DOCUMENT-SPECIFIC CONTEXT PLSA LANGUAGE MODEL FOR SPEECH RECOGNITION

Md Akmal Haidar and Douglas O'Shaughnessy

INRS-EMT, 800 de la Gauchetiere Ouest, Bureau 6900, H5A 1K6, Montreal (QC), Canada

haidar@emt.inrs.ca, dougo@emt.inrs.ca

ABSTRACT

In this paper, we introduce a document-specific context probabilistic latent semantic analysis (DCPLSA) model for speech recognition. This is an extension of a CPLSA model [1] where the probability of word is conditioned only on topics. The CPLSA model uses the bigram counts that are the number of appearances of the bigrams in the corpus. These counts are the sum of the bigram counts in different documents where they could appear to describe different topics. We encounter this problem in the CPLSA model and introduce the document-specific CPLSA model (DCPLSA) where the probability of a word is conditioned on both topic and document. We carried out experiments on a continuous speech recognition (CSR) task using the Wall Street Journal (WSJ) corpus and have seen that the proposed DCPLSA approach yields significant reduction in both perplexity and word error rate (WER) measurements over the other approaches used in the literature.

Index Terms— Topic models, bigram PLSA models, speech recognition, context-based PLSA language model, statistical language model

1. INTRODUCTION

Statistical n-gram language models (LMs) have been used successfully in speech recognition and many other applications. They capture only the short-range information in the language and suffer from a shortage of long-range information, which limits performance. To handle the long-range information, many approaches have been tried such as cache-based LMs [2] and trigger-based LM adaptation [3]. Recently, latent topic analysis has been used broadly to compensate for the weaknesses of n-gram models. Several techniques such as Latent Semantic Analysis (LSA) [4, 5], PLSA [6, 7], and Latent Dirichlet Allocation (LDA) [8] have been studied to extract the latent semantic information (topics) from a training corpus. These methods have been used successfully for speech recognition [5, 7, 9, 10, 11, 12, 13, 14]. A bigram LDA topic model has been recently investigated [15], where the word probabilities are conditioned on their preceding history context and the topic probabilities are conditioned on the documents. A similar model, but in the PLSA framework, called the bigram PLSA model was introduced recently [16]. An updated bigram PLSA model (UBPLSA) was proposed in [17] where the topic is further conditioned on the bigram history context to the original bigram PLSA model [16]. In the UBPLSA model, only the seen bigram probabilities are trained. This approach is not practical as it assigns zero probability to the unseen bigrams in the training and yields incorrect topic probabilities of the unseen test document. To overcome the limitation of the UBPLSA model, a context-based PLSA (CPLSA) model [1] was introduced.

In this paper, we extend our previous work [1] and propose a new document-specific context PLSA (DCPLSA) model. The CPLSA model [1] uses the sum of bigrams in all documents to compute the word probabilities for topics. However, words in the bigrams may describe different topics in different documents. For example, the bigram *White House* can occur in a document where it describes a real estate topic. Also, it can occur in another document that describes a political topic. Therefore, the probability of word given only the topics may not give the appropriate results. This motivates us to introduce the DCPLSA model where the word probabilities are trained by conditioning on the topics and the documents. We have seen significant improvement using perplexity and word error rate (WER) measurement. However, the DCPLSA model requires more complexity and memory requirement than the CPLSA model.

The rest of the paper is organized as follows. Section 2 describes the PLSA, the UBPLSA, and the CPLSA models. The proposed DCPLSA model is described in section 3. The calculation of the n-gram probabilities of the unseen test document is illustrated in section 4. A comparison of the UBPLSA, CPLSA and DCPLSA models is studied in section 5. In section 6, the time complexity and memory requirements of the UBPLSA, CPLSA and DCPLSA models are analysed. The experimental details are described in section 7. Finally the conclusions are explained in section 8.

2. REVIEW OF PLSA, UBPLSA, AND CPLSA MODELS

2.1. PLSA Model

The PLSA model [7] extracts semantic information from a corpus in a probabilistic framework. It uses an unobserved topic variable with each observation, i.e., with each occurrence of a word in a document. It is assumed that the document and the word are independent conditioned on the state of the latent topic variable. It models each word in a document as a sample from a mixture model, where the mixture models can be viewed as representations of topic distributions. Therefore, a document is generated as a mixture of topic distributions and reduced to a fixed set of topics. Each topic is a distribution over words. The model [7] can be described in the following procedure. First a document d_l $(l = 1, 2, \dots, N)$ is selected with probability $P(d_l)$. A topic t_k $(k = 1, 2, \dots, K)$ is then chosen with probability $P(t_k|d_l)$, and finally a word w_j $(j = 1, 2, \dots, M)$ is generated with probability $P(w_j|t_k)$. The probability of word w_j given a document d_l can be estimated as:

$$P(w_j|d_l) = \sum_{k=1}^{K} P(w_j|t_k)P(t_k|d_l). \qquad (1)$$

The model parameters $P(w_j|t_k)$ and $P(t_k|d_l)$ are computed by using the expectation maximization (EM) algorithm [7].

2.2. UBPLSA Model

The PLSA model yields unigram models for topics. To improve the performance, a bigram PLSA model [16] was introduced where the bigram probabilities for topics were trained instead of unigrams in the PLSA model. Before describing the UBPLSA model, the previous bigram PLSA model is briefly explained. Instead of $P(w_j|t_k)$ in Equation 1, the bigram PLSA model uses $P(w_j|w_i, t_k)$ in computing the probability of word w_j given the bigram history w_i and the document d_l:

$$P(w_j|w_i, d_l) = \sum_{k=1}^{K} P(w_j|w_i, t_k)P(t_k|d_l). \qquad (2)$$

The model parameters are computed using the EM procedure [16].

The UBPLSA model was recently proposed in [17], which outperforms the previous bigram PLSA model [16]. Here, the topic probability is further conditioned on the bigram history context. It can model the topic probability for the document given a context, using the word co-occurrences in the document. In this model, the probability of the word w_j given the document d_l and the word history w_i is computed as:

$$P(w_j|w_i, d_l) = \sum_{k=1}^{K} P(w_j|w_i, t_k)P(t_k|w_i, d_l). \qquad (3)$$

The model parameters are computed using the EM procedure [17].

2.3. CPLSA Model

The problem of the UBPLSA model is that it uses only seen bigrams for training. Therefore, it cannot compute all the possible bigram probabilities in the training phase. It results in incorrect topic probabilities of the test document. This is because the model cannot compute topic probabilities for some history contexts that are present both in the training and test sets. To overcome the limitations of the UBPLSA model, the CPLSA model was introduced [1].

The CPLSA model is similar to the original PLSA model except the topic is further conditioned on the history context as is the UBPLSA model. Using this model, we can compute the bigram probability using the unigram probabilities of topics as:

$$P(w_j|w_i, d_l) = \sum_{k=1}^{K} P(w_j|t_k)P(t_k|w_i, d_l). \qquad (4)$$

The parameters of the model are computed as:
E-step:

$$P(t_k|w_i, w_j, d_l) = \frac{P(w_j|t_k)P(t_k|w_i, d_l)}{\sum_{k'} P(w_j|t_{k'})P(t_{k'}|w_i, d_l)}, \qquad (5)$$

M-step:

$$P(w_j|t_k) = \frac{\sum_{i'} \sum_{l'} n(w_{i'}, w_j, d_{l'})P(t_k|w_{i'}, w_j, d_{l'})}{\sum_{j'} \sum_{i'} \sum_{l'} n(w_{i'}, w_{j'}, d_{l'})P(t_k|w_{i'}, w_{j'}, d_{l'})}, \qquad (6)$$

$$P(t_k|w_i, d_l) = \frac{\sum_{j'} n(w_i, w_{j'}, d_l)P(t_k|w_i, w_{j'}, d_l)}{\sum_{k'} \sum_{j'} n(w_i, w_{j'}, d_l)P(t_{k'}|w_i, w_{j'}, d_l)}. \qquad (7)$$

From Equations 5 and 7, we see that the model can compute all the possible bigram probabilities of the seen history context in the training set. Therefore, the model can overcome the problem of computing topic probabilities of the test document using the UBPLSA model, which causes the problem in the computation of the bigram probabilities of the test document.

3. PROPOSED DCPLSA MODEL

In the CPLSA model, the word probabilities for topics are computed using the sum of the bigram events in all training documents where the words may appear to describe different topics in different documents. Therefore, the word probabilities given only the topics will not give proper results. In this section, we describe a new topic model where the document-specific word probabilities for topics are trained. The DCPLSA model is similar to the original CPLSA model except that the document-based word probabilities for topics are computed instead of the global word probabilities for topics in the CPLSA model. To better understand the model, the matrix decomposition of the DCPLSA model is described in Figure 1. Using this model, we can compute the bigram probability for a document as:

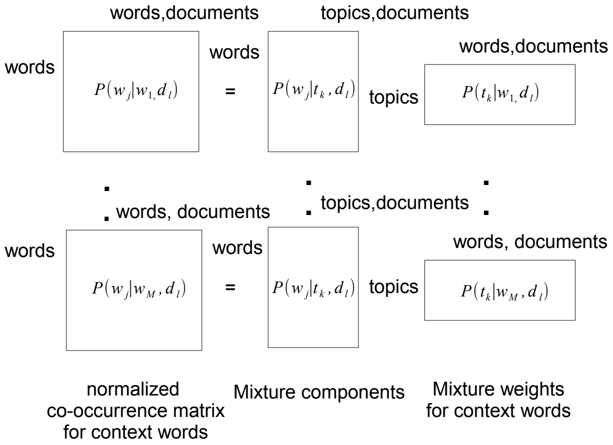

Fig. 1. Matrix decomposition of the DCPLSA model

$$P(w_j|w_i, d_l) = \sum_{k=1}^{K} P(w_j|t_k, d_l)p(t_k|w_i, d_l). \qquad (8)$$

The parameters of the model are computed as:
E-step:

$$P(t_k|w_i, w_j, d_l) = \frac{P(w_j|t_k, d_l)P(t_k|w_i, d_l)}{\sum_{k'} P(w_j|t_{k'}, d_l)P(t_{k'}|w_i, d_l)}, \qquad (9)$$

M-step:

$$P(w_j|t_k, d_l) = \frac{\sum_{i'} n(w_{i'}, w_j, d_l)P(t_k|w_{i'}, w_j, d_l)}{\sum_{j'} \sum_{i'} n(w_{i'}, w_{j'}, d_l)P(t_k|w_{i'}, w_{j'}, d_l)}, \qquad (10)$$

$$P(t_k|w_i, d_l) = \frac{\sum_{j'} n(w_i, w_{j'}, d_l)P(t_k|w_i, w_{j'}, d_l)}{\sum_{k'} \sum_{j'} n(w_i, w_{j'}, d_l)P(t_{k'}|w_i, w_{j'}, d_l)}. \qquad (11)$$

4. N-GRAM PROBABILITIES OF THE TEST DOCUMENT

We used the folding-in procedure [7] to compute the n-gram probabilities of the test document d_t using the above models. For the PLSA model, we keep the unigram probabilities for topics $P(w_j|t_k)$ fixed and used them to compute the topic probabilities $P(t_k|d_t)$ using EM iterations and then compute the unigram probabilities $P(w_j|d_t)$ using Equation 1. In the UBPLSA model, the bigram probabilities $P(w_j|w_i, t_k)$ remain unchanged while computing the topic probabilities $P(t_k|w_i, d_t)$ using EM iterations. The bigram probabilities $P(w_j|w_i, d_t)$ are then computed using Equation 3. However, the topic probabilities $P(t_k|w_i, d_t)$ for some histories w_i were assigned zeros, as the training model gives zero probabilities to the unseen bigrams in the training model [17]. Therefore, some bigrams of the test document with history context w_i were assigned zero probabilities. The problem is solved by the CPLSA model, which is able to assign probabilities to all the bigrams of the seen history context in the training set. In the CPLSA model, $P(w_j|t_k)$ remains fixed in the EM iterations of the test phase in computing $P(t_k|w_i, d_t)$. Finally, the bigram probabilities $P(w_j|w_i, d_t)$ are computed using Equation 4.

For the DCPLSA, we have word probabilities $P(w_j|t_k, d_l)$ for topics of each training document d_l. During testing, we kept $P(w_j|t_k, d_l)$ unchanged and used them to compute $P(t_k|w_i, d_t, d_l)$ for the test document d_t.

The seen bigram probabilities of the test document d_t are then computed as:

$$\begin{aligned} P(w_j|w_i, d_t) &= \sum_{l=1}^{N} P(w_j|w_i, d_t, d_l)P(d_l|w_i) \\ &= \sum_{l=1}^{N}(\sum_{k=1}^{K} P(w_j|t_k, d_l)P(t_k|w_i, d_t, d_l)) \times \\ &\frac{C(w_i, d_l)}{\sum_{l=1}^{N} C(w_i, d_l)} \end{aligned} \tag{12}$$

where $C(w_i, d_l)$ is the count of w_i in the training document d_l. However, for some seen bigrams of the test document, the words of the bigram cannot be found together in any of the training documents. Their probabilities are computed as:

$$P(w_j|w_i, d_t) = \sum_{l=1}^{N}(\sum_{k=1}^{K} P(w_j|t_k, d_l)P(t_k|w_i, d_t, d_l))P(d_l) \tag{13}$$

where $P(d_l) = 1/N$.

The zero probabilities of the obtained matrix $P(w_j|w_i, d_t)$ are then computed by using back-off smoothing. To capture the local lexical regularities, the model is then interpolated with a back-off trigram background model.

5. COMPARISON OF UBPLSA, CPLSA & DCPLSA MODELS

The UBPLSA, CPLSA and DCPLSA models are differentiated by the word probabilities. The bigram probabilities for topics, the unigram probabilities for topics, and the unigram probabilties given the topics and documents are trained, for the UBPLSA, the CPLSA and the DCPLSA models respectively. The CPLSA model requires less memory and complexity than the other models. The memory and complexity requirements for the DCPLSA model are less than the UBPLSA model if the number of seen bigrams is higher than the product of the number of vocabulary words and the documents. As the UBPLSA model and the CPLSA model, the proposed DCPLSA model can also be extended to the n-gram ($n > 2$) case with increasing complexity and memory space requirements.

6. COMPLEXITY ANALYSIS OF THE UBPLSA, CPLSA AND DCPLSA MODELS

The numbers of free parameters for the UBPLSA, CPLSA and DCPLSA models are $M(M-1)K + (K-1)MN$, $(M-1)K + (K-1)MN$, and $(M-1)KN + (K-1)MN$ respectively. Here, M, K, and N represent the number of words, the number of topics and the number of documents, respectively. From the above discussion, we note that the CPLSA model needs fewer parameters, hence requires smaller memory space than the other models. The DCPLSA model requires fewer parameters than the UBPLSA model as long as the number of documents N is less than the number of vocabulary words M.

In the E-step of the EM algorithm, we have to compute $P(t_k|w_i, w_j, d_l)$ for all i, j, k, l. Therefore, the time complexity of the UBPLSA model [17], the CPLSA model [1] and the DCPLSA model is $O(M^2NK)$. The time complexities for the M-step are $O(KNB)$, $O(MNK)$ and $O(MN^2K)$ for the UBPLSA, the CPLSA and the proposed DCPLSA models respectively. Here, B is the average number of word pairs in the training documents [17]. The size of B is obviously greater than the size of M. Therefore, the CPLSA model also needs less training time than the other models. The DCPLSA model can require less training time than the UBPLSA model as long as $M \times N$ is less than B.

7. EXPERIMENTS

7.1. Data and Experimental Setup

We randomly selected 500 documents from the '87-89 WSJ corpus [18] for training the PLSA, the UBPLSA, the CPLSA and the DCPLSA models. The total number of words in the documents is 224,995. We used the 5K non-verbalized punctuation closed vocabulary from which we removed the MIT stop word list [19] and the infrequent words that occur only once in the training documents. After these removals, the total number of vocabulary is 2628. We could not consider more training documents due to the higher computational cost and huge memory requirements for the UBPLSA model [17] and the DCPLSA models. However, trigram models give better results than the bigram models when more training data are considered. As a small amount of training data can be considered in the UBPLSA and DCPLSA models, the reliability of trigrams decreases more severely than that of bigrams and the bigrams are more robust than the trigrams [20]. For these reasons, we train only the bigram UBPLSA, CPLSA and DCPLSA models. Also, we used the same number of documents for the PLSA and CPLSA models for valid comparison. To capture the local lexical regularity, the topic models are interpolated (defined as + in the tables and figures) with a back-off trigram background (B) model. The trigram background model is trained on the '87-89 WSJ corpus using the back-off version of the Witten-Bell smoothing; 5K non-verbalized punctuation closed vocabulary and the cutoffs 1 and 3 on the bigram and trigram counts respectively are incorporated. The interpolation weights are computed by optimizing on the held-out data. We used the acoustic model from [21] in our experiments. The acoustic model is trained by using all WSJ and TIMIT [22] training data, the 40 phones set of the CMU dictionary [23], approximately 10000 tied-states, 32 gaussians per state and 64 gaussians per silence state. The acoustic waveforms are parameterized into a 39-dimensional feature vector consisting of 12 cepstral coefficients plus the 0^{th} cepstral, delta and delta delta coefficients, normalized using cepstral mean subtraction ($MFCC_{0-D-A-Z}$). The experiments are evaluated on the evaluation test, which is a total of 330 test utterances from the November 1992 ARPA CSR benchmark test data for vocabularies of 5K words [24, 25]. The results are described by using the perplexity and WER measurements.

7.2. Experimental Results

We tested the above LM approaches for various sizes of topics. We performed the experiments five times and the results are averaged. The perplexity results are described in Table 1.

Table 1. Perplexity results of the topic models

Language Model	20 Topics	40 Topics
Background (B)	69.0	69.0
B+PLSA	62.0	61.9
B+UBPLSA	59.0	58.7
B+CPLSA	57.5	55.8
B+DCPLSA	55.5	53.8

From Table 1, we can note that the perplexities are decreased with increasing topic size. The B+UBPLSA model outperforms the B+PLSA [7] models and the B+CPLSA model shows better results than the B+PLSA [7] and the B+UBPLSA [17] models respectively. The proposed B+DCPLSA model outperforms the B+PLSA [7], the B+UBPLSA [17] and the B+CPLSA [1] models respectively. The B+DCPLSA model yields perplexity reduction of about 19.6% (69.0 to 55.5), 10.5% (62.0 to 55.5), 5.9% (59.0 to 55.5) and 3.5% (57.5 to 55.5) for 20 topics and about 22.0% (69.0 to 53.8), 13.1% (61.9 to 53.8), 8.3% (58.7 to 53.8) and 3.6% (55.8 to 53.8) for 40 topics, over the background (B) model, B+PLSA model [7], the B+UBPLSA [17] and the B+CPLSA [1] approaches respectively.

We performed the paired t-test on the perplexity results of the above models with a significance level of 0.01. The p-values for different topic sizes are described in Table 2. From Table 2, we can note that all p-values are less than the significance level of 0.01. Therefore, the perplexity improvements of the proposed DCPLSA model over the CPLSA model [1] are statistically significant. Also, the CPLSA model [1] is statistically better than the UBPLSA model [17].

Table 2. p-values obtained from the paired t test on the perplexity results

Language Model	20 Topics	40 Topics
B+UBPLSA and B+CPLSA	6.0E-11	2.8E-14
B+CPLSA and B+DCPLSA	6.5E-12	3.1E-13

We evaluated the WER experiments using lattice rescoring. In the first pass, we used the back-off trigram background language model for lattice generation. In the second pass, we applied the interpolated form of the PLSA, UBPLSA, CPLSA and DCPLSA models for lattice rescoring. The experimental results are explained in Figure 2. From Figure 2, we can note that the proposed DCPLSA model yields significant WER reductions of about 25% (4.0% to 3.0%), 14.3% (3.5% to 3.0%), 9.1% (3.3% to 3%) and 6.25% (3.2% to 3.0%) for 20 topics and about 27.5% (4.0% to 2.9%), 17.1% (3.5% to 2.9%), 14.7% (3.4% to 2.9%) and 9.4% (3.2% to 2.9%) for 40 topics, over the background model, PLSA model [7], the UBPLSA [17] and the CPLSA [1] approaches respectively.

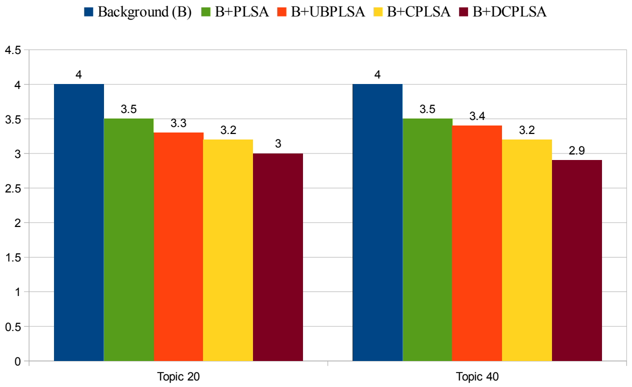

Fig. 2. WER results for different topic sizes

We also performed a paired t test on the WER results for the interpolated models with a significance level of 0.01. The p-values of the test are explained in Table 3. From Table 3, we can see that the p-values are smaller than the significance level of 0.01. Therefore, the WER improvements of the proposed DCPLSA model are statistically significant.

Table 3. p-values obtained from the paired t test on the WER results

Language Model	20 Topics	40 Topics
B+UBPLSA and B+CPLSA	4.7E-06	9.3E-06
B+CPLSA and B+DCPLSA	6.9E-06	1.5E-07

8. CONCLUSIONS

In this paper, we introduce a new document-specific CPLSA (DCPLSA) language model for speech recognition. This is an extended work of the CPLSA [1] model, which was investigated to overcome the limitations of an UBPLSA [17] model. As the UBPLSA model assigns probabilities to the seen bigrams only in the training phase, the model gives zero topic probabilities for some history context of the test document that are seen in the training set. Therefore, some of the bigram probabilities of the test document cannot be computed using the training model, which is not practical. The CPLSA model can compute all the possible bigram probabilities of the seen history context in the training set. It helps to find the topic weights of the unseen test documents correctly and hence gives the correct bigram probabilities to the test document. However, the CPLSA model trains the unigram probabilities for topics by using the sum of bigram events in all documents where the words may appear to describe different topics in different documents. This may yield inappropriate word probabilities for topics. We identify this problem in the CPLSA model and propose the DCPLSA model where document-wise unigram probabilities for topics are trained and have seen significant perplexity and WER reductions using the WSJ corpus over the other approaches.

9. REFERENCES

[1] M. A. Haidar and D. O'Shaughnessy, "Comparison of a bigram PLSA and a novel Context-based PLSA language model for speech recognition", in *Proc. of ICASSP*, pp. 8440-8444, 2013.

[2] R. Kuhn and R. D. Mori, "A Cache-Based Natural Language Model for Speech Recognition", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 12(6), pp. 570-583, 1990.

[3] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling", *Computer, Speech and Language*, vol. 10(3), pp. 187-228, 1996.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, vol. 41(6), pp. 391-407, 1990.

[5] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling", *IEEE Trans. on Speech and Audio Processing*, vol. 88, No. 8, pp. 1279-1296, 2000.

[6] T. Hofmann, "Probabilistic Latent Semantic Analysis", in *Proc. of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 289-296, 1999.

[7] D. Gildea and T. Hofmann, "Topic-Based Language Models Using EM", in *Proc. of EUROSPEECH*, pp. 2167-2170, 1999.

[8] D. M. Blei, A. Y.Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[9] D. Mrva and P. C. Woodland, "A PLSA-based Language Model for conversational telephone speech", in *Proc. of ICSLP*, pp. 2257-2260, 2004.

[10] Y.-C. Tam and T. Schultz, "Dynamic Language Model Adaptation Using Variational Bayes Inference", in *Proc. of INTERSPEECH*, pp. 5-8, 2005.

[11] Y.-C. Tam and T. Schultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals", in *Proc. of INTERSPEECH*, pp. 2206-2209, 2006.

[12] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using N-gram weighting", in *Proc. of CCECE*, pp. 857-860, 2011.

[13] M. A. Haidar and D. O'Shaughnessy, "LDA-based LM adaptation using latent semantic marginals and minimum discrimination information", in *Proc. of EUSIPCO*, pp. 2040-2044, 2012.

[14] M. A. Haidar and D. O'Shaughnessy, "Topic N-gram Count Language Model for Speech Recognition", in *Proc. of IEEE SLT Workshop*, pp. 165-169, 2012.

[15] H. M. Wallach, "Topic Modeling: Beyond bag-of-words", in *Proc. of ICML*, pp. 977-984, 2006.

[16] J. Nie, R. Li, D. Luo, and X. Wu, "Refine bigram PLSA model by assigning latent topics unevenly", in *Proc. of the IEEE Workshop on ASRU*, pp. 141-146, 2007.

[17] M. Bahrani and H. Sameti, "A New Bigram PLSA Language Model for Speech Recognition", *Euraship Journal on Signal Processing*, pp. 1-8, 2010.

[18] "CSR-II (WSJ1) Complete", Linguistic Data Consortium, Philadelphia, 1994.

[19] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[20] H. Yamamoto, S. Isogai, and Y. Sagisaka, "Multi-class composite n-gram language model", *Speech Communication*, vol. 41, pp. 369-379, 2003.

[21] K. Vertanen, "HTK Wall Street Journal Training Recipe",http://www.inference.phy.cam.ac.uk/kv227/htk/

[22] J. S. Garofolo, et al,"TIMIT Acoustic-Phonetic Continuous Speech Corpus" Linguistic Data Consortium, Philadelphia, 1993.

[23] "The Carnegie Mellon University (CMU) Pronounciation Dictionary", http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[24] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus", in *Proc. of ICSLP*, pp. 899-902,1992.

[25] P.C. Woodland, J.J. Odell, V. Valtchev and S.J. Young, "Large Vocabulary Continuous Speech Recognition Using HTK", in *Proc. of ICASSP*, pp. II:125-128, 1994.