1. INTRODUCTION

Statistical n-gram language models have been successfully developed for continuous speech recognition and many other applications. The n-gram models suffer from the insufficiencies of long-distance information, which limit the model performance. In [10], the long-distance information was captured by using cache language model in which the short-term pattern in historical words was continuously merged for word prediction. In [4], the multiple distant associated words were identified in the association pattern language models. In addition, the weakness of n-gram model was compensated by extracting the latent semantic information. Bellegarda [1] applied the latent semantic analysis (LSA) to build large-span language model. The semantic regularities extracted from the probabilistic LSA (PLSA) were also incorporated in n-gram model for speech recognition [7]. PLSA was applied to model the collected documents. The unseen documents could not be characterized. The parameter size was increased significantly with a large amount of documents. Blei et al. [2] presented the latent Dirichlet allocation (LDA) by merging the Dirichlet priors for topic mixtures. The unseen documents were generalized by LDA parameters that were estimated by the variational inference method. LDA was employed for transfer learning [13] from labeled and unlabeled documents. Typically, LDA extracted the topic information at document level and were used for building topic-based language model [12].

In addition, we built the latent Dirichlet language model (LDLM) [5] in which the topic structure of n-gram events of a training corpus was directly incorporated in language model. The word ordering was considered for speech recognition. LDLM was smoothed by the shared topic parameters which were estimated by maximizing the marginal data likelihood. Although LDLM tackled the data sparseness issue and bag-of-words problem in LDA language model, the useful information outside of n-gram window was not considered. In this study, we are motivated to build the topic cache language model (TCLM) to capture the long-distance information for LDLM. TCLM is established by continuously updating the topic statistics from a large-span historical context. Importantly, the large-span topic statistics are integrated in TCLM through a Bayesian framework where the multinomial likelihood of a topic sequence and the Dirichlet prior assumption of Markov chain in n-gram models by considering the self trigger using all historical words. Analogous to the cache memory in hardware technology, the cache language model increases the word probability if the word occurs frequently in the large-span history. The pioneering work of cache model was proposed in [10] and applied for class-based language model [3].

2. SURVEY OF CACHE LANGUAGE MODELS

The basic idea of cache language model aims to release the assumption of Markov chain in n-gram models by considering the self trigger using all historical words. The cache model was employed for word prediction. In [4], the multiple distant associated words were identified in the association pattern language models. In addition, the weakness of n-gram model was compensated by extracting the latent semantic information. Bellegarda [1] applied the latent semantic analysis (LSA) to build large-span language model. The semantic regularities extracted from the probabilistic LSA (PLSA) were also incorporated in n-gram model for speech recognition [7]. PLSA was applied to model the collected documents. The unseen documents could not be characterized. The parameter size was increased significantly with a large amount of documents. Blei et al. [2] presented the latent Dirichlet allocation (LDA) by merging the Dirichlet priors for topic mixtures. The unseen documents were generalized by LDA parameters that were estimated by the variational inference method. LDA was employed for transfer learning [13] from labeled and unlabeled documents. Typically, LDA extracted the topic information at document level and were used for building topic-based language model [12].

In addition, we built the latent Dirichlet language model (LDLM) [5] in which the topic structure of n-gram events of a training corpus was directly incorporated in language model. The word ordering was considered for speech recognition. LDLM was smoothed by the shared topic parameters which were estimated by maximizing the marginal data likelihood. Although LDLM tackled the data sparseness issue and bag-of-words problem in LDA language model, the useful information outside of n-gram window was not considered. In this study, we are motivated to build the topic cache language model (TCLM) to capture the long-distance information for LDLM. TCLM is established by continuously updating the topic statistics from a large-span historical context. Importantly, the large-span topic statistics are integrated in TCLM through a Bayesian framework where the multinomial likelihood of a topic sequence and the Dirichlet prior of topic mixture vector are combined. The dynamic topics of historical words are updated in an online fashion. The experiments on continuous speech recognition illustrate the superiority of TCLM to other methods.
was interpolated with baseline $n$-gram for language model adaptation. In addition, a topic mixture language model was calculated by combining $M$ topic-dependent $n$-grams as [8]

$$p_{TM}(w_i | w_{i-1}^{j-1}) = \sum_{j=1}^{M} \lambda_j p_j(w_i | w_{i-1}^{j-1})$$  \hspace{1cm} (3)

where $p_j(w_i | w_{i-1}^{j-1})$ denoted the topic-dependent $n$-gram for topic $j$. This model was merged with a cache model through a dynamic adaptation procedure. Each mixture component had its own cache. The cache-based topic mixture model was calculated by

$$p_{TM}^c(w_i | w_{i-1}^{j-1}) = \sum_{j=1}^{M} \lambda_j [\mu p_j(w_i | w_{i-1}^{j-1}) + (1-\mu)p_j^c(w_i | w_{i-1}^{j-1})]$$  \hspace{1cm} (4)

with an interpolation weight $\mu$ and the topic-dependent cache model $p_j^c(w_i | w_{i-1}^{j-1})$ estimated by counting the occurrence of $n$-gram events in the cache consisting of all historical sentences classified to topic $j$. The topic posterior probability due to historical sentences was considered to calculate the cache probability. Each word in the cache contributed to different topic-dependent cache models.

### 3. TOPIC CACHE LANGUAGE MODEL

We have surveyed three cache language models based on the class $n$-gram, word $n$-gram and topic mixture language model. In what follows, we present a new topic cache model for latent Dirichlet language modeling.

#### 3.1 Latent Dirichlet language model

LDLM [5] acts as a Bayesian topic language model in which the prior density of topic variable is characterized by $n$-gram events $\{w_i, w_{i-1}^{j-1}\}$. Using LDLM with $V$ vocabulary words, the historical words $w_{i-1}^{j-1}$ are first represented by a $(n-1)V \times 1$ vector $h_{i-1}^{j-1}$ consisting of $n-1$ block subvectors with the entries of the seen words set to one and those of unseen words set to zeros. The order of historical words is represented in $h_{i-1}^{j-1}$. We can use a linear discriminant function [5] to evaluate the historical words appearing in different topics $k$ by

$$g_k(h_{i-1}^{j-1}) = a_k^T h_{i-1}^{j-1}$$  \hspace{1cm} (5)

and $A = [a_1 \cdots a_k]$ is a parameter matrix. This function reflects the topic posterior probability $p(k | h_{i-1}^{j-1})$ which is essential for predicting the topic information for unseen history. LDLM is constructed as a Bayesian model by compensating the uncertainty due to the latent topics $k$ and topic mixtures $\theta$. The topic $k$ in LDLM is drawn from a history-dependent Dirichlet prior $\theta = [\theta_1, \cdots, \theta_k]^T \sim \text{Dir}(\alpha_1^{h_{i-1}^{j-1}})$. Each word $w_i$ is predicted according to a multinomial distribution with a parameter $\beta$ which is associated with a topic $k_i$. This topic is chosen from a topic mixture $\theta$ with a Dirichlet prior density determined by the historical words $w_{i-1}^{j-1}$ and the projection matrix $A$. The joint probability of word $w_i$, topic $k_i$ and topic mixture vector $\theta$ conditioned on history $h_{i-1}^{j-1}$ and LDLM parameters $\{A, \beta\}$ is computed by

$$p(w_i, k_i, \theta | h_{i-1}^{j-1}, A, \beta) = p(w_i | k_i, \beta) p(k_i | \theta) p(\theta | h_{i-1}^{j-1}, A).$$  \hspace{1cm} (6)

The parameters $\{A, \beta\}$ were estimated using the variational Bayes expectation maximization (VB-EM) algorithm [5]. The $n$-gram probability is expressed in a form of marginal likelihood as

$$p(w_i | h_{i-1}^{j-1}, A, \beta) = \sum_{k_i=1}^{K} p(w_i | k_i, \beta) p(k_i | \theta) p(\theta | h_{i-1}^{j-1}, A)$$

$$= \sum_{k_i=1}^{K} \beta_{k_i} a_{k_i}^T h_{i-1}^{j-1}.$$  \hspace{1cm} (7)

The integral in (7) is operated over a Dirichlet distribution $p(\theta | h_{i-1}^{j-1}, A)$ and is obtained as a distribution mean. Inherently, LDLM parameters are smoothed in the latent topic space. The smoothed model with Bayesian treatment is assured with good model regularization. This LDLM can be viewed as a kind of class-based $n$-gram. However, the class-based $n$-gram in [3] adopted the additional measurement, e.g. mutual information. In contrast, LDLM simultaneously identifies the hidden classes and estimates the model parameters by maximizing the marginal likelihood given the model structure.

#### 3.2 Topic cache language model

In LDLM procedure, the topic mixtures $\theta$ are drawn from history $w_{i-1}^{j-1}$ using the Dirichlet distribution with parameter $A^T h_{i-1}^{j-1}$. The topic probability $p(k_i | \theta)$ is calculated. The word $w_i$ is predicted by incorporating the categorical or multinomial parameters $\beta = [\beta_{a_i} a_k]$. In this procedure, only the short-term history $w_{i-1}^{j-1}$ is considered in word prediction. To compensate the insufficient long-distance information in LDLM, we strive towards continuously updating the topic information and consistently building a topic cache language model (TCLM) under Bayesian framework. TCLM is seen as a cache extension of LDLM. Using TCLM, the topic mixtures are not only generated from historical words but also from the topics $h_{i-1}^{j-1}$ and LDLM parameters $\{A, \beta\}$. The integral in (7) is operated over a Dirichlet distribution $p(\theta | h_{i-1}^{j-1}, A)$ and is obtained as a distribution mean. Inherently, LDLM parameters are smoothed in the latent topic space. The smoothed model with Bayesian treatment is assured with good model regularization. This LDLM can be viewed as a kind of class-based $n$-gram. However, the class-based $n$-gram in [3] adopted the additional measurement, e.g. mutual information. In contrast, LDLM simultaneously identifies the hidden classes and estimates the model parameters by maximizing the marginal likelihood given the model structure.
where \( p(k_i | \mathbf{h}_{i-1}, \mathbf{A}) = p(k_i | \mathbf{h}_{i-1}, \mathbf{A}) = \prod_{k=1}^{K} \beta_k \delta(k, k_i) \) and the term \( \sum_{k=1}^{K} \delta(k, k_i) \) denotes the number of occurrences of topic \( k \) in topic sequence \( k_{i-1} \). Since the summation in (8) is computationally expensive, here, we simplify the calculation by considering a single best topic sequence \( k_{i-1} \), which is recursively detected from word \( w_{i-1} \) and the best preceding topics \( k_{i-2} \) by Bayesian formulation. Accordingly, the word probability given its history using TCLM is calculated based on the graphical model shown in Figure 1. The topic mixture vector is drawn from the historical using TCLM is calculated based on the graphical model Bayesian formulation. Accordingly, the word probability given its history using TCLM is calculated based on the graphical model where \( \text{detected from word } w_{i-1} \) and the best preceding topics \( k_{i-2} \) and dropping the term \( \sum_{k=1}^{K} \delta(k, k_i) \) factor 10 is applied to discount the distant topic information. The topic associated with the farther word has a smaller impact on the prediction of current word. The order of topic sequence is weighted. In the experiments, the tuning parameters \( \rho \) and \( \tau \) were found from development data and were applied in (10) to detect the best topics. When comparing (7) of LDLM and (11) of TCLM, it is interesting that TCLM is established by additionally incorporating the topic unigram of all preceding words. The more frequent a topic occurs, the greater weight this topic contributes to the model. In the case of \( \rho = 0 \), TCLM reduced to LDLM. If a very large \( \rho \) is used, TCLM is comparable to build a topic-based cache, which is different from the word-based cache in previous cache language models [6][8][10].

Attractively, the proposed TCLM is engaged in a tight combination of LDLM and a cache topic unigram. In contrast to previous methods using a linear interpolation, the new hybrid model is theoretically established by Bayesian framework where the Dirichlet priors are merged in drawing the topic information and the Dirichlet posterior probability is produced by combining the multinomial likelihood of a cache model due to the property of conjugate prior. It is important to marginalize the cache posterior probability over the uncertainty of topic mixtures to guarantee model generalization. The cache modeling based on Bayesian framework in TCLM is illustrated by Figure 2. When comparing the word-based cache language model and TCLM, we find that the forgetting factor \( \tau \) in (11) is comparable to the term \( \exp(-\epsilon) \) in (2). The proposed TCLM is naturally normalized due to the fact that the mean vector of a Dirichlet distribution is picked up from the integral operation. This is different from (2) that an additional normalization factor \( \eta \) was used in word-based cache model [6].

The sources of historical words and historical topics are used to predict the topic of a word. In (10), the Dirichlet prior and the multinomial distribution of topic sequence \( k_{i-2} \) are merged in a new Dirichlet distribution and so the marginalization is feasible for the new Dirichlet. The recursion \( \hat{k}_i \to \hat{k}_i \to \cdots \to \hat{k}_{i-1} \) is accordingly implemented to find \( \hat{k}_{i-1} \). By substituting (9) into (8) and dropping the term \( p(k_{i-1} | w_{i-1}) \) that is independent of \( w_i \), TCLM is obtained by

\[
p(w_i | w_{i-1}, \mathbf{A}, \beta) = \sum_{k=1}^{K} \beta_k h_{i-1}^{w_{i-1}} A(k, k_i) \prod_{k=1}^{K} \beta_k \delta(k, k_i) \frac{\Gamma(\sum_{k=1}^{K} \beta_k)}{\prod_{k=1}^{K} \Gamma(\beta_k) \prod_{j=1}^{\min(k, j)} \delta(\hat{k}_j)}
\]

Again, the product of Dirichlet distribution \( p(\mathbf{h}_{i-1}, \mathbf{A}) \) and multinomial distribution \( \prod_{k=1}^{K} \delta(k, k_i) \) is expressed as a new Dirichlet distribution. The integral in (11) is equivalent to picking up the mode of new Dirichlet distribution. This is an attractive property of using Dirichlet density as a conjugate prior. The derivation of (11) is similar to that of (10). In (11), a weighting factor \( 0 < \rho \leq 1 \) is empirically introduced to balance the factors between \( p(\mathbf{h}_{i-1}, \mathbf{A}) \) and \( \prod_{k=1}^{K} \delta(k, k_i) \), and a forgetting factor \( 0 < \tau \leq 1 \) is applied to discount the distant topic information. The topic associated with the farther word has a smaller impact on the prediction of current word. The order of topic sequence is weighted. In the experiments, the tuning parameters \( \rho \) and \( \tau \) were found from development data and were applied in (10) to detect the best topics. When comparing (7) of LDLM and (11) of TCLM, it is interesting that TCLM is established by additionally incorporating the topic unigram of all preceding words. The more frequent a topic occurs, the greater weight this topic contributes to the model. In the case of \( \rho = 0 \), TCLM reduced to LDLM. If a very large \( \rho \) is used, TCLM is comparable to build a topic-based cache, which is different from the word-based cache in previous cache language models [6][8][10].

Adaptively, the proposed TCLM is engaged in a tight combination of LDLM and a cache topic unigram. In contrast to previous methods using a linear interpolation, the new hybrid model is theoretically established by Bayesian framework where the Dirichlet priors are merged in drawing the topic information and the Dirichlet posterior probability is produced by combining the multinomial likelihood of a cache model due to the property of conjugate prior. It is important to marginalize the cache posterior probability over the uncertainty of topic mixtures to guarantee model generalization. The cache modeling based on Bayesian framework in TCLM is illustrated by Figure 2. When comparing the word-based cache language model and TCLM, we find that the forgetting factor \( \tau \) in (11) is comparable to the term \( \exp(-\epsilon) \) in (2). The proposed TCLM is naturally normalized due to the fact that the mean vector of a Dirichlet distribution is picked up from the integral operation. This is different from (2) that an additional normalization factor \( \eta \) was used in word-based cache model [6].

4. EXPERIMENTS

The Wall Street Journal (WSJ) corpus [11] was used to evaluate the proposed TCLM for continuous speech recognition (CSR).
The SI-84 training set was adopted to estimate the HMM parameters. The acoustic feature vector consisted of twelve Mel-frequency cepstral coefficients and one log energy and their first and second derivatives. The triphone models were built for 39 phones and one background silence. Each triphone model had three states with eight Gaussian mixtures. The HTK toolkit was used for acoustic model training and lattice generation. The ’87-89 WSJ corpus with 38M words was used to train the modified Kneser-Ney backoff trigram [9] using the SRILM toolkit. We used the 20k non-verbalized punctuation, closed vocabulary. There were 333 test utterances sampled from the November 1992 ARPA CSR benchmark test data. These test utterances were used to examine different models in terms of model perplexity and word error rate (WER). A development set consisting of 4,002 utterances was provided [11]. In CSR implementation, the baseline trigram was used to generate the 100-best lists. Different topic-based trigrams were interpolated with baseline trigram (denoted by baseline LM) using an interpolation weight, and were applied for N-best rescoring. The cache trigram [6] (denoted by cache LM), LDA LM [12], LDLM and TCLM were carried out for comparison. The tuning parameters $\{\rho, \tau\}$ in TCLM, the topic sizes in LDA LM, LDLM and TCLM and the interpolation weight between topic-based model and baseline LM were empirically determined from development data according to the metric of perplexity. In VB-EM procedure, the initial values in $\mathbf{b}$ were set to $1/V$ and those in $\mathbf{a}$ were randomly set in the range $[0, 1]$. Totally, ten VB-EM iterations were run.

<table>
<thead>
<tr>
<th>WER (%)</th>
<th>Baseline LM</th>
<th>Cache LM</th>
<th>LDA LM</th>
<th>LDLM</th>
<th>TCLM</th>
<th>Error Rate Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12.9</td>
<td>12.7</td>
<td>12.2</td>
<td>12.0</td>
<td>11.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>5.4</td>
<td>7.0</td>
<td>7.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First, we performed the evaluation of model perplexity. The perplexities of baseline LM and cache LM were measured as 146 and 143, respectively. Using LDA LM and LDLM, the perplexities were reduced to 128 and 126, respectively. The perplexities of TCLM with $K=200$ using different $\tau$ and $\rho$ are displayed in Figure 3. The lowest perplexity 120.6 was obtained by using TCLM in case of $\tau=0.75$ and $\rho=0.001$. In evaluation of speech recognition, we report WERs (%) and error rate reductions (%) of using different LMs over baseline LM in Table 1. The baseline LM had a WER of 12.9% and the cache LM attained WER as 12.7%. LDLM obtained WER of 12.0% which was lower than that using LDA LM. When merging large-span topic information, the resulting TCLM further reduced WER to 11.9%. In addition, we performed the match-pairs test and counted the misrecognized words in each test utterance by various methods. The P value was measured as 0.0052, 0.016 and 0.048 by using TCLM relative to cache LM, LDA LM and LDLM, respectively. At a significance level of 0.05, TCLM was significantly better than the other methods.

5. CONCLUSIONS

We extended the topic-based language model by incorporating the large-span topic cache in generation of topic mixtures. The proposed TCLM was built through the Bayesian learning. The word prediction in test sentence was significantly improved due to the smoothing process via topic-based modeling and the extraction of long-distance information via cache modeling. We obtained the improvements in perplexity as well as WER by using TCLM compared to baseline LM, cache LM and LDA LM on using WSJ corpus. In future studies, we will incorporate the cache topics into the maximum entropy or neural network language model. The computation complexity of different methods will be evaluated.

6. REFERENCES