

IMPROVED LANGUAGE MODEL ADAPTATION USING EXISTING AND DERIVED EXTERNAL RESOURCES

Pi-Chuan Chang and Lin-Shan Lee

Dept. of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

pcchang@speech.ee.ntu.edu.tw

ABSTRACT

Adaptation of language models to obtain better parameters for the topics addressed by the spoken documents to be recognized has been a key issue for speech recognition. In this paper, we propose to collect existing as well as derived external resources for improved language model adaptation. The derived external resources are those retrieved based on the baseline transcriptions for the input spoken documents from the Internet using some search engine. The design of queries for such purposes are also analyzed in this paper, in which the special structure of Chinese language is considered. The obtained existing and derived external resources are then used in the model adaptation under a *Clustering-Classification* framework. Very encouraging results were obtained in the preliminary experiments with two test sets: broadcast news and interview recording.

1. INTRODUCTION

Language modeling has been one of the few core components in speech recognition systems. Among the various approaches for language modeling, the statistical N -gram has been the most popular and successful. To train an appropriate N -gram language model for a target speech recognition task, large quantities of text corpus, which is homogeneous (in topic or style) to the target task, must be available. Language model adaptation techniques, on the other hand, serve the purpose of tuning the language models with only limited quantities of adaptation text corpus homogeneous to the target task.

In this paper, we aim at the use of two different types of external resources as the adaptation text corpus: existing and derived. Two test sets will be recognized in the experiments presented below: the broadcast news and an interview recording. The existing external resource used is collected from the Yahoo! News [1], while the derived resource is the documents retrieved from the Internet by dynamically querying a search engine based on the baseline transcriptions of the spoken documents to be recognized. These existing

and derived resources are used with a proposed *Clustering-Classification* framework, in which two commonly used machine learning techniques are integrated, the document clustering and document classification, as illustrated in Fig. 1. Given a large collection of text corpus C , we first apply the document clustering technique to divide the corpus C into N homogeneous sub-corpora C_1, C_2, \dots, C_N , each with some topic T_1, T_2, \dots, T_N . Then we take these sub-corpora as the training corpus for document classification. Some document classification techniques are used to train document classifier, and then the target spoken documents, after processed by a baseline recognizer, can be classified among the topics T_1, T_2, \dots, T_N . After the *Clustering-Classification* process, we adapt our background language model with the framework shown in Fig. 2. A set of foreground models are first trained with the sub-corpora C_1, C_2, \dots, C_N , each for the topics T_1 to T_N respectively. They are then integrated with a background language model. The appropriate integrated models is then chosen for the target spoken documents based on the homogeneity with respect to the sub-corpora C_1, \dots, C_N , as was performed during the document classification process in Fig. 1.

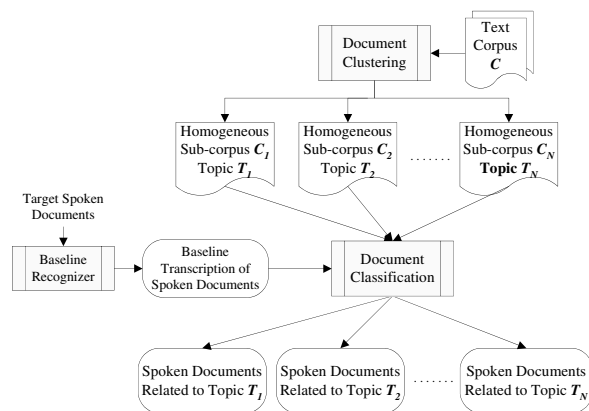


Fig. 1. The *Clustering-Classification* framework for obtaining homogeneous data

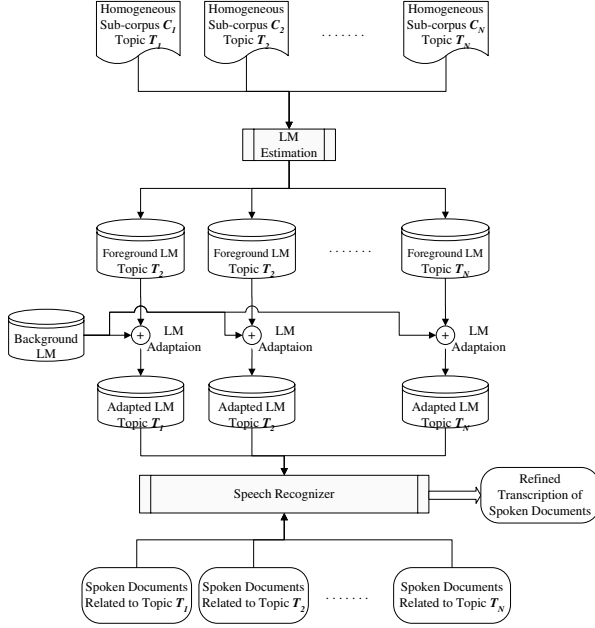


Fig. 2. Language model adaptation framework

In the literature, information retrieval and machine learning techniques have been proposed in order to select homogeneous data [2, 3]. The proposed *Clustering-Classification* framework has the similar flavor, however we put more emphasis on the influence of different external resources. Also, we proposed different query-construction approaches for deriving resources and gave further discussions.

In what follows, Section 2 presents the *Clustering-Classification* framework in more detail, and how the language adaptation works with this framework. Section 3 describes how the existing and derived resources can be obtained, including different query-construction approaches for retrieving derived resources. Section 4 analyzes the characteristics of the existing and derived external resources, based on the experimental results. Section 5 gives the concluding remarks.

2. THE CLUSTERING-CLASSIFICATION AND LANGUAGE MODEL ADAPTATION FRAMEWORK

Although it seems that more training data is always better for language modeling, it is really not the case. Large quantities of training data with variable quality (or heterogeneity) may actually hurt the purity of the language model and therefore the system performance. It is thus desirable to use corpora which are more homogeneous in quality but less in quantity, rather than those which are large in quantity but heterogeneous in quality. This is the basic idea of the *Clustering-Classification* framework, i.e., to divide the large corpus C

into N more sub-corpora C_1, \dots, C_N , associate the target spoken document with one of these sub-corpus, and then adapt the language model to a more appropriate direction for the target spoken document.

2.1. Document Clustering

The document clustering is to divide a large corpus C into N more homogeneous sub-corpora C_1, \dots, C_N , each with topics T_1, \dots, T_N respectively. The document clustering algorithm used in this paper is a variant of K-means: *bisecting K-means*. *Bisecting K-means* is found to be more efficient and able to produce clusters of relatively uniform size [4]. The *bisecting K-means* algorithm takes several steps. In each iteration, first pick up a cluster for splitting. Then we bisect the cluster into 2 sub-clusters using the basic K-means algorithm. Repeat this process until the desired number of clusters is reached. The *bisecting K-means* algorithm is implemented in CLUTO (A Clustering Toolkit) [5], which we used as a component of our overall framework.

2.2. Document Classification

The objective of document classification is to assign a target spoken document (with some baseline transcription obtained with some baseline recognizer) into one or more predefined topics (T_1, \dots, T_N as mentioned above). A Naive Bayes classifier was used in this paper for this purpose. The assumption behind Naive Bayes classifier is that the features chosen should be independent of each other, but this seems unreasonable when we use words as features. However, it was argued that the lack of independence does not necessarily mean that the classifier will perform poorly [6]. Furthermore, the computation efficiency of this classifier has made it suitable for large-scale tasks.

The Naive Bayes classifier is constructed from the training data. It is to estimate the probability for each class given the document feature vector. Using the Bayes theorem, we can estimate the probability that a document with a feature vector d belongs to a topic T_j as:

$$P(T_j|d) = \frac{P(T_j)P(d|T_j)}{P(d)} \quad (1)$$

When choosing among the topics, we only need to find the topic T_j^* which gives the maximum score, and $P(d)$ is the same across all topics, thus

$$T_j^* = \operatorname{argmax}_{T_j} P(T_j|d) = \operatorname{argmax}_{T_j} P(T_j)P(d|T_j) \quad (2)$$

In our experiments, the program *Rainbow* implementing a Naive Bayes classifier in the Bow toolkit was used for document classification purpose [7]. The program was slightly modified to deal with Chinese characters.

2.3. The Language Model Adaptation

With the document clustering, N homogeneous sub-corpora C_1, \dots, C_N are obtained each with a specific topic T_1, \dots, T_N . With the document classification, the target spoken document is associated to one or more topics. As shown in Fig. 2, we can now integrate the clustered sub-corpora and the classified topic-labeled spoken documents in the overall language model adaptation framework.

We first estimate the foreground language models for each clustered sub-corpus or each topic, and then interpolated each of them with a background language model. These foreground language models give greater emphasis on the data that better matches the topics associated with the target spoken documents, while the background language model gives better balance on common usage of the language.

3. EXISTING AND DERIVED EXTERNAL RESOURCES

The text corpus C in Fig. 1 used to generate the homogeneous sub-corpora C_1, \dots, C_N is the primary resource for the language model adaptation. It can be constructed with various external resources. Here two types of external resource are discussed: existing and derived. By existing resources we mean those already collected and made available. By derived resource we mean those obtained based on the input baseline transcription for the spoken documents, or the resources derived with some knowledge about the input speech.

3.1. Existing External Resource

The existing external resource used in this paper is collected from Yahoo! News [1], which continuously collects major news sources of Taiwan. This site constantly stores up news articles, which are similar to our first test set of broadcast news in both topic and style. We therefore take it as the existing resource in the following experiments.

3.2. Derived External Resource

Given the input speech (and its baseline transcription, which may include quite many errors due to the poor language model in the baseline recognizer), better resources which can be used to update the language model may be derived. In this sense, the Internet may be the most complete resource, and the most direct approach to derive the desired resource helpful to the target spoken document is to retrieve the relevant documents via the search engine Google [8] by entering queries related to the baseline transcription of the input speech documents. It may take some time to obtain such derived resources, thus the approach discussed here may be inadequate for some real-time applications. However, there also exist many tasks for which the correct transcriptions of

spoken documents are very helpful even if obtained some time later. So the approach discussed here may be applied in those cases. The key issue of gathering such text data over the Internet is then to construct the appropriate queries to be entered to the search engine. Due to the imperfect nature of speech recognition transcriptions, two different query-construction approaches were tested here.

3.2.1. Disjunctive 3-word queries

All the tests performed in this study are for Mandarin Chinese. Since there is no clear boundaries between Chinese words and new Chinese words can be arbitrarily constructed by concatenating different characters, new named entities in Chinese which are not included in the original lexicon for the baseline recognizer are often segmented into several mono-character words, because all characters are also included in the original lexicon as mono-character words. Therefore, some metrics discussed in the literature for choosing “meaningful” words may not apply here.

Considering both the correctness and size of the data retrieved in the preliminary tests, one naive query-construction approach is to select every three adjacent words in the baseline transcription of the spoken documents, and enter them as a disjunctive 3-word query to Google. The reason we use disjunctive (*ORed*) queries instead of conjunctive (*ANDed*) queries is that these words are in the baseline transcription and thus some of the words may be incorrect, and conjunction makes the restriction tighter. As a result, the data we collected may be too few or very irrelevant if conjunctive queries are used.

3.2.2. Conjunctive Queries Filtered by Confidence Measure

As mentioned above, the baseline transcriptions produced by the baseline recognizer always include some errors. It is therefore reasonable to use some filtering mechanism, to make sure that most of the text we collect will not be misleading. A query-construction approach based on some confidence measure is thus proposed here. The confidence measure $C(w)$ for a word hypothesis w used in this paper is the *posterior word probabilities*, which can be computed by summing over all histories and futures of w [9],

$$\begin{aligned}
 C(w) &= p(w|x_1^T) \\
 &= \sum_{h_2^{m-1}} \sum_{f_1^{m-2}} \frac{\Phi(h_2^{m-1}; w) \cdot \Psi(w; f_1^{m-2})}{p(x_1^T) \cdot (x_\tau^t | w)} \\
 &\quad \cdot \prod_{n=1}^{m-2} p(f_n | h_{n+1}^{m-1} w f_1^{n-1})
 \end{aligned} \tag{3}$$

where $\Phi(h_2^{m-1}; w)$ is the forward probability that the last hypothesis of a sequence of m word hypotheses is w and

h_2^{m-1} is the history of w , and $\Psi(w; f_1^{m-2})$ is the backward probability that the first hypothesis of a sequence of m word hypotheses is w and f_1^{m-2} is the future of w .

We tag each word hypothesis in the word graphs for baseline transcriptions with its confidence measure, find the successive word sequence with high enough confidence measure, and enter these sequences as a conjunctive query to Google. As an example shown in Fig. 3, the solid edges are the word hypotheses with higher confidence measures. In this example, the word sequence ABC and the word D are selected as queries.

Since we already selected the word hypotheses with higher confidence measures, the queries of variable number of words would be more trustworthy than simply windowing through the transcriptions as in the disjunctive 3-word queries.

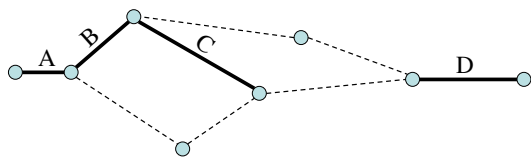


Fig. 3. Query construction based on the word graph by filtering with confidence measures

4. EXPERIMENTAL RESULTS

Extensive experiments were performed to analyze the achievable improvements for the approaches proposed here. Since the goal is to have better speech recognition and all the test were on Mandarin Chinese, the performance metric used here is the character error rate (CER) for Chinese characters and syllable error rate (SER) for Mandarin syllables.

For all the experiments presented below, the large vocabulary recognizer for Mandarin Chinese developed at National Taiwan University [10] was used. The acoustic models consist of 151 Initial-Final sub-syllabic units. (112 Initials, 38 Finals, and a silence). Here Initial is the initial consonant of a syllable, and Final is everything in the syllable following the Initial, primarily the vowel/diphthong part plus optional medial and nasal ending. Each Initial model has three states and each Final model has four states. In each state, 16 Gaussian-mixtures were used. The baseline background language model is a trigram based on a lexicon of 60K words and estimated on a 40M-character text news corpus collected from the Central News Agency (CNA) at Taipei during 1997-1999, smoothed with Good-Turing discounting by SRI Language Modeling Toolkit (SRILM) [11]. Two test sets were used. The first is the Chinese broadcast news (CBN) recorded from News98 radio station at Taipei

[12] in September 2002. It includes news of 18 days with total length of 3.7 hours. The second test set is an interview recording speech corpus with a total length of 34 minutes, also collected from CNA. But the topic of the interview has nothing to do with the news, so this test set is mismatched with our baseline background language model as well as the existing external resource in both topic and style.

4.1. Experiments on the First Test Set of Broadcast News

The first experiment was performed on the test set of broadcast news. The results with the existing and derived external resources were compared based on the *Clustering-Classification* framework. The results are listed in Table 1, and plotted as functions of number of clusters in Fig. 4. For the case of derived external resources, the disjunctive 3-word queries mentioned in Section 3.2.1 were used here for simplicity. The approach of conjunctive queries filtered by confidence measure will be discussed later on. In Fig. 4, the two solid curves show the CER and SER for the language model adapted by derived external resource, while the two dotted curved those adapted by existing external resources. It should be noted that the CER is much higher than the SER. This normal for transcriptions of broadcast news in Mandarin Chinese. Due to the many new named entities and out of vocabulary (OOV) words in the news, CER is always higher even if the SER can be much lower.

4.1.1. The Effect of the Clustering-Classification Framework

The results listed in Table 1 shows how the *Clustering-Classification* framework works on both existing and derived external resources. We can see that by using the external resources, no matter existing or derived, the character and syllable error rates were both improved. Also, with the *Clustering-Classification* framework, both the existing and derived resources achieved better performance. It is also observed that for the broadcast news test data used here, the best number of clusters for existing resource is 200, which results in 7.37% CER reduction and 5.86% SER reduction with respect to the result when no *Clustering-Classification* framework is applied, although the performance for number of clusters ranging between 50 and 500 are very close. As for derived resource, the best number of clusters is between 50 and 100.

4.1.2. Existing External Resource vs. Derived External Resource

From the data in Table 1 and Fig. 4, it is clear that the existing external resources always achieve better performance than the derived external resources. This is natural for the case here because both the existing external resources and

Language Models		CER		SER	
Baseline LM		23.17		13.17	
Adapted with Existing resource	1 cluster	18.05		10.58	
	10 clusters	17.32	(4.04%)	10.20	(3.59%)
	50 clusters	16.76	(7.15%)	9.94	(6.05%)
	100 clusters	16.75	(7.20%)	9.96	(5.86%)
	200 clusters	16.72	(7.37%)	9.96	(5.86%)
500 clusters	16.86	(6.59%)	10.07	(4.82%)	
Adapted with Derived resource	1 cluster	20.18		11.65	
	10 clusters	19.80	(1.88%)	11.43	(1.89%)
	50 clusters	19.83	(1.73%)	11.46	(1.63%)
	100 clusters	19.85	(1.64%)	11.40	(2.15%)
	200 clusters	20.29	(-0.55%)	11.71	(-0.52%)
500 clusters	20.88	(-3.47%)	11.97	(-2.75%)	

Table 1. Character and syllable error rates (error rate reduction with respect to 1 cluster, or the *Clustering-Classification* process was not applied for the first test set of broadcast news.

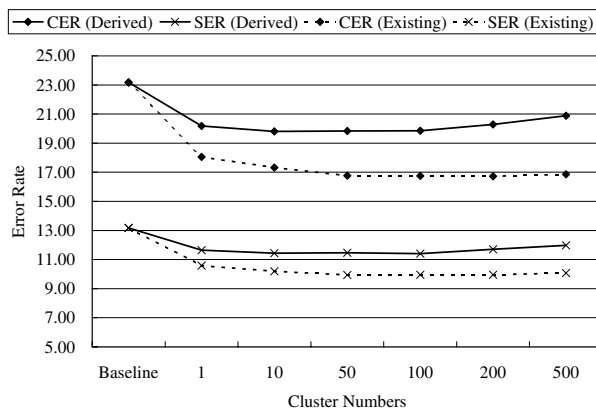


Fig. 4. Character and syllable error rates for the first test set of broadcast news plotted as functions of cluster numbers

the target spoken documents are news, and the existing external resources are a quite complete set of news. On the other hand, the derived external resources may include some misleading topics resulted from the incorrect transcriptions, and the correct supporting news may not be as complete as those in the existing resources.

4.1.3. Different Query-Construction Approaches

We also compared the different query-construction approaches which we discussed in 3.2.1 (disjunctive 3-word queries, **DISJ3**) and 3.2.2 (filtered by confidence measures, **FCM**). The total number of queries and the derived external resources we collected in the tests by two different query-construction approaches are depicted in Table 2. It can be found from this table that by filtering queries with the confidence measure, apparently the quantity of the derived external resources we collected is reduced by a factor of 4.

Regarding the achievable performance for these two query-construction approaches, we designed preliminary ex-

Query-construction approaches	Queries	Characters
Disjunctive 3-grams (DISJ3)	31108	17092041
Filtered by Confidence Measure (FCM)	7301	3754509

Table 2. Number of queries used and number of characters retrieved for the derived external resource (for the first test set of broadcast news) using different query-construction approaches

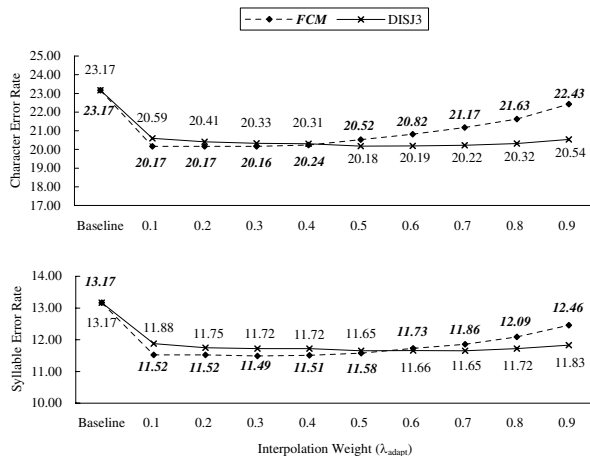


Fig. 5. CER and SER with respect to λ_{adapt} for the first test set of broadcast news

periments to see the respective improvements obtained. We set the the linear interpolation weight for the adaptation data λ_{adapt} used in the language model adaptation from 0.1 to 0.9. The results are shown in Fig. 5. We can see that the best performance of **FCM** (CER=20.16%, SER=11.49%) is slightly better than that of **DISJ3** (CER=20.18%, SER=11.65%), even though **FCM** used much less quantity of data. This shows the advantage of the **FCM** query-construction approach. In Fig. 5, we also observed that the performance of **FCM** is more sensitive to the interpolation weight λ_{adapt} . This is probably because **FCM** collected much less quantity of text data; these data are more specific to the topic addressed, but are rather unbalanced in general concern. Therefore, the performance degrades when λ_{adapt} of **FCM** becomes too larger. On the other hand, **DISJ3** might collect many “noisy” data which are not as specific to the topics addressed by the spoken documents, but the larger quantity of data resulted in a more stable performance curve while λ_{adapt} was adjusted.

LM Data Sources		CER		SER	
Baseline LM		72.52		56.11	
Iteration 1	1 cluster	70.44	(2.87%)	55.19	(1.64%)
	5 clusters	69.66	(3.94%)	54.74	(2.44%)
	10 clusters	69.56	(4.08%)	54.90	(2.16%)
Iteration 2	1 cluster	69.95	(3.54%)	54.82	(2.30%)
	5 clusters	68.77	(5.17%)	53.98	(3.80%)
	10 clusters	69.31	(4.43%)	54.69	(2.53%)

Table 3. Character and syllable error rates (error rate reduction with respect to *Baseline* for the second test set of CNA interview recording.)

4.2. Experiments on the Second Test Set of CNA Interview Recording

In addition to broadcast news, we also performed experiments on the second test set of CNA interview recording. The topic addressed in the interview has nothing to do with the news. It is thus intuitive that the existing resources of news won't be helpful for this case, but we can observe the effect of obtaining derived resources for adapting the language model to a brand-new domain. First of all, the baseline transcriptions were performed using the baseline recognizer, whose acoustic and language models were all trained from the broadcast news. We can see from the results listed in Table 3 that the baseline CER and SER are both very high when compared with the results of broadcast news in Table 1.

With the baseline transcriptions were generated, we retrieved the derived external resources using the conjunctive queries filtered by the confidence measure (**FCM**) and performed the adaptation with the *Clustering-Classification* framework with 1, 5, and 10 clusters. Since the target spoken documents are highly mismatched with the baseline language model, we performed the procedure twice to see how much extra benefit we could get. As can be found in Table 3, in the first iteration a highest CER reduction of 4.08% was obtained, while in the second iteration, we can see that the error rate can be further reduced, with a highest achievable CER reduction of 5.17%. This reveals the fact that the more correct characters we have at hand, the more accurate derived external resources we may obtain. In other words, the very limited improvements obtained here may be due to the very poor baseline transcription we started with. A baseline recognizer better matched to the target spoken documents may give much better results.

5. CONCLUSION

In this paper, we propose to use existing as well as derived external resources for language model adaptation with a *Clustering-Classification* framework. We also go deep

into the considerations of different query-construction approaches, and we verified that the quality of derived external resources can be enhanced using queries filtered by confidence measure. Detailed analysis and convincing experiments were performed on broadcast news as well as an interview recording.

6. REFERENCES

- [1] "Yahoo! Kimo News portal," <http://tw.news.yahoo.com/>.
- [2] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *(COLT): Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, 1998.
- [3] Langzhou Chen, Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. ICASSP*, Hong Kong, April 2003, pp. I-220-223.
- [4] Michael Steinbach, George Karypis, and Vipin Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000.
- [5] George Karypis, "Cluto: A clustering toolkit," Tech. Rep. #02-017, University of Minnesota, Department of Computer Science, August 2002.
- [6] Pedro Domingos and Michael Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103-130, 1997.
- [7] Andrew Kachites McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," <http://www.cs.cmu.edu/mccallum/bow>, 1996.
- [8] "Google," <http://www.google.com/>.
- [9] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 9, pp. 288-298, March 2001.
- [10] Yi-Cheng Pan, "One-pass and word-graph-based search algorithms for large vocabulary continuous mandarin speech recognition," M.S. thesis, National Taiwan University, 2001.
- [11] Andreas Stolcke, "SRI language modeling toolkit," version 1.3.3, <http://www.speech.sri.com/projects/srilm/>.
- [12] "News 98 FM-98.1," <http://www.news98.com.tw/>.