprisingly few descriptions exist in literature that define silent HMMs in a formal way. In this section, we formally define silent HMMs.

Let $\mathbf{x}_{1:T} = x_1, \ldots, x_T$ be the sequence of observations and $\mathcal{X}$ be the domain of each observation ($x_t \in \mathcal{X}$). We denote the states that correspond to each observation x_t by $\mathbf{z}_t = z_{t,1}, \ldots, z_{t,|\mathbf{z}_t|}$. In silent HMMs, $\mathbf{z}_t$ can be a series of states that contain multiple silent states that precede a normal state producing x_t. Figure 1 illustrates the relationship between x_t and $\mathbf{z}_t$. $z_{t,1}, \ldots, z_{t,|\mathbf{z}_t|-1}$ are all silent states and $z_{t,|\mathbf{z}_t|}$ is a normal state.

A silent HMM is defined by a tuple $(\mathcal{X}, Q, C, R, \boldsymbol{\pi}, A, \Theta)$. Q is a finite set of states. Silence assignment $C : Q \to \{0, 1\}$ is a mapping that designates silent states. Each state is either a silent state or a normal state. $C(q) = 1$ indicates that the state $q \in Q$ is a silent state and $C(q) = 0$ means q is a normal state. The set of normal states is denoted by $Q_n = \{q \in Q | C(q) = 1\}$ and the set of silent states is denoted by $Q_s = \{q \in Q | C(q) = 0\}$.

R is a predicate that defines a transition topology. The domain of R is $Q \times Q$. If $R(q_1, q_2)$ is true, the transition from q_1 to q_2 is allowed. In the rest of the paper, we also use $q_1 \xrightarrow{R} q_2$ to indicate $R(q_1, q_2)$ is true.

The joint likelihood of $\mathbf{x}_{1:T}$ and $\mathbf{z}_{1:T}$ is described as follows.

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(x_t | z_{t,|\mathbf{z}_t|}). \quad (1)$$

Since $\mathbf{z}_t = z_{t,1}, \ldots, z_{t,|\mathbf{z}_t|}$, $p(\mathbf{z}_t | \mathbf{z}_{t-1})$ is the joint probability given as below.

$$p(\mathbf{z}_t | \mathbf{z}_{t-1}) = p(z_{t,1} | z_{t-1,|\mathbf{z}_{t-1}|}) \prod_{\tau=1}^{|\mathbf{z}_t|} p(z_{t,\tau} | z_{t,\tau-1}). \quad (2)$$

When $t = 1$, the first term in Eq (2), $p(z_{1,1} | z_{0,|\mathbf{z}_0|})$, is defined as an initial state probability. $\boldsymbol{\pi}$ is a $|Q|$ dimensional vector that represents the initial state distribution. A is a $|Q| \times |Q|$ matrix of which A_{q_1,q_2} indicates the transition probability from q_1 to q_2; e.g., $p(z_{t,\tau} = q_2 | z_{t,\tau-1} = q_1) = A_{q_1,q_2}$. For q_1, q_2 such that $R(q_1, q_2)$ is false, A_{q_1,q_2} is restricted to being zero. $\Theta = \{\boldsymbol{\theta}_q\}_{q \in Q_n}$ is parameters of the emission distribution $p(x_t | z_{t,|\mathbf{z}_t|})$ for each normal state.

$\mathcal{X}, Q, C, R$ are meta-parameters of silent HMMs, which are not trainable from data. These meta-parameters reflect prior knowledge of a structure of sequence data. In the following sections, we show that there is a set of meta-parameters that makes the likelihood function of silent HMM identical to the likelihood function of HSMMs and HHMMs.

3 Hidden Semi-Markov Models

3.1 Model Definition

A hidden semi-Markov model (HSMM) is a probabilistic automaton that allows a state to emit multiple observations. Figure 2 illustrates a hidden state sequence of HSMMs. HSMMs explicitly consider a probabilistic distribution of the duration. For example, in Figure 2, the duration of the

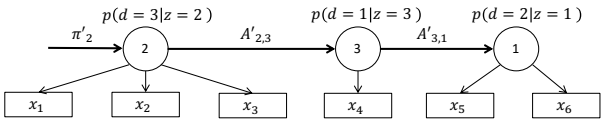

Figure 2: Example of a hidden state sequence of HSMM

first state $z_1 = 2$ is $d_1 = 3$, meaning the state keeps emitting three observations (x_1, x_2, and x_3) following the i.i.d. distribution of $p(x|z = 2)$. The duration of each state is stochastically determined depending on the state. While multiple ways to define the distribution of duration $p(d|z)$ have been proposed (Yu, 2010), we use categorical distributions with D possible classes to represent $p(d|z)$ where $D \in \mathbb{N}$ is the maximum duration.

An HSMM is defined by a tuple $(\mathcal{X}, Q', D, \boldsymbol{\pi}', A', \Phi, \Theta')$. $\mathcal{X}$ is a domain of observations, Q' is a set of states, and $D \in \mathbb{N}$ is a maximum duration. A' is a transition probability matrix where the transition probability from the state i to j is $A'_{i,j}$. $\boldsymbol{\pi}'$ is an initial probability vector where the initialization probability of the state i is π'_i. $\Phi = \{\boldsymbol{\phi}_i\}_{i \in Q'}$ is a set of parameters of duration distribution where $p(d|z) = \phi_{z,d}$. Θ' is a set of parameters for the emission distributions.

Let $\mathbf{x} = x_1, \ldots, x_T$ be a sequence of observations, $\mathbf{z} = z_1, \ldots, z_n$ be a sequence of hidden states, and $\mathbf{d} = d_1, \ldots, d_n$ be a sequence of duration variables. We use n to indicate the length of the hidden state sequence, which is not necessarily equal to T. Instead, $\sum_{\tau=1}^{n} d_\tau$ must be equal to T. The likelihood function of an HSMM is defined as follows;

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:n}, \mathbf{d}_{1:n}) = \pi'_{z_1} \prod_{\tau=2}^{n} A'_{z_{\tau-1}, z_\tau} \prod_{\tau=1}^{n} \phi_{z_\tau, d_\tau} \prod_{t=1}^{T} p(x_t|\boldsymbol{\theta}_{z_{c(t)}}), \quad (3)$$

where $c(t)$ is a function that returns the index of the hidden state that corresponds to the observation x_t.

$\mathcal{X}, Q', D$ are meta-parameters of HSMMs that are not trainable from data. In the next section, we demonstrate how an HSMM that has meta-parameters $\mathcal{X}, Q', D$ can be equivalently represented as a silent HMM.

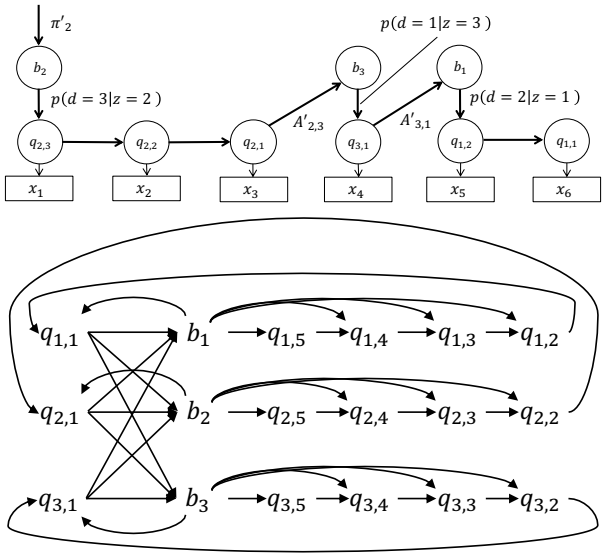

Figure 3: **(Top)** Hidden state sequence of the silent HMM that is equivalent to the state sequence of the HSMM in Figure 2 **(Bottom)** State transition diagram

3.2 Expressivity of HSMMs and Silent HMMs

Given an HSMM that has meta-parameters $\mathcal{X}, Q', D$, we can obtain an equivalent silent HMM that has meta-parameters $(\mathcal{X}, Q, C, R)$. Figure 3 depicts the representation of the transition dynamics of an HSMM by a silent HMM. The duration of each state is represented explicitly by a transition throughout "countdown states." A countdown state $q_{i,d}$ only changes to $q_{i,d-1}$. The state of the last count $q_{i,1}$ changes to b_j, which indicates a silent state that represents the beginning of the state j in HSMM. The transition probabilities from b_j correspond to the duration probability $p(d|z = j)$.

Here, we explain the proposed mapping from a tuple of meta-parameters (Q', D) of HSMMs to a tuple of meta-parameters (Q, C, R) of the equivalent silent HMMs. First, Q is constructed as a union of a set of countdown states Q_c and a set of beginning states Q_b. We define Q_c and Q_b as follows.

$$Q_c = \{q_{i,d}\}_{i \in Q', 1 \leq d \leq D}$$
$$Q_b = \{b_i\}_{i \in Q'}.$$

The whole set of states in the silent HMM is defined as $Q = Q_b \cup Q_c$. The elements in Q_c are nor-

mal states and the elements in Q_b are silent states.

$$C(q) = \begin{cases} 0 & q \in Q_c \\ 1 & q \in Q_b. \end{cases}$$

The transition topology R is defined as depicted in Figure 3 **(Bottom)**. More formally:

$$\forall i, d(b_i \xrightarrow{R} q_{i,d})$$
$$\forall i, j(q_{i,1} \xrightarrow{R} b_j)$$
$$\forall i, d > 1(q_{i,d} \xrightarrow{R} q_{i,d-1}).$$

To show the equivalency of the given HSMM and the obtained silent HMM, we also specify a surjective mapping from the distributions in the silent HMM parameterized by $(\boldsymbol{\pi}, A, \Theta)$ to the distributions in the HSMM parameterized by $(\boldsymbol{\pi}', A', \Phi, \Theta')$.

- The D-dimensional categorical distribution in the silent HMM for transition from the state $b_i \in Q_b$ parameterized by $\mathbf{A}_{b_i}$ is mapped into the categorical distribution in the HSMM for the duration of the state i parameterized by $\boldsymbol{\phi}_i$.
- The $|Q|$-dimensional categorical distribution in the silent HMM for transition from the state $q_{i,1}$ parameterized by $\mathbf{A}_{q_{i,1}}$ is mapped into the categorical distribution in the HSMM for the transition from the state i parameterized by $\mathbf{A}'_i$.
- The emission distribution of the state $q_{i,d}$ in the silent HMM is mapped into the emission distribution of the state i in the HSMM for any $d \in D$.

Note that the destination of the transition from $q_{i,d}$ is only $q_{i,d-1}$ for any $d > 1$; therefore, the transition probability from $q_{i,d}$ to $q_{i,d-1}$ is always one.

Lemma 1. *The likelihood function of the silent HMM constructed in the way described above is equivalent to the likelihood function of the given HSMM.*

This lemma can be proved straightforwardly by mapping random variables as shown in Figure 3 **(Top)** and putting mapped parameters in Eqs. (1) and (2).

Theorem 1. *The expressivity of silent HMMs is better than the expressivity of HSMMs. In other words, the mapping from an HSMM to a silent HMM that makes the likelihood function equivalent is injective and not surjective.*

Proof of being injective is easy: We can confirm that different HSMMs have different likelihood functions. If the mapping is not injective, two HSMMs with different likelihood functions are mapped into the same silent HMM. This contradicts the Lemma 1. Being not surjective is obvious; for silent HMMs, we can set different meta-parameters from ones explained above.

This result is useful in practice because we can use an implementation of the silent HMMs when we want to use HSMMs. We do not need to implement the training algorithm and the Viterbi algorithm just for HSMMs.

4 Hierarchical HMMs

4.1 Model Definition

A hierarchical HMM (HHMM) is a probabilistic automaton that simulates multiple Markov chains that have a hierarchical relationship. Figure 4 illustrates the dynamics of an HHMM that has three hierarchy levels. A hidden state sequence is in each level. Each state sequence can be terminated probabilistically when the sequence reaches a special End state. The state at level d is allowed to change to another state at time step t only when the state sequences at all the lower levels are terminated. If a state sequence at level d is terminated at time step t, a state sequence is initialized again at the next time step $t + 1$. Only the states at the bottom level emit the observation. The probabilistic distribution of state transitions and observation emissions depend on the combination of the states at all the upper levels[1]. For example, the state transition from the bottom state 2 to state 1 at time step $t = 1$ in Figure 4 depends on all the upper states, namely, state 2 at the top level $d = 1$ and state 1 at the middle level $d = 2$.

An HHMM is defined by a tuple $(\mathcal{X}, N, L, \boldsymbol{\pi}'', A'', \Theta'')$ where N is the number of states in each Markov chain and L is the number of levels. The HHMM in Figure 4 has $N = 2$ and $L = 3$. When the states at level 1 to $d - 1$ are $\mathbf{k} = (k_1, \ldots, k_{d-1})$, the state transition probability from the state i to the state j at level d is denoted by $A''^{\mathbf{k}}_{i,j}$ and the state initialization probability of the state i at level d is represented by $\pi''^{\mathbf{k}}_i$. We consider a special symbol End as

[1] Another version of HHMMs shares the probabilistic distributions among the states that have different upper states (Bui et al., 2004). Although we could extend the discussion in this section to adapt to this version, we do not go into detail due to the length limitations.

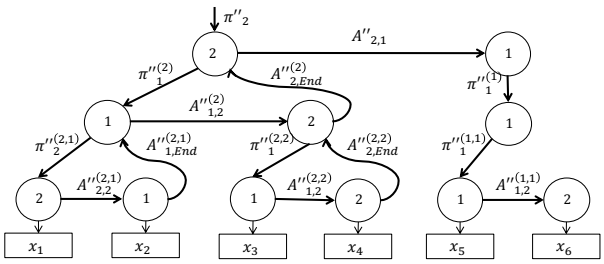

Figure 4: Example of a hidden state sequence of hierarchical HMM

another destination of state transition, which triggers the termination of the state sequence at that level. The transition parameters A'' satisfy the following condition for all i and $\mathbf{k}$.

$$\sum_{j \in \mathbb{N}_{\leq N} \cup \{End\}} A''^{\mathbf{k}}_{i,j} = 1,$$

where $\mathbb{N}_{\leq N}$ is a set of natural numbers that are less than or equal to N (representing the set of states on the Markov chain at that level).

The formulation of the likelihood function of HHMMs is complicated because the length of state sequence is different at each level. To simplify the situation, we apply a variable conversion proposed by (Murphy and Paskin, 2002) that explicitly considers random variables that represent the state at each time step for all levels. Formally, we define $\mathbf{z}^d = z_1^d, \ldots, z_T^d$ as a sequence of hidden states at level d. For example, the state sequences in Figure 4 are represented as $\mathbf{z}^1 = 2,2,2,2,1,1$, $\mathbf{z}^2 = 1,1,2,2,1,1$, and $\mathbf{z}^3 = 2,1,1,2,1,2$. We also consider a set of binary auxiliary variables $\{f_t^d\}$ that indicate if the state sequence at level d is terminated at time step t. For example, $\mathbf{f}_1^{1:L} = 0,0,0$, $\mathbf{f}_2^{1:L} = 0,0,1$, $\mathbf{f}_3^{1:L} = 0,0,0$, $\mathbf{f}_4^{1:L} = 0,1,1$. f_t^d has to be 0 whenever $f_t^{d+1} = 0$ because the state does not change at level d if the state sequence at the lower level $d+1$ is not terminated.

Using this representation, we can formulate the likelihood function of HHMMs as Eq. (4). For simplicity, we define $f_t^{L+1} = 1$. The first factor (a) corresponds to an initialization of the state sequences at time step $t = 1$. The second factor (b) indicates a product of termination probabilities, a transition probability, and initialization probabilities for each time step. For example, consider the case of $t = 4$ for the state sequences in Figure 4. Since $\mathbf{f}_4^{1:L} = 0,1,1$, we calculate the product of two termination probabilities $A''^{(2,2)}_{2,End}, A''^{(2)}_{2,End}$ (for $d = 3$ and $d = 2$), one transition probability $A''_{2,1}$ (for $d = 1$), and two initialization probabilities $\pi''^{(1)}_1, \pi''^{(1,1)}_1$ (for $d = 2$, $d = 3$). The third factor (c) is a product of emission probabilities for all observations.

Since the dynamics of HHMMs are complex, an inference algorithm needs to be reformulated as a specialized algorithm. Several inference methods have been proposed, such as a modified inside-outside algorithm (Fine et al., 1998), an inference based on dynamic Bayesian network (Murphy and Paskin, 2002), a method based on a variable conversion (Wakabayashi and Miura, 2012), etc. The unsupervised training of HHMMs produces finite state machines that reflect hierarchical sequential patterns on letter sequences in natural language text (Fine et al., 1998), musical pitch structure (Weiland et al., 2005), etc.

4.2 Expressivity of HHMMs and Silent HMMs

Given an HHMM that has meta-parameters $\mathcal{X}, N, L$, we show a method of obtaining a silent HMM that has the equivalent likelihood function. First, we represent the combination of the states in a tree structure as shown in Figure 6 because the probabilistic behaviors in HHMMs depend on the combination of states in all the upper levels. We denote the set of nodes in this tree, excluding the special $ROOT$ node by Ω. Let $parent : \Omega \rightarrow \Omega \cup \{ROOT\}$ be a function that maps a node to its parent node. We denote the children of the node $\omega \in \Omega$ by $child(\omega) = \{v | parent(v) = \omega\}$ and the siblings by $sib(\omega) = child(parent(\omega))$. We also denote the set of leaf nodes by $\Omega_l = \{v \in \Omega | child(v) = \phi\}$ and the set of non-leaf nodes by $\Omega_n = \Omega - \Omega_l$.

Figure 7 shows an equivalent representation of the hidden states of the HHMM in Figure 4, which illustrates the basic idea for obtaining a silent HMM that has an identical likelihood function. Each state in the silent HMMs corresponds to a node in Figure 6. A leaf node is represented as a normal state denoted by q, and a non-leaf node is represented as a silent state. Termination of a state sequence is represented by a state transition to a silent state denoted by e at an upper level. A state transition at an upper level is represented by a state transition from a silent state denoted by e to another silent state denoted by b. An initialization of a state at a lower level is represented by a state

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}^{1:L}, \mathbf{f}_{1:T}^{1:L}) =$$
$$\underbrace{\prod_{d=1}^{L} \pi''^{\mathbf{z}_1^{1:d-1}}_{z_1^d}}_{(a)} \underbrace{\prod_{t=1}^{T-1} \prod_{d=1}^{L} \left(A''^{\mathbf{z}_t^{1:d-1}}_{z_t^d, End} \right)^{f_t^d} \left(A''^{\mathbf{z}_t^{1:d-1}}_{z_t^d, z_{t+1}^d} \right)^{f_t^{d+1}(1-f_t^d)} \left(\pi''^{\mathbf{z}_{t+1}^{1:d-1}}_{z_{t+1}^d} \right)^{f_t^d}}_{(b)} \underbrace{\prod_{t=1}^{T} p(x_t | \boldsymbol{\theta}''_{\mathbf{z}_t^{1:L}})}_{(c)} \quad (4)$$

Figure 5: Likelihood function of HHMMs

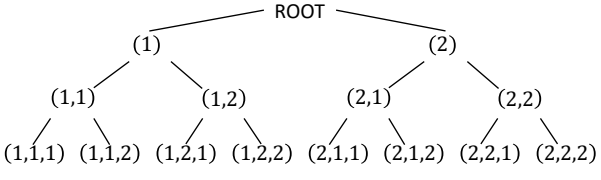

Figure 6: Tree structure that expresses the combination of states in the HHMM with $N = 2, L = 3$.

transition from a silent state denoted by b to a state at a lower level.

We propose a mapping from a tuple of meta-parameters (N, L) of HHMMs to a tuple of meta-parameters (Q, C, R) of silent HMMs to make equivalent likelihood functions. First, we construct the tree structure shown in Figure 6 from N and L and obtain the sets of nodes Ω, Ω_l, and Ω_n. We define a set of production states Q_q, a set of beginning states Q_b, and a set of ending states Q_e as follows:

$$Q_q = \{q_\omega\}_{\omega \in \Omega_l}$$
$$Q_b = \{b_\omega\}_{\omega \in \Omega_n}$$
$$Q_e = \{e_\omega\}_{\omega \in \Omega_n}.$$

The whole set of states in the silent HMM is $Q = Q_q \cup Q_b \cup Q_e$. The elements in Q_b and Q_e are silent states and elements in Q_p are normal states.

$$C(q) = \begin{cases} 1 & q \in Q_b \cup Q_e \\ 0 & q \in Q_q \end{cases}$$

The transition topology R is designed like in Figure 7 **(Bottom)**. More formally:

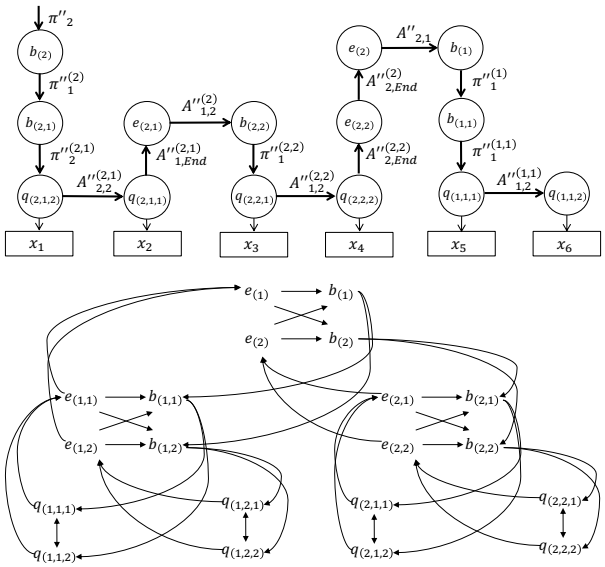

Figure 7: **(Top)** Hidden state sequence of the silent HMM that is equivalent to the state sequence of the hierarchical HMM in Figure 4. **(Bottom)** State transition diagram.

$$\forall \omega \in \Omega_n, \forall \omega' \in child(\omega)(\omega' \in \Omega_n \implies b_\omega \xrightarrow{R} b_{\omega'})$$
$$\forall \omega \in \Omega_n, \forall \omega' \in child(\omega)(\omega' \in \Omega_l \implies b_\omega \xrightarrow{R} q_{\omega'})$$
$$\forall \omega \in \Omega_l, \forall \omega' \in sib(\omega)(q_\omega \xrightarrow{R} q_{\omega'})$$
$$\forall \omega \in \Omega_l(q_\omega \xrightarrow{R} e_{parent(\omega)})$$
$$\forall \omega \in \Omega_n, \forall \omega' \in sib(\omega)(e_\omega \xrightarrow{R} b_{\omega'})$$
$$\forall \omega \in \Omega_n(e_\omega \xrightarrow{R} e_{parent(\omega)}).$$

The transition from b_ω corresponds to the initialization of the lower state sequence. The transition from q_ω or e_ω to $e_{parent(\omega)}$ indicates the termination of the state sequence at that level.

To show the equivalency of the likelihood function, we also specify a mapping from the distributions in the silent HMM parameterized by $\boldsymbol{\pi}, A, \Theta$

to the distributions in the given HHMM parameterized by $\boldsymbol{\pi}'', A'', \Theta''$.

- The N-dimensional categorical distribution in the silent HMM for transition from the state $b_{(i_1,\ldots,i_d)} \in Q_b$ is mapped into the categorical distribution in the HHMM for state initialization parameterized by $\boldsymbol{\pi}''^{(i_1,\ldots,i_d)}$.
- The $(N+1)$-dimensional categorical distribution in the silent HMM for the transition from the state $q_{(i_1,\ldots,i_d)} \in Q_q$ and $e_{(i_1,\ldots,i_d)} \in Q_e$ is mapped into the categorical distribution in the HHMM for the state transition from i_d parameterized by $\mathbf{A}_{i_d}^{''(i_1,\ldots,i_{d-1})}$. The state transition to $e_{(i_1,\ldots,i_{d-1})}$ is mapped into $A_{i_d,End}^{''(i_1,\ldots,i_{d-1})}$.
- The N-dimensional categorical distribution in the silent HMM for initialization probabilities parameterized by $\boldsymbol{\pi}$ is mapped into

Lemma 2. *The likelihood function of the silent HMM constructed in the way described above is equivalent to the likelihood function of the given HHMM.*

Proof. The proof of this lemma is based on a comparison between the likelihood function of the silent state sequence in Eq. (1) (2) and factors $(a), (b), (c)$ in Eq. (4).

Factor (a) For $t = 1$, the length of the silent state sequence is exactly L because the sequence starts from a state in Q_b at the top level and follows links to a state at the next lower level. As we defined above, the distribution in the silent HMM for the transition from the state in Q_b is mapped into the state initialization distribution in the HHMM parameterized by $\boldsymbol{\pi}$. This product is identical to the first factor (a) in Eq. (4).

Factor (b) This factor depends on the values $\mathbf{f}_t$. We can say that $\mathbf{f}_t$ is a variable that indicates the level that holds a state transition. Let $l(\mathbf{f}_t)$ be the level that holds a state transition. For example, when $\mathbf{f}_t^{1:L} = 0, 0, 1$, the level 2 holds a state transition and $l(\mathbf{f}_t) = 2$. Given $\mathbf{f}_t$, the silent state sequence for the time step t is $e_{(z_t^1,\ldots,z_t^{L-1})}$, $\ldots$, $e_{(z_t^1,\ldots,z_t^{l(\mathbf{f}_t)})}$, $b_{(z_{t+1}^1,\ldots,z_{t+1}^{l(\mathbf{f}_t)})}$, $\ldots$, $b_{(z_{t+1}^1,\ldots,z_{t+1}^{L-1})}$, $q_{(z_{t+1}^1,\ldots,z_{t+1}^{L})}$. By putting the mapped parameters into the product of the transition probabilities in this trajectory, we can confirm that the probability is identical to the factor (b) in Eq. (4) for any $\mathbf{f}_t$.

Factor (c) Factor $\prod_{t=1}^{T} p(x_t|q_{(z_t^1,\ldots,z_t^L)})$ in Eq. (1) is identical to the factor (c). □

Theorem 2. *The expressivity of silent HMMs is better than the expressivity of HHMMs.*

The theorem can be proved in the same way as Theorem 1. We emphasize again that this result is useful because we can use an implementation of the silent HMMs when we want to use HHMMs. This generalization also brings more flexibility to the modeling of PFSMs that will allow us to explore new useful classes of sequence models in future work.

5 Inference of silent HMMs

5.1 Silent Circuit Constraint

In this section, we discuss an inference algorithm used for EM training of silent HMMs. For inference of silent HMMs, we need to be careful of an infinite length of state sequence that possibly happen by an infinite loop of transitions between silent states. Explicit consideration of an infinite loop of state transitions obviously complicates an inference algorithm. In this paper, we impose a sufficient condition on meta-parameters (Q, C, R) that ensures the length of a state sequence is finite.

To derive the condition, we consider *silent transition topology*, a directed graph representing possible silent state transitions. The graph is obtained from the directed graph representation of R by omitting outlinks from all the normal states. More formally:

Definition 1 (Silent transition topology)**.** *Let Q be a set of states, C be a mapping that indicates the silence assignment, and R be a transition topology. Let R_s be a set of edges defined as follows:*

$$R_s = \{(q_1, q_2) \in Q_s \times Q | q_1 \xrightarrow{R} q_2\}.$$

*A directed graph $G_s = (Q, R_s)$ is **silent transition topology** induced by (Q, C, R).*

Figure 8 shows an example of a silent transition topology. A silent transition topology represents all the possible transitions allowed in a state sequence at a single time step, $\mathbf{z}_t$. Based on the set of meta-parameters in Figure 8 **(Left)**, we can say a state sequence $\mathbf{z}_t = q_1, q_3, q_4$ never happens at a single time step because q_3 is a normal state that produces an observation. Therefore, the state sequence must split into $\mathbf{z}_t = q_1, q_3$ and $\mathbf{z}_{t+1} = q_4$. The state transition topology (Figure 8 **(Right)**)

clearly expresses this property, since there is no link from q_3 to q_4.

We can use the definition of silent transition topology to derive a sufficient condition that ensures the length of state sequence at a single time step to be finite.

Definition 2 (Silent circuit constraint)**.** *A set of meta-parameters (Q, C, R) satisfies **silent circuit constraint** if the silent transition topology induced by (Q, C, R) does not contain any circuits.*

Theorem 3. *If a silent HMM has meta-parameters satisfying the silent circuit constraint, $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ is always zero whenever $|\mathbf{z}_t| > |Q_s| + 1$ for any t and $\mathbf{z}_{t-1}$.*

Proof. $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ is greater than 0 only when $\mathbf{z}_t$ is a path in the silent transition topology because transition probabilities from q_1 to q_2 are restricted to being zero when $\neg q_1 \xrightarrow{R} q_2$. Since the silent transition topology contains no circuits and normal states have no outlinks, the length of a path in the silent transition topology is at most $|Q_s| + 1$. From these facts, $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = 0$ when $|\mathbf{z}_t| > |Q_s| + 1$. □

Silent HMMs that satisfy the silent circuit constraint form a subclass of general silent HMMs. The following theorems show that the constrained silent HMMs have more expressive power than HSMMs and HHMMs.

Theorem 4. *The silent HMM constructed from a given HSMM by using the method explained in Section 3.2 satisfies the silent circuit constraint.*

Theorem 5. *The silent HMM constructed from a given HHMM by using the method explained in Section 4.2 satisfies the silent circuit constraint.*

These theorems are easily proven by checking that the silent transition topologies contain no circuits. Based on these results, we can apply efficient inference algorithms (explained in the next section) to silent HMMs that are equivalent to HSMMs and HHMMs.

5.2 Inference Algorithms

The inference of silent HMMs indicates a calculation of the expectations of hidden states $\mathbf{z}$ given a sequence of observations $\mathbf{x}$. We describe a modified forward-backward algorithm for the inference of silent HMMs. The forward-backward algorithm is an inference algorithm for normal HMMs based on efficient computation of forward and backward probabilities. In this paper, we explain the calculation of forward probabilities to handle the existence of silent states to apply the algorithm to silent HMMs. For details on the forward-backward algorithm, please refer to (Rabiner, 1989).

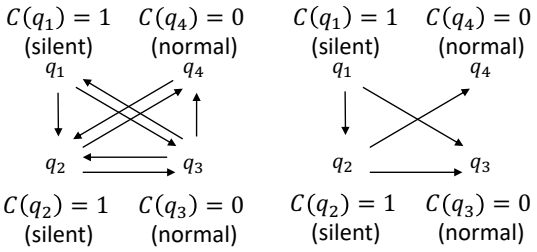

Figure 8: **(Left)** A set of meta-parameters (Q, C, R) that satisfies the silent circuit constraint. **(Right)** Silent transition topology obtained omitting the outlinks from the normal states.

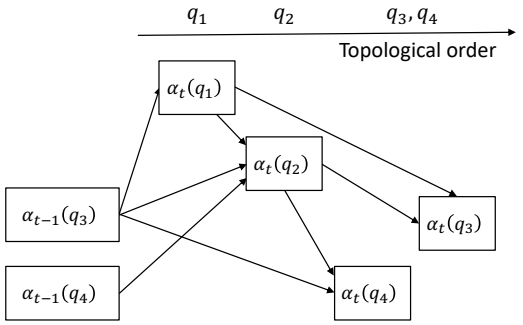

Figure 9: Propagation of the forward probabilities in the silent HMM that has a transition topology in Figure 8. We can calculate the forward probabilities in $O(T|Q|^2)$ with the same complexity as normal HMMs by processing in the topologically sorted order of the states.

The forward probability of state q_i at time step t is defined as the joint probability of the observations until the time step t. For silent HMMs, we divide cases for silent states and normal states as follows.

$$\alpha_t(q) = \begin{cases} p(z_t = q, x_{1:t-1}) & C(q) = 1(silent) \\ p(z_t = q, x_{1:t}) & C(q) = 0(normal). \end{cases}$$

While multiple transitions can be involved in a single time step t in silent HMMs, the forward probabilities can be efficiently calculated by following the topological order of states in the silent transition topology. Given a silent HMM with meta-parameters (Q, C, R), we obtain a silent transition topology and apply the topological sort

algorithm to the directed graph of the silent transition topology. The obtained topological order reflects the possible order of transitions throughout the silent states at a single time step t. Figure 9 shows the flow of the computation of α_t for the silent HMM that has the transition topology expressed in Figure 8. By following the topological order we decided in this way, the computation of α_t for each state can be done with the computational complexity $O(|Q|^2)$, which has the same computational complexity as the normal HMMs. The recursive formula is derived as follows:

$$\alpha_t(q) = \sum_{q' \in Q_n} \alpha_{t-1}(q') A_{q',q} + \sum_{q' \in \mathcal{T}(q) \cap Q_s} \alpha_t(q') A_{q',q},$$

where $\mathcal{T}(q)$ is a set of states that are earlier than q in the topological order imposed on R.

While we are omitting the case for the backward probabilities due to length limitations, we can efficiently calculate the backward counterpart and apply the forward-backward algorithm to ensure the inference is the same computational complexity as normal HMMs. We are not detailing the algorithm that estimates the most likely hidden state sequence, but we can obtain the Viterbi algorithm straightforwardly by replacing the forward computation in the Viterbi algorithm for HMMs (Rabiner, 1989) with the method we explained above.

6 Conclusion

In this paper, we provided formal descriptions of silent HMMs and proposed methods to obtain silent HMMs that are equivalent to given HSMMs and HHMMs. We believe that our results establish a firm foundation to use silent HMMs as a unified framework for PFSM modeling.

Future work includes developing PFSMs for modeling structures in natural language (e.g., morphological structure) by combining the structural assumptions in HSMMs and HHMMs in the framework of silent HMMs. Other future work is more advanced Bayesian extensions of silent HMMs incorporating Dirichlet prior (Foti et al., 2014) and nonparametric Bayesian prior (Beal et al., 2002; Heller et al., 2009). These extensions coule enable us to estimate the number of states during the training process, offering more powerful PFSM modeling methods.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19K20333 and 16H02904.

References

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. 2002. The infinite hidden markov model. In *Advances in Neural Information Processing Systems 14*, pages 577–584.

Hung H. Bui, Dinh Q. Phung, and Svetha Venkatesh. 2004. Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the 19th national conference on Artifical intelligence*, pages 324–329.

Sean R. Eddy. 1998. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763.

Razieh Ehsani, Berke Ozenc, and Ercan Solak. 2017. A fst description of noun and verb morphology of azarbaijani turkish. In *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing*, pages 62–68.

Shai Fine, Yoram Singer, and Naftali Tishby. 1998. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62.

Nick Foti, Jason Xu, Dillon Laird, and Emily Fox. 2014. Stochastic variational inference for hidden markov models. In *Advances in Neural Information Processing Systems 27*, pages 3599–3607.

Katherine Heller, Yee W. Teh, and Dilan Gorur. 2009. Infinite hierarchical hidden markov models. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pages 224–231.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. Unsupervised grammar induction with depth-bounded pcfg. *Transactions of the Association for Computational Linguistics*, 6:211–224.

Anders Krogh, I. Saira Mian, and David Haussler. 1994. A hidden Markov model that finds genes in E.coli DNA. *Nucleic Acids Research*, 22(22):4768–4778.

Michael D. Moore and Michael I. Savic. 2004. Speech reconstruction using a generalized hsmm (ghsmm). *Digital Signal Processing*, 14(1):37–53.

Kevin P. Murphy and Mark A. Paskin. 2002. Linear Time Inference in Hierarchical HMMs. In *Advances in Neural Information Processing Systems 14*, pages 833–840.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Shashikant Sharma and Anil Kumar Singh. 2017. Word transduction for addressing the oov problem in machine translation for similar resource-scarce languages. In *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing*, pages 56–63.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden markov models. *Transactions of the Association for Computational Linguistics*, 4:245–257.

M. Ines Torres. 2013. Stochastic bi-languages to model dialogs. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, pages 9–17.

Kei Wakabayashi. 2018. Segmentation-based unsupervised phrase detection. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*, pages 138–142.

Kei Wakabayashi and Takao Miura. 2012. Forward-backward activation algorithm for hierarchical hidden markov models. In *Advances in Neural Information Processing Systems 25*, pages 1493–1501.

Mengqiu Wang and Christopher D. Manning. 2012. Probabilistic finite state machines for regression-based MT evaluation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 984–994.

Michèle Weiland, Alan Smaill, and Peter C. Nelson. 2005. Learning musical pitch structures with hierarchical hidden markov models. Technical report, University of Edinburgh.

Shun-Zheng Yu. 2010. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243.