INTEGRATING RECOGNITION AND RETRIEVAL WITH USER FEEDBACK: A NEW FRAMEWORK FOR SPOKEN TERM DETECTION

Hung-yi Lee and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University

ABSTRACT

People usually consider recognition and retrieval as two cascaded independent modules for spoken term detection. Retrieval techniques were assumed to be applied on top of some ASR output, with performance depending on ASR accuracy. In this paper, we propose a new framework: to integrate the two parts into a single task. This can be achieved by adjusting the acoustic model parameters, borrowing the principle of Minimum Classification Error (MCE), based on user feedback. The modified acoustic models then give updated posterior probabilities for the lattice-based structures used in spoken term detection. Encouraging results were obtained on a bilingual course lecture corpus in preliminary experiments.

Index Terms— Spoken Term Detection, Discriminative Training

1. INTRODUCTION

Spoken term detection is to return a list of spoken segments containing the term requested by the user. It has been considered as a key technology for voice-based information retrieval, which is believed to be very important for the network era when people try to access the multimedia content based on its audio signals. Spoken term detection is usually accomplished by the following steps. 1-best transcriptions or lattices of spoken segments are first generated by automatic speech recognition (ASR). Lattices are preferred since they include multiple recognition hypotheses, especially when the accuracy in 1-best transcriptions is relatively low. Very often lattices are converted into sausage-like structures in order to make the indexing task easier and save the required memory space. Some efficient sausage-like structures have been shown to be not only memory saving, but able to maintain or even improve the retrieval performance. Good examples of such sausage-like lattice-based structures include Position Specific Posterior Lattices (PSPL)[1], Confusion Networks (CN)[2, 3] and Time-based Merging for Index (TMI)[4], etc. Indexing and search is then performed over these ASR output (sausage-like structures or 1-best transcriptions), actually very similar to those over text documents. As a result, many text document retrieval technologies can be transplanted onto spoken term detection easily. For example, the approach of learning to rank previously developed for text document retrieval has been successfully used on spoken term detection [5]. Such techniques using sausage-like structures are referred to as lattice-based spoken term detection in this paper. However, in the past people usually consider recognition and retrieval as two cascaded independent modules, and assumed they should be individually optimized. For example, it is usually believed that retrieval performance depends heavily on ASR accuracy. Also many spoken term detection techniques were proposed assuming they should be applied on top of some ASR output. In this paper, we consider a new framework to integrate the two parts of recognition and retrieval together into a single task.

There are apparent limitations when considering only the retrieval process applied on top of the ASR output, since ASR output is the only representation of spoken segments the retrieval process can use (using other information such as prosody and speaker information is out of the scope of this paper). When the recognition performs very bad, for example the correct word hypothesis are included in the lattice but with very low posterior probabilities, it is difficult to detect the spoken term even with lattices. As a result, spoken term detection performance is inevitably dominated by the ASR performance. However, in many practical applications, it is difficult to obtain acoustic and language models robust enough for the huge quantities of target spoken segments generated in different applications of different domains. In such cases even very robust retrieval approaches are not able to compensate for the recognition errors.

Some research works have in fact considered the recognition and retrieval process as a whole to try to improve retrieval performance, but efforts deliberating on the interaction between the recognition and retrieval processes are still very limited. Considering the recognition error pattern by a confusion matrix during retrieval has been a very good example [6]. In this approach people tried to infer the correct words actually appearing in the spoken segments from the erroneous ASR transcriptions. People have also observed that although word accuracy is an excellent metric to evaluate recognition performance, it is not directly related to the retrieval performance [7, 8, 9]. For example, those words frequently used as query terms should be correctly recognized while recognition errors for function words almost have no impact on retrieval performance. As a result, word significance was carefully considered during decoding [7, 8]. Also, Minimum Classification Error (MCE [10]) discriminative training method has been used by considering the word significance [9]. The interaction of retrieval and recognition processes was also proposed previously [11], in which when an out-of-vocabulary (OOV) query term is entered, the OOV query term is dynamically inserted into the possible position in the lattice to take into account the OOV query. The query was also expanded by relevant feedback [12].

In this paper, we propose to integrate the recognition and retrieval modules as a whole. The parameters of the acoustic models used for recognition are first adjusted according to the user feedback. The posterior probabilities on the lattice-based structures for spoken segments to be retrieved are then rescored by the new set of acoustic models to improve the retrieval performance for queries entered in the future. This technique can be very helpful for a search engine aiming at indexing spoken segments available on many web sites over the Internet with various acoustic/linguistic conditions, for which adapting the acoustic/language models for the various acoustic/linguistic conditions is almost impossible. This is

Hung-yi Lee was supported from Sept 2008 to Aug 2009 by the National Taiwan University Advanced Speech Technologies Scholarship.

different from some successful spoken document retrieval systems currently already on-line, for which the spoken segments to be retrieved are primarily for a specific application task and therefore a better set of acoustic and language models are obtainable. With this approach proposed here, acoustic models can be adjusted and posterior probabilities updated based on user feedback. This can be an important step forward towards a more robust spoken segment retrieval technologies.

Below the proposed approach is presented in Section 2, and the experimental results in Section 3. Section 4 are the concluding remarks.

2. PROPOSED APPROACH

2.1. Overview of the Proposed Approach

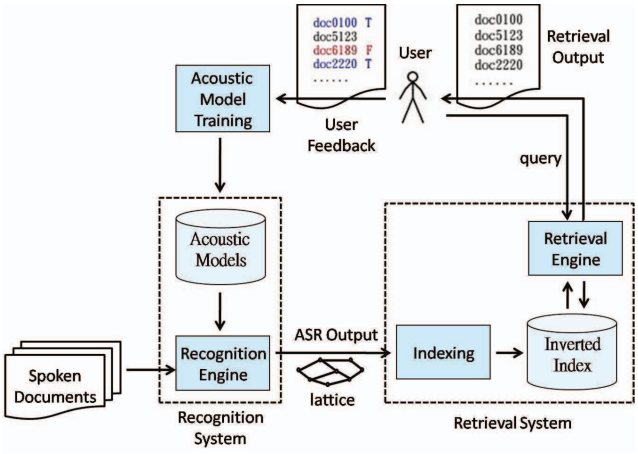

Fig. 1. The framework of the proposed approach.

The framework of the proposed approach is shown in Figure 1. Just as the standard spoken term detection procedures, the spoken document collection is first divided into many spoken segments, and each spoken segment is transcribed into a lattice by a recognition engine based on a set of acoustic models. All the lattices are then indexed for fast retrieval in the retrieval system. This is a cascade of two modules, recognition and retrieval. Here we introduce the user feedback to construct a loop, with which the two modules can be jointly improved step by step. In other words, when a query is entered by a user, the retrieval system searches over the inverted index, and offer a ranked list of matched spoken segments to the user. If the user gives some feedback to the system, for example he selects items 1 and 4 shown in Figure 1 as relevant but item 3 as irrelevant, a new set of acoustic models is then estimated based on the feedback. The new set of acoustic models is then used to update the posterior probability on the lattice-based structures and then the inverted index are modified accordingly. Apparently for frequently entered query terms, which very often appear in search engines, the retrieval performance can be improved step by step when queries including the same term are entered by the users repeatedly.

2.2. Lattices and Posterior Probability

Each observation sequence X for a spoken segment is transcribed into a lattice $\mathcal{L}$ which is a weighted directed acyclic graph (DAG) $\{\mathcal{N}, \mathcal{A}\}$, where $\mathcal{N}$ is the set of nodes containing the time information, and $\mathcal{A}$ is the set of arcs including its word or subword hypotheses, acoustic likelihood and language model score. Below we assume the lattice is based on words, although extension to subword units is straightforward. Let $word(a)$ denote the word hypotheses of arc a, X_a the observation sequence corresponding to arc a, and $head(a)$ and $tail(a)$ the two nodes of arc a on the two ends.

The posterior probability of every arc a in the lattice $\mathcal{L}$ is then computed,

$$P(a|X,\theta) = \frac{\alpha(head(a))P(X_a|a,\theta)P(a)\beta(tail(a))}{\beta_{start}}, \quad (1)$$

where θ is the acoustic model set, $P(X_a|a,\theta)$ is the acoustic likelihood of the observation sequence X_a given arc a and the acoustic model set θ, $P(a)$ the language model score of arc a, $\alpha(head(a))$ and $\beta(tail(a))$ the forward and backward probabilities accumulated to the nodes $head(a)$ and $tail(a)$, respectively, and β_{start} the sum of posterior probabilities for all paths in the lattice $\mathcal{L}$ recognized from X based on θ.

2.3. Indexing and Retrieval

When a query, Q, is entered, a ranked list of spoken segments X based on the relevant score $S(Q, X|\theta)$ is returned by the system,

$$S(Q,X|\theta) = \sum_{a\in\mathcal{A}, word(a)=Q} P(a|X,\theta), \quad (2)$$

where $P(a|X,\theta)$ is defined in Equation (1) and $\mathcal{A}$ is the arc set of the lattice $\mathcal{L}$ recognized from X using the model set θ. Here we assume the query Q has only a single word through this paper for simplicity in the initial work, although extension to longer queries is not difficult. Practically, the reduction of the lattice into sausage-like structures such as PSPL or CN does not influence the value of $S(Q, X|\theta)$ when the query has only a single word.

2.4. User Feedback

Here we assume for each item returned by the retrieval system, the user can select to click "relevant" (the query word Q is included in the segment) or "irrelevant" (the query word Q is not included in the segment), or not to click. If a spoken segment is labelled relevant with the query term Q, the observation sequence is denoted X_T^Q, or a positive example; If labelled irrelevant with Q, the observation sequence is denoted X_F^Q, or a negative example.

2.5. Acoustic Model Training

Using the training data feedback from the user, we wish to find a new set of acoustic model parameters θ which minimize the objective loss function defined in Equation (3),

$$loss(\theta) = \sum_{X_F^Q} \sum_{X_T^Q} l(d(X_F^Q, X_T^Q|\theta)), \quad (3)$$

$$d(X_F^Q, X_T^Q|\theta) = S(Q, X_F^Q|\theta) - S(Q, X_T^Q|\theta), \quad (4)$$

where $S(Q, X|\theta)$ is as defined in Equation (2), and $l(.)$ is a sigmoid function. The spirit of minimizing $loss(\theta)$ is equal to trying to make the relevant score for every X_T^Q larger than the relevant score for every X_F^Q. The acoustic parameters can then be adjusted iteratively to minimize Equation (3) by gradient decent from Equations (5) to (9) below. The acoustic model set used to generate the original lattice serves as the initial model, θ^0.

$$\theta^{i+1} = \theta^i - \mu\frac{\partial loss(\theta^i)}{\partial\theta} \quad (5)$$

$$\frac{\partial loss(\theta^i)}{\partial \theta} = \sum_{X_F^Q} \sum_{X_T^Q} \frac{\partial l(\theta^i)}{\partial \theta} \quad (6)$$

$$\frac{\partial l(\theta^i)}{\partial \theta} = \frac{\partial l(d(\theta^i))}{\partial d(\theta)} \frac{\partial d(\theta^i)}{\partial \theta} \quad (7)$$

$$\frac{\partial d(\theta^i)}{\partial \theta} = \frac{\partial S(Q, X_F^Q|\theta^i)}{\partial \theta} - \frac{\partial S(Q, X_T^Q|\theta^i)}{\partial \theta} \quad (8)$$

$$\frac{\partial S(Q, X|\theta)}{\partial \theta} = \sum_{a \in \mathcal{A}, word(a)=Q} \frac{\partial P(a|X, \theta^i)}{\partial \theta} \quad (9)$$

In Equation (9), we have to perform partial differentiation on the posterior probability of an arc a. However, it is not easy to perform partial differentiation on the posterior probability $P(a|X, \theta)$ in Equation (1). Thus we use the confusion network [3] to approximate the lattice structure [13]. All the arcs in the lattice are clustered into a sequence of arc clusters. But different from the conventional confusion network [3], here we do not merge the arcs with the same word hypothesis into a single arc, but keep them as they are instead. In this way the partial differentiation in Equation (9) can be performed easily. Note that we still use the exact posterior probability in Equation (1) when computing the relevant score, and the approximate posterior probability from confusion network mentioned here is only used for acoustic model parameter estimation.

Assume arc a' represents a certain arc in cluster $\mathcal{C}$ which contains arc a. Consider the special structure of the confusion network mentioned above, we can write β_{start} in Equation (1) as in Equation (10) below because in the confusion network $head(a)$ and $tail(a)$ is exactly the same as $head(a')$ and $tail(a')$, respectively.

$$\beta_{start} = \alpha(head(a))\beta(tail(a))(\sum_{a' \in \mathcal{C}} P(a'|X_{a'}, \theta)P(a')) \quad (10)$$

Substitute Equation (10) into Equation (1), we have an approximate posterior probability in Equation (11).

$$P(a|X, \theta) = \frac{P(a|X_a, \theta)P(a)}{\sum_{a' \in \mathcal{C}} P(a'|X_{a'}, \theta)P(a')} \quad (11)$$

Partial differentiation over Equation (11) with respect to θ is now trivial.

2.6. Updating Posterior Probabilities

After a new set of acoustic models is obtained, we use the new models to recompute the acoustic likelihood of the arcs in those lattices whose corresponding spoken segments are returned by the query. We then compute the new posterior probabilities of all arcs in the lattices whose acoustic likelihood has been changed, and update the inverted index accordingly.

3. EXPERIMENTS

3.1. Experimental Setup

We used the recorded lectures for a course of Digital Speech Processing offered in National Taiwan University in 2006 as our corpus, which included 45 hours of audio and was divided into about 30000 spoken segments. All these segments were produced in the host language of Mandarin Chinese, but embedded with many words (in particular the terminologies for the course) produced in the guest language of English. Such code-switching phenomena are normal in the courses offered in Taiwan. We split the lecture corpus into two parts: 12 hours for acoustic model training and 33 hours for retrieval tests.

A phone set of 74 phonemes was used for transcription, which is a direct combination of 35 Mandarin phonemes and 39 English phonemes without any effort of merging, even if some Mandarin phonemes are very similar to some English phonemes. We extracted 39 MFCC features, and trained speaker independent cross-word triphone acoustic models using HTK, which were then further adapted to the voice of the course instructor. The lexicon is a combination of a monolingual Chinese dictionary and all English words in slides which includes many terminologies. Because of the lack of bilingual corpus, specially those matched to the topic (technical content of the course) and the style (spontaneous monologue) for the task here, we were not able to train bilingual language models. So a Chinese trigram language model was trained from the Mandarin Giga-word corpus released by Linguistic Data Consortium. We also trained an English unigram language model from the slides, and linearly interpolated it with the Chinese trigram model. These bilingual acoustic/language models and lexicon were used in the transcription.

Each spoken segment in the 33 hours of corpus for retrieval tests was transcribed into a lattice with a beam width of 50, and then transformed into a PSPL structure. Our system indexed those PSPL's and performed retrieval on them. We manually selected 40 single word English queries and 34 single word Chinese queries as our testing query set. We used mean average precision (MAP[14]) as our evaluation measure for retrieval performance evaluation.

When a testing query was entered to the system, the system returned a spoken segment list. A user then randomly selected a part of the returned spoken segments and labelled those segments as either relevant to the query or not. Those label segments were collected and used to estimate a new set of acoustic models, and the system used the new acoustic model set to rescore the PSPL's of those returned segments without labelling (all labelled segments were considered training segments thus not used in retrieval performance evaluation in all the tests below). The inverted index were changed due to modified likelihoods of arcs in some lattices and PSPL's. After all queries in the query set had been entered and the corresponding lattices and PSPL's had been rescored by the new sets of acoustic models, the same set of testing queries were entered again. We compare the MAP retrieval performance over all the spoken segments not labelled previously (those labelled by the user were considered as training set) before and after rescoring.

3.2. Experimental Result

Table 1 lists the average number of positive and negative training examples when different percentages of returned segments were labelled by the user feedback. As can be found, for Chinese queries, only less than 14 segments for each query in average $(6.4 + 7.4)$ were labelled and used in the acoustic model training for the 50% feedback case, and so on. On the other hand, for English queries, roughly 50 segments $(10.8 + 38.1)$ were labelled and used in the acoustic model training for the 50% feedback case. There were more false alarms for English queries than Chinese queries.

In Table 2, we compare the MAP performance of the proposed approach with the baseline system without feedback. It can be found that our proposed approach significantly improved the MAP score for all percentages of feedback no matter in English or Mandarin queries. Also, we observe that although Chinese queries used much less training examples than English queries, the performance for

Table 1. *Number of training examples used for different percentages of feedback for English and Chinese queries.*

	feedback percentage of labelled segments	average number of positive examples	average number of negative examples
English Query	50%	10.8	38.1
	40%	8.4	30.3
	30%	6.2	22.6
	20%	4.0	15.0
Chinese Query	50%	6.4	7.4
	40%	5.4	6.3
	30%	4.2	4.9
	20%	3.0	3.5

Chinese queries were higher than those for English queries. We are not sure if the unbalance of positive and negative examples limited the improvement of English queries.

Table 2. *MAP of baseline and proposed approach with different feedback percentages and query types.*

	feedback percentage	baseline	proposed approach	absolute improvement in MAP
English Query	50%	56.71	58.61	1.90
	40%	55.75	56.99	1.24
	30%	49.71	50.30	0.59
	20%	44.78	45.25	0.47
Chinese Query	50%	67.32	68.90	1.58
	40%	67.03	70.32	3.29
	30%	65.80	68.29	2.49
	20%	65.76	67.86	2.10

Table 3 lists the overall word accuracy for the different cases in Table 2. It can be found that although the proposed approach significantly improved the retrieval performance, the word accuracy actually remained unchanged or even slightly decreased with the updated acoustic models. This is not surprising because in the model training process the object function aimed for better retrieval performance, which is not directly related to the overall word accuracy of the transcription. In fact, we found that even the recognition accuracy for the query terms was not improved either. However, we found the posterior probabilities of the word arcs for the query terms in the relevant lattices were increased by the feedback in many cases, which means the quality of the lattices or PSPLs were actually improved.

Table 3. *Word accuracy of baseline and proposed approach with different feedback percentages and query types.*

	feedback percentage	baseline	proposed approach
English Query	50%	51.60	51.60
	40%	48.97	48.94
	30%	49.05	49.04
	20%	49.06	49.06
Chinese Query	50%	52.69	52.63
	40%	52.47	52.38
	30%	52.34	52.25
	20%	52.44	52.36

4. CONCLUDING REMARKS

In this paper, we propose to use the feedback from the user to estimate better acoustic models for retrieval purposes, and show that in this way the retrieval performance can be improved significantly.

More work is left for the future. For example, most state-of-the-art discriminative training methods are based on optimizing the word accuracy, phone accuracy, or similar, but word accuracy is not always related to retrieval performance. It is possible to develop retrieval oriented discriminative training approach aiming for better retrieval performance. Clearly there are still many opportunities for future advances in the area of spoken term detection.

5. REFERENCES

[1] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *ACL*, 2005.

[2] T. Hori, I.L. Hetherington, T.J. Hazen, and J.R. Glass, "Open vocabulary spoken utterance retrieval using confusion networks," in *ICASSP*, 2007.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other application of confusion networks," in *Computer Speehc and Language, vol. 14, no. 4, pp. 373-400*, 2000.

[4] Peng Yu, Yu Shi, and Frank Seide, "Approximate word-lattice indexing with text indexers: Time-anchored lattice expansion," in *ICASSP*, 2008.

[5] C.-H. Meng, H.-Y. Lee, and L.-S. Lee, "Improved lattice-based spoken document retrieval by directly learning from the evaluation measures," in *ICASSP*, 2009.

[6] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *SIGIR*, 2000.

[7] H. Nanjo and T. Kawahara, "A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding," in *ICASSP*, 2005.

[8] T. Shichiri, H. Nanjo, and T. Yoshimi, "Minimum bayes-risk decoding with presumed word significance for speech based information retrieval," in *ICASSP*, 2008.

[9] Q. Fu and B.-H. Juang, "Automatic speech recognition based on weighted minimum classification error (W-MCE) training method," in *ASRU*, 2007.

[10] B.-H. Juang, W. Chou, and C.-H. Lee., "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Signal Processing*, vol. 5(3), pp. 257–265, 1997.

[11] J. Shao, R.-P. Yu, Q. Zhao, Y. Yan, and F. Seide, "Towards vocabulary-independent speech indexing for large-scale repositories," in *Interspeech*, 2008.

[12] Wade Shen, Christopher M. White, and Timothy J. Hazen, "A comparison of query-by-example methods for spoken term detection," in *INTERSPEECH*, 2009.

[13] K. Thambiratnam and F. Seide, "Unsupervised lattice-based acoustic model adaptation for speaker-dependent conversational telephone speech transcription," in *INTERSPEECH*, 2009.

[14] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees, "The trec spoken document retrieval track: A success story," in *in Text Retrieval Conference (TREC) 8*, 2000, pp. 16–19.