EUSIPCO 2013 1569741253

PLSA ENHANCED WITH A LONG-DISTANCE BIGRAM LANGUAGE MODEL FOR SPEECH RECOGNITION

Md. Akmal Haidar and Douglas O'Shaughnessy

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

ABSTRACT

We propose a language modeling (LM) approach using background n-grams and interpolated distanced n-grams for speech recognition using an enhanced probabilistic latent semantic analysis (EPLSA) derivation. PLSA is a bag-of-words model that exploits the topic information at the document level, which is inconsistent for the language modeling in speech recognition. In this paper, we consider the word sequence in modeling the EPLSA model. Here, the predicted word of an n-gram event is drawn from a topic that is chosen from the topic distribution of the $(n$-1$)$ history words. The EPLSA model cannot capture the long-range topic information from outside of the n-gram event. The distanced n-grams are incorporated into interpolated form (IEPLSA) to cover the long-range information. A cache-based LM that models the re-occurring words is also incorporated through unigram scaling to the EPLSA and IEPLSA models, which models the topical words. We have seen that our proposed approaches yield significant reductions in perplexity and word error rate (WER) over a PLSA based LM approach using the Wall Street Journal (WSJ) corpus.

Index Terms— language model, topic model, speech recognition, cache-based LM, long-distance n-grams

1. INTRODUCTION

Statistical n-gram LMs play a vital role for speech recognition and many other applications. They use the local context information by modeling text as a Markovian Sequence. However, the n-gram LMs suffer from shortages of long-range information, which degrade performance. Cache-based LM was one of the earliest efforts to capture the long-range information. Here, the model increases the probability of the words that appear earlier in a document when predicting the next word [1]. Recently, latent topic modeling techniques have been used broadly for topic based language modeling to compensate for the weaknesses of the n-gram LMs. Several techniques such as Latent Semantic Analysis (LSA) [2], PLSA [3], and latent Dirichlet allocation (LDA) [4] have been studied to extract the latent semantic information from a training corpus. All these methods are based on a bag-of-words assumption. LSA performs word-document matrix decomposition to extract the semantic information for different words and documents. In PLSA and LDA, semantic properties of words and documents can be shown in probabilistic topics. The PLSA latent topic parameters are trained by maximizing the likelihood of the training data using an expectation maximization (EM) procedure and have been successfully used for speech recognition [3, 5]. The LDA model has been used successfully in recent research work for LM adaptation [6, 7]. A bigram LDA topic model, where the word probabilities are conditioned on their preceding context and the topic probabilities are conditioned on the documents, has been recently investigated [8]. A similar model but in the PLSA framework, called a bigram PLSA model, was introduced recently [9]. An updated bigram PLSA model was proposed in [10] where the topic is further conditioned on the bigram history context. A topic-based language model was proposed where the topic information was obtained from n-gram history through Dirichlet distribution [11] and from long-distance history (topic cache) through multinomial distributions [12].

In [13], a PLSA technique enhanced with long-distance bigrams was used to incorporate the long-term word dependencies in determining word clusters. This motivates us to present LM approaches for speech recognition using distanced n-grams. In this paper, we use default n-grams using enhanced PLSA derivation to form the EPLSA n-gram model. Here, the observed n-gram events contain the history words and the predicted word. The EPLSA model extracts the topic information from history words and the current word is then predicted based on the topic information of the history words. However, the EPLSA model does not capture the topic information from outside of the n-gram events. We propose interpolated distanced n-grams (IEPLSA) and cache based models to capture the long-term word dependencies into the EPLSA model. The n-gram probabilities of the IEPLSA model are computed by mixing the component distanced word probabilities for topics and the interpolated topic information for histories. Furthermore, a cache-based LM is incorporated into the EPLSA and IEPLSA models as the cache-based LM models a different part of the language than EPLSA/IEPLSA models.

The rest of this paper is organized as follows. Section 2 is used to review the PLSA. The proposed EPLSA and IEPLSA models are described in section 3. In section 4, a comparison

of PLSA, bigram PLSA, EPLSA and IEPLSA models is illustrated. The unigram scaling of the cache-based model to the topic models is explained in section 5. Section 6 is used to describe the experiments. Finally, the conclusions are explained in section 7.

2. PLSA MODEL

The PLSA model [3] can be described in the following procedure. First a document D_j $(j = 1, 2, \ldots, N)$ is selected with probability $P(D_j)$. A topic z_k $(k = 1, 2, \ldots, K)$ is then chosen with probability $P(z_k|D_j)$, and finally a word w_i $(i = 1, 2, \ldots, V)$ is generated with probability $P(w_i|z_k)$. Here, the observed variables are w_i and D_j whereas the unobserved variable is z_k. The joint distribution of the observed data can be described as:

$$P(D_j, w_i) = P(D_j)P(w_i|D_j) = P(D_j)\sum_{k=1}^{K} P(z_k|D_j)P(w_i|z_k), \tag{1}$$

where the word probability $P(w_i|D_j)$ can be computed as:

$$P(w_i|D_j) = \sum_{k=1}^{K} P(z_k|D_j)P(w_i|z_k). \tag{2}$$

The model parameters $P(w_i|z_k)$ and $P(z_k|D_j)$ are computed by using the EM algorithm [3].

3. PROPOSED EPLSA AND IEPLSA MODELS

3.1. EPLSA

Representing a document D_j as a sequence of words, the joint distribution of the document and the previous $(n\text{-}1)$ history words h of the current word w_i can be described as [13]:

$$P(D_j, h) = P(h) \prod_{w_i \epsilon D_j} P_d(w_i|h), \tag{3}$$

where $P_d(w_i|h)$ is the distanced n-gram model. Here, d represents the distance between the words in the n-grams. Therefore, the probability $P_d(w_i|h)$ can be computed similar to the PLSA derivation [3, 13]. For $d = 1$, $P_d(w_i|h)$ is the default background n-gram and we define it as the enhanced PLSA (EPLSA) model. The graphical model of the EPLSA model can be described in Figure 1. The equations for the EPLSA model are:

$$P_{EPLSA}(w_i|h) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|h), \tag{4}$$

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

$$P(z_k|h, w_i) = \frac{P(w_i|z_k)P(z_k|h)}{\sum_{k'=1}^{K} P(w_i|z_{k'})P(z_{k'}|h)}, \tag{5}$$

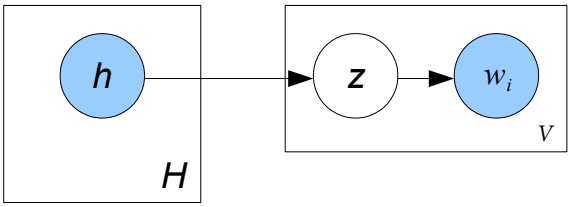

Fig. 1. *The graphical model of the EPLSA model. The shaded circle represents the observed variables. H and V describe the number of histories and the size of vocabulary.*

M-step:

$$P(w_i|z_k) = \frac{\sum_h n(h, w_i)P(z_k|h, w_i)}{\sum_{i'}\sum_h n(h, w_{i'})P(z_k|h, w_{i'})}, \tag{6}$$

$$P(z_k|h) = \frac{\sum_{i'} n(h, w_{i'})P(z_k|h, w_{i'})}{\sum_{k'}\sum_{i'} n(h, w_{i'})P(z_{k'}|h, w_{i'})}. \tag{7}$$

3.2. IEPLSA

The EPLSA model does not capture the long-distance information. To incorporate the long-range characteristics, we used the distanced n-grams in the EPLSA model. Incorporating the interpolated distance n-grams in the EPLSA, the model can be written as [13]:

$$P_{IEPLSA}(w_i|h) = \sum_{k=1}^{K}[\sum_{d} \lambda_d P_d(w_i|z_k)]P(z_k|h), \tag{8}$$

where λ_d are the weights for each component probability estimated on the held-out data using the EM algorithm and $P_d(w_i|z_k)$ is the word probabilities for topic z_k obtained by using the distanced n-grams in the IEPLSA training. d represents the distance between words in the n-gram events. $d = 1$ describes the default n-grams. For example, the distanced n-grams of the phrase "Speech in Life Sciences and Human Societies" are described in Table 1 for the distance $d = 1, 2$.

Table 1. *Distanced n-grams for the phrase "Speech in Life Sciences and Human Societies"*

Distance	Bigrams	Trigrams
d=1	*Speech in, in Life, Life Sciences, Sciences and, and Human, Human Societies*	*Speech in Life, in Life Sciences, Life Sciences and, Sciences and Human, and Human Societies*
d=2	*Speech Life, in Sciences, Life and, Sciences Human, and Societies*	*Speech Life and, in Sciences Human, Life and Societies*

The parameters of the IEPLSA model can be computed as: E-step:

$$P_d(z_k|h,w_i) = \frac{P_d(w_i|z_k)P(z_k|h)}{\sum_{k'=1}^{K} P_d(w_i|z_{k'})P(z_{k'}|h)}, \quad (9)$$

M-step:

$$P_d(w_i|z_k) = \frac{\sum_h n_d(h,w_i)P_d(z_k|h,w_i)}{\sum_{i'}\sum_h n_d(h,w_{i'})P_d(z_k|h,w_{i'})}, \quad (10)$$

$$P(z_k|h) = \frac{\sum_{i'}\sum_d \lambda_d n_d(h,w_{i'})P_d(z_k|h,w_{i'})}{\sum_{k'}\sum_{i'}\sum_d \lambda_d n_d(h,w_{i'})P_d(z_k|h,w_{i'})}. \quad (11)$$

4. COMPARISON OF PLSA, PLSA BIGRAM AND EPLSA/IEPLSA

PLSA [3] is a bag-of-words model where the document probability is computed by using the topic structure at the document level. This is inappropriate for the language model in speech recognition. PLSA bigram models were introduced where the bigram probabilities for each topic are modeled and the topic is conditioned on the document [9] or bigram history and the document [10]. In either approach, the models require V distributions for each topic, where V is the size of the vocabulary. Therefore, the size of the parameters grows exponentially with increasing n-gram order. In contrast, the EPLSA/IEPLSA models developed the word distributions given the history words. The history information is used to form the topic distributions, then the probability of the predicted word is computed given the topic information of the histories. Therefore, the parameter number grows linearly with V [11].

5. INCORPORATING THE CACHE MODEL THROUGH UNIGRAM SCALING

A Cache-based language model was used to increase the probability of words appearing in a document that are likely to re-occur in the same document. The unigram cache model for a given history $h_c = w_{i-M}, \ldots, w_i$, where M is the cache size, is defined as:

$$P_{cache}(w_i) = \frac{n(w_i,h_c)}{n(h_c)} \quad (12)$$

where $n(w_i,h_c)$ is the number of occurrences of the word w_i within h_c and $n(h_c) <= M$ is the number of words within h_c that belongs to the vocabulary V [14].

The EPLSA/IEPLSA models capture topical words. The models are then interpolated with a background n-gram model to capture the local lexical regularities as:

$$P_L(w_i|h) = (1-\gamma)P_{EPLSA/IEPLSA}(w_i|h) + \gamma P_{Background}(w_i|h). \quad (13)$$

As the cache-based LM (i.e., models re-occurring words) is different from the background model (i.e., models short-range information), EPLSA and IEPLSA models (i.e., model topical words), we can integrate the cache model to adapt the $P_L(w_i|h)$ through unigram scaling as [15, 16]:

$$P_{Adapt}(w_i|h) = \frac{P_L(w_i|h)\delta(w_i)}{Z(h)}, \quad (14)$$

with

$$Z(h) = \sum_{w_i} \delta(w_i).P_L(w_i|h). \quad (15)$$

where $Z(h)$ is a normalization term, which guarantees that the total probability sums to unity, $P_L(w_i|h)$ is the interpolated model of the background and the EPLSA/IEPLSA model and $\delta(w_i)$ is a scaling factor that is usually approximated as:

$$\delta(w_i) \approx (\frac{\alpha P_{cache}(w_i) + (1-\alpha)P_{Background}(w_i)}{P_{Background}(w_i)})^{\beta}, \quad (16)$$

where β is a tuning factor between 0 and 1. In our experiments we used the value of β as 1. We used the same procedure as [15] to compute the normalization term. To do this, an additional constraint is employed where the total probability of the observed transitions is unchanged:

$$\sum_{w_i:observed(h,w_i)} P_{Adapt}(w_i|h) = \sum_{w_i:observed(h,w_i)} P_L(w_i|h).$$

The model $P_L(w_i|h)$ has standard back-off structure and the above constraint, so the model $P_{Adapt}(w_i|h)$ has the following recursive formula:

$$P_{Adapt}(w_i|h) = \begin{cases} \frac{\delta(w_i)}{Z_o(h)}.P_L(w_i|h) & \text{if } (h,w_i) \text{ exists} \quad (17) \\ b(h).P_{Adapt}(w_i|\hat{h}) & \text{otherwise} \quad (18) \end{cases}$$

where

$$Z_o(h) = \frac{\sum_{w_i:observed(h,w_i)} \delta(w_i).P_L(w_i|h)}{\sum_{w_i:observed(h,w_i)} P_L(w_i|h)} \quad (19)$$

and

$$b(h) = \frac{1 - \sum_{w_i:observed(h,w_i)} P_L(w_i|h)}{1 - \sum_{w_i:observed(h,w_i)} P_{Adapt}(w_i|\hat{h})} \quad (20)$$

where $b(h)$ is the back-off weight of the context h to ensure that $P_{Adapt}(w_i|h)$ sums to unity. $\hat{h}$ is the reduced word history of h. The term $Z_o(h)$ is used to do normalization similar to Equation 15 except the summation is considered only on the observed alternative words with the equal word history h in the LM [6].

6. EXPERIMENTS

6.1. Data and experimental setup

LM adaptation approaches are evaluated using the Wall Street Journal (WSJ) corpus [17]. The SRILM toolkit [18] and the HTK toolkit [19] are used for generating the LMs and computing the WER respectively. The '87-89 WSJ corpus is used to train language models. The models are trained using the WSJ 5K non-verbalized punctuation closed vocabulary. A tri-gram background model is trained using the modified Kneser-Ney smoothing incorporating the cutoffs 1 and 3 on the bi-gram and tri-gram counts respectively. To reduce the computational and memory requirements using MATLAB, we trained only the bi-gram EPLSA and IEPLSA models. For IEPLSA models, we considered bigrams for $d = 1, 2$. A fixed cache size of $M = 400$ is used for the cache-based LM. The acoustic model from [20] is used in our experiments. The acoustic model is trained by using all WSJ and TIMIT [21] training data, the 40 phone set of the CMU dictionary [22], 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 $(MFCC_{0-D-A-Z})$. We evaluated the cross-word models. The values of the word insertion penalty, beam width, and the language model scale factor are -4.0, 350.0, and 15.0 respectively [20]. The development and the evaluation test sets are the $\mathbf{si\_dt\_05.odd}$ (248 sentences from 10 speakers) and the Nov'93 Hub 2 5K test data from the ARPA November 1993 WSJ evaluation (215 sentences from 10 speakers) [17, 23]. The interpolation weights λ_d, γ and α are computed using the $compute\text{-}best\text{-}mix$ program from the SRILM toolkit. They are tuned on the development test set. The results are noted on the evaluation test set.

6.2. Experimental Results

We used the folding-in procedure [3] to compute the PLSA, EPLSA and IEPLSA model probabilities. We keep the unigram (Equations 2, 4 and 8) probabilities for topics of PLSA, EPLSA and IEPLSA, and λ_d of component probabilities for IEPLSA unchanged, and used them to compute $P(z_k|D)$ for the test document D of the PLSA model and $P(z_k|h)$ for the test document histories of the EPLSA and IEPLSA models. The language models for PLSA, EPLSA and IEPLSA are then computed using (Equations 2, 4 and 8). The remaining zero probabilities of the obtained matrix $P_{EPLSA/IEPLSA}(w_i|h)$ are computed by using back-off smoothing. The EPLSA and IEPLSA models are interpolated with a back-off trigram background model to capture the local lexical regularities. Furthermore, a cache-based LM that models re-occurring words is integrated through unigram scaling (Equations 17 and 18) with the EPLSA and IEPLSA models, which describe topical words. We compared our approaches with a PLSA based LM approach [3] using unigram scaling where the PLSA unigrams are used in place of cache unigrams in Equation 16 and denoted as Background*PLSA.

We tested the proposed approach for various sizes of topics. The perplexity results are described in Table 2. From Table 2, we can note that all the models outperform the background model and the performances are better with increasing topics. The proposed EPLSA and IEPLSA models outperform the PLSA models in every form (stand-alone, interpolated, unigram scaling).

Table 2. *Perplexity results of the language models*

Language Model	40 Topics	80 Topics
Background	70.26	70.26
PLSA	517.77	514.78
EPLSA	192.91	123.32
IEPLSA	101.19	93.02
Background*PLSA	66.63	66.50
Background+EPLSA	62.92	59.74
Background+IEPLSA	55.12	55.10
(Background+EPLSA)*CACHE	57.98	55.06
(Background+IEPLSA)*CACHE	50.71	50.69

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 LM adaptation approaches for lattice rescoring. The experimental results are explained in Figure 2. From Figure 2, we can note that the proposed EPLSA model yields significant WER reductions of about 10.93% (7.59% to 6.76%) and 8.64% (7.40% to 6.76%) for 40 topics, and about 15.41% (7.59% to 6.42%) and 13.00% (7.38% to 6.41%) for 80 topics, over the background model and the PLSA [3] approaches respectively. For the IEPLSA models, the WER reductions are about 19.50% (7.59% to 6.11), 17.43% (7.40% to 6.11%), and 9.61% (6.76% to 6.11%) for 40 topics and about 20.28% (7.59% to 6.05), 18.02% (7.38% to 6.05%), and 5.76% (6.42% to 6.05%) for 80 topics, over the background model, PLSA [3] and EPLSA approaches respectively. The integration of cache based models improves the performance as it carries different information (captures the dynamics of word occurrences in a cache) than the EPLSA and IEPLSA approaches. The cache unigram scaling of the IEPLSA approach gives 6.74% and 1.63% WER reductions over the cache unigram scaling of the EPLSA approach for 40 and 80 topics respectively. We can note that the addition of cache models improves the performance of EPLSA (6.76% to 6.52% for 40 topics and 6.42% to 6.13% for 80 topics) more than for IEPLSA (6.11% to 6.08% for 40 topics and 6.05% to 6.03% for 80 topics). This might be due to the fact that the IEPLSA approach captures long-range information using the

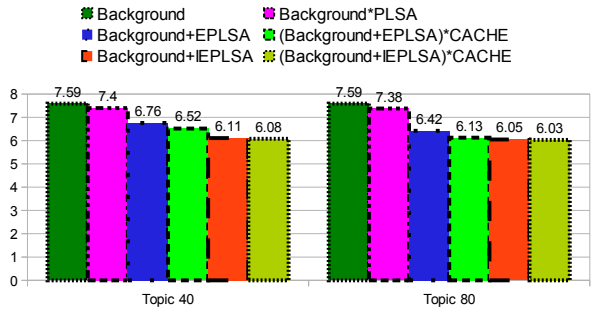

Fig. 2. *WER Results of the Language Models*

interpolated distanced bigrams.

7. CONCLUSIONS

In this paper, the background n-grams and the interpolated distanced n-grams are used to derive the EPLSA and IEPLSA models respectively for speech recognition. The EPLSA model extracted the topic information from the $(n\text{-}1)$ history words in calculating the n-gram probabilities. However, it does not capture the long-range semantic information from outside of the n-gram events. The IEPLSA model overcomes the shortcomings of EPLSA by using the interpolated long-distance n-grams that capture the long-term word dependencies. Using the IEPLSA, the topic information for the histories are trained using the interpolated distanced n-grams. The model probabilities are computed by weighting the component word probabilities for topics and the interpolated topic information for the histories. We have seen that the proposed EPLSA and IEPLSA approaches yield significant perplexity and WER reductions over the PLSA-based LM approach using the WSJ corpus. Moreover, we incorporate a cache-based model into the EPLSA and IEPLSA models using unigram scaling for adaptation and have seen improved performances. However, cache unigram scaling of the EPLSA gives much better performance over the EPLSA than the cache unigram scaling of the IEPLSA over the IEPLSA. This proves that the IEPLSA approach captures long-range information of the language.

8. REFERENCES

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

[2] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling", in *Proc. of the IEEE*, vol. 88, No. 8, pp. 1279-1296, 2000.

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

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

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

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

[7] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation and Dynamic Marginals", in *Proc. of EUSIPCO*, pp. 1480-1484, 2011.

[8] H. M. Wallach, "Topic Modeling: Beyond bag-of-words", in *Proc. of the 23rd International Conference of Machine Learning (ICML'06)*, pp. 977-984, 2006.

[9] 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.

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

[11] J-T. Chien and C-H. Chueh, "Latent Dirichlet Language Model for Speech Recognition", in *Proc. of IEEE SLT workshop*, pp. 201-204, 2008.

[12] C-H. Chueh and J-T. Chien, "Topic Cache Language Model for Speech Recognition", in *Proc. of ICASSP*, pp.5194-5197, 2010.

[13] N. Bassiou and C. Kotropoulos, "Word Clustering PLSA enhanced with Long Distance Bigrams", in *Proc. of International Conferance on Pattern Recognition*, pp. 4226-4229, 2010.

[14] A. Vaiciunas and G. Raskinis, "Cache-based Statistical Language Models of English and Highly Inflected Lithuanian, in *Informatica*, Vol. 17(1), pp. 111-124, 2006.

[15] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals", in *Proc. of EUROSPEECH*, pp. 1971-1974, 1997.

[16] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Topic Dependent Class-Based N-gram Language Model", in *IEEE Trans. on Audio, Speech and Language Processing*, pp. 1513-1525, Vol. 20. No.5, 2012.

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

[18] A. Stolcke, "SRILM- An Extensible Language Modeling Toolkit", in *Proc. of ICSLP*, vol. 2, pp. 901-904, 2002.

[19] S. Young, P. Woodland, G. Evermann and M. Gales, "The HTK toolkit 3.4.1", http://htk.eng.cam.ac.uk/, Cambridge Univ. Eng. Dept. CUED.s

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

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

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

[23] 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.