SPOKEN LANGUAGE RECOGNITION WITH RELEVANCE FEEDBACK

Rong Tong[1,2], *Haizhou Li*[1,2], *Bin Ma*[1], *Eng Siong Chng*[2] *and* Siu-Yeung Cho[2]
[1]Institute for Infocomm Research, Singapore 119613,
[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798
{tongrong, hli, mabin}@i2r.a-star.edu.sg aseschng@ntu.edu.sg assycho@ntu.edu.sg

ABSTRACT

This paper applies relevance feedback technique in spoken language recognition task, in which we consider a test utterance as a test query. Assuming that we have a labeled multilingual corpus, we exploit the retrieved utterances from such a reference corpus to automatically augment the test query. Note that successful spoken language recognition relies on sufficient query data. The proposed method is especially effective for short query by expanding the query at a low cost. Experiments show that unsupervised relevance feedback reduces the relative equal-error-rate by 16.2%, 4.9% and 10.2% on NIST LRE 1996, 2003 and 2005 databases respectively for 3-second trials.

Index Terms— Spoken language recognition, relevance feedback, vector space model

1. INTRODUCTION

Automatic spoken language identification (LID) is the process of determining the identity of the language corresponding to a given set of test utterances. Studies show that LID becomes more accurate as more test samples are available. It is observed that in NIST LRE evaluation tasks, the equal-error-rates of LID drop substantially when the test sample is reduced from 30 seconds to 10 or 3 seconds in length [1,2]. We usually have plenty of labeled training samples in terms of hours. On the other hand, we have test samples as short as few seconds. In light of this, we are prompted to think of ways to automatically increase the amount of test samples by making use of the labeled training samples.

Building a LID system, we typically use a large training set to train a model for each language. We further use a held-out data set, which behaves similarly to the test data, also known as the development set, to fine-tune the LID classifier. In this paper, we discuss a novel LID approach motivated by the relevance feedback technique in information retrieval. In this approach, we use the development data set as the reference corpus to augment the test query.

Relevance feedback and the selection of search terms for query expansions are the major research areas in information retrieval research [3,4]. The automatic query expansion techniques utilize the text of a user query and retrieved documents that are relevant to the user as input for different techniques [5,6] to derive a set of search terms for a new search. In this way, the original query is augmented with new terms and new statistics that are learnt from the initial retrieved results. The focus of automatic query expansion is on formulating the algorithms and automatic mechanisms that select and weight search terms for query expansion.

Motivated by the idea of relevance feedback, we propose a novel approach to LID, in which speech query expansion strategies are studied in the framework of vector-based LID [1,7].

This paper is organized as follows. In Section 2, we briefly introduce a vector-based spoken language recognition system which serves as the workbench of our study. In Section 3, we describe three query expansion strategies. In Section 4, we present the experiment results. Finally, we conclude in Section 5.

2. VECTOR SPACE MODELING

Suppose that we have a speech recognition frontend consisting of F parallel phone recognizers (PPR) $V = \{V_1,...,V_f,...V_F\}$, where V_f represents the set of m_f phones from language *f*. For each phone sequence generated from a phone recognizer, we derive a spoken document vector from the phone *n*-gram counts, also known as *acoustic word* [1] counts. We can derive a large composite document vector by stacking F vectors resulting from the individual phone recognizers to obtain a *bag-of-sounds* vector of $m_1 \times m_1 + ... + m_F \times m_F$ dimensions in the case of bigrams. This process is illustrated in Figure 1. In this way, the utterances for each language are represented as a collection of such vectors. Details of *bag-of-sounds* method can be found in [1].

The LID task can be seen as a multiclass classification problem in a high dimensional *bag-of-sounds* vector space. To reduce the dimensionality of *bag-of-sounds* vectors, we further adopt pairwise support vector machines (SVM) to form an ensemble classifier. Given *M* languages, we build *M(M-1)/2* pairwise SVMs. The outputs of the SVM ensemble

classifier convert the *bag-of-sounds* vectors to a lower dimensional space with each dimension being characterized by an independent SVM. The dimension-reduced vector is also referred to as the output vector, denoted as $\mathbf{y}$. The pairwise SVM partitions are illustrated in Figure 2. For the *CallFriend* database which consists of M=15 language/accents, we can build an output vector of 105 dimensions.

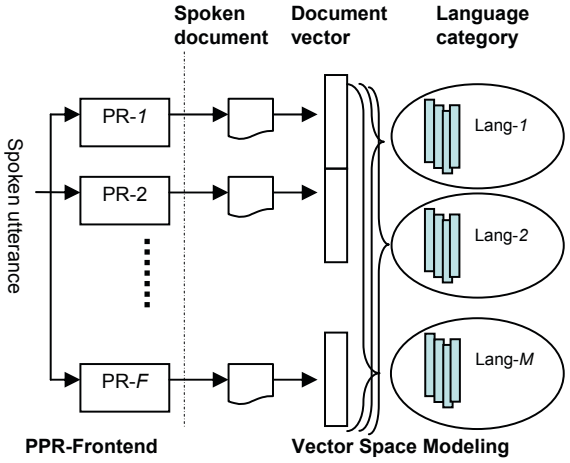

Figure 1: LID systems with parallel phone recognition frontend and vector space modeling backend

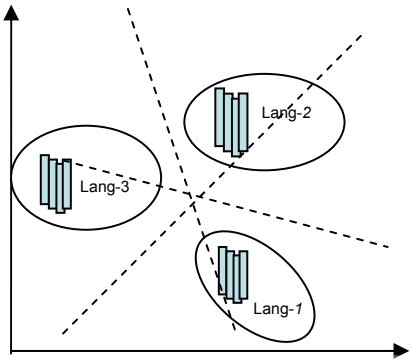

Figure 2. An ensemble classifier with 3 pairwise SVMs

After dimensionality reduction, we construct Gaussian mixture models (GMM) for each target language m^+ and their competing languages m^-. As such, for each target language, we build a pair of GMMs $\{m^+, m^-\}$. For language verification, we need to evaluate the probability of a hypothesized language model m^+ for a given test utterance O in the form of output vector $\mathbf{y}$, $P(m^+ | \mathbf{y})$. However, the output of a GMM system gives $P(\mathbf{y} | m^+)$. By making the assumption that all languages are equiprobable, we approximate the *posterior* probability $P(m^+ | \mathbf{y})$ by Bayes' theorem:

$$\log P(m^+ | \mathbf{y}) = \log P(\mathbf{y} | m^+) - \log P(\mathbf{y} | m^-) \qquad (1)$$

Eq.(1) gives a relative log-likelihood score between the target language and its competing languages. Intuitively, $\log P(m^+ | \mathbf{y})$ reflects how the target model overtakes the competing models with respect to the input utterance, thus serving as the confidence score of a test vector $\mathbf{y}$ being hypothesized as m^+.

3. SPEECH QUERY EXPANSION

We prepare a reference database from *CallFriend* development set. We select 200 utterances from each language resulting in a reference corpus of 3,000 utterances for 15 languages/accents. Each utterance is about 3 seconds in length. To measure the similarity between two output vectors $\mathbf{y}_1$ and $\mathbf{y}_2$, we calculate the cosine distance as in Eq.(2) between them.

$$d(\mathbf{y}_1, \mathbf{y}_2) = \frac{\mathbf{y}_1 \cdot \mathbf{y}_2}{|\mathbf{y}_1||\mathbf{y}_2|} \qquad (2)$$

In this way we can rank the utterances in the reference corpus by their similarity or relevance with regard to the query. In text-based information retrieval, one is able to tell relevant documents from irrelevant ones. As such, the terms from the relevant documents are used to enhance the query while those from non-relevant documents are suppressed. However, in the LID scenario, human-assisted relevance feedback is not possible because we can not assume the existence of such a human agent at run-time. The challenge is therefore to find an automatic query expansion technique that makes use of a reference corpus. One possible way is to use the ranking of utterances as the indication of relevancy.

In language detection or verification application, ultimately we would like to derive a confidence score for a given query as defined in Eq.(1). The expanded query is expected to provide more reliable statistics than the initial query. Motivated by the idea of text query expansion, we will study ways to extract relevant statistics from the reference corpus to augment the initial speech query.

3.1. Expansion with relevant utterances

Given a test query, an easiest way is to sort the reference corpus by similarity distance as defined in Eq.(2). In this way, we can use the top N-best choices to augment the test query. Note that the top N-best results are not necessarily from the same language. This is fine because what we want is the relevant statistics that are related to the test query. The process is illustrated in Figure 3, where shaded, meshed and clear bars indicate vectors from 3 different languages.

In Section 2, we discuss two cascaded vectorization, the high dimensional *bag-of-sounds* vectors followed by the output vectors for dimensional reduction. The former is composed of phone *n*-gram counts, each representing an *acoustic word* [1] as an indexing term (Figure 1); the latter is composed of pairwise SVM outputs (Figure 2). The

query expansion can be applied to both vectors. For computational efficiency, we prefer to operate on the output vectors. However, if indexing term reweighting is needed, we have no choice but to operate on *bag-of-sounds* vectors, in that case, the SVM outputs need to be recomputed.

The query expansion here is implemented as a weighted sum over the output vectors of input query and N-best choices.

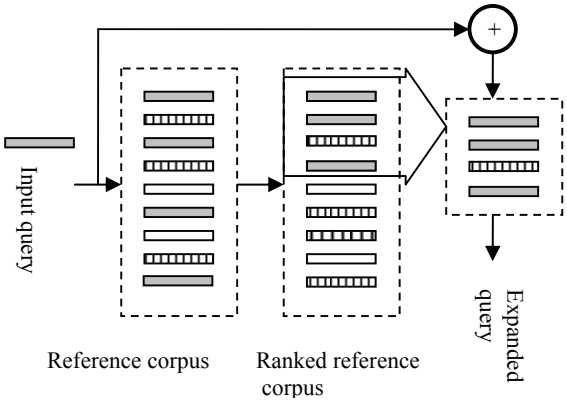

Figure 3. Speech query expansion with relevant utterances

3.2. Expansion with relevant clusters

Note that longer utterance provides more reliable statistics. We further propose using clusters of the reference utterances instead of single utterances for query expansion. In this case, we group utterances of the reference corpus in each language into clusters using k-means clustering approach during preparation of reference corpus. Given a test query, we rank the clusters by similarity distance and use the top N-best clusters to augment the test query. Similar to the N-best utterances, the top N clusters may or may not belong to the same language. The process can be illustrated in Figure 4 where inner dotted boxes represent clusters where the query expansion is operated over the output vectors.

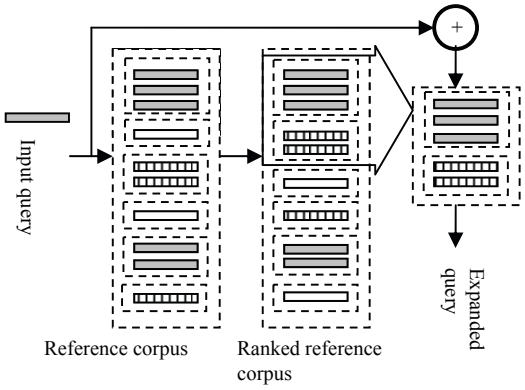

Figure 4. Speech query expansion with relevant clusters

3.3. Expansion by reweighting terms

In [6], Rocchio formulated a way to combine initial retrieved documents with test query using vector space model and shown very positive results. Motivated by the idea of reweighting the query terms in Rocchio's approach, we operate on the *bag-of-sounds* vector in this case. The query term expansion can be described by Eq. (3).

$$w_1 = w_0 + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{l=1}^{n_2} \frac{S_l}{n_2} \qquad (3)$$

where w_0 is the original bigram counts in the *bag-of-sounds* vector, R_k denotes the bigram count observed in kth relevant utterance, S_l denotes that observed in the lth non-relevant utterances, n_1 and n_2 are the numbers of relevant and non-relevant utterances respectively. β and γ are the parameters that control the contributions of relevant and non-relevant feedback. The reweighted *bag-of-sounds* vectors are further processed to obtain the output vectors for GMM modeling. Eq.(3) can be illustrated in Figure 5.

4. EXPERIMENTS

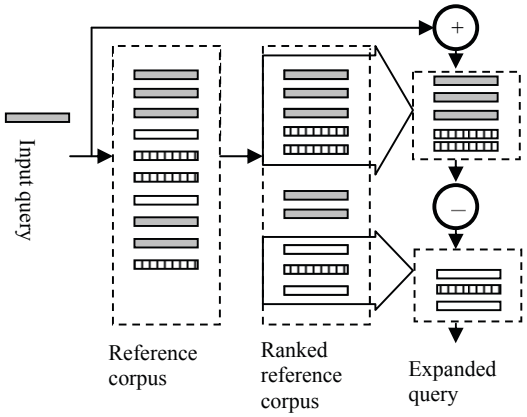

Figure 5. Speech query expansion by reweighting *bag-of-sounds* terms

We follow the experiment setup in the NIST LRE tasks[1]. In the 1996 and 2003 tasks, 12 known languages are tested, with Russian being the out-of-language (OOL) in the 2003 test. In the 2005 task, 8 languages, a subset of the 1996 and 2003 languages, are tested, with German being the OOL.

Training sets came from three corpora [1], namely: (i) the 3-language IIR-LID database; (ii) the 6-language OGI-

[1] http://www.nist.gov/speech/tests/index.htm

TS (Multilanguage Telephone Speech) database; and (iii) the 12-language LDC CallFriend database. Both IIR-LID and OGI-TS are telephone speech with phonetic transcriptions. They are used for acoustic modeling. In addition, the *CallFriend* database was used for constructing *bag-of-sounds vectors* and designing classifiers [1]. It contains telephone conversations of the same 12 languages as are in the 1996 and 2003 NIST LRE tasks, with 3 languages having 2 accented versions. As a result, we have $M = 15$. The three databases are independent of each other.

In NIST LRE tasks, there are 3 different duration settings, 3, 10, and 30 seconds. In this paper, we only conduct experiments on test queries of 3 seconds. The 1996, 2003 and 2005 evaluation data consist of 1,503, 1,280 and 3,662 test sessions respectively. In classifier design, each conversation in the *CallFriend* database is segmented into overlapping sessions, resulting in about 12,000 sessions per language. In the baseline experiment, we don't use query expansion. The results are previously reported in [1].

A series of systematic experiments on the GMM size suggest us 512 mixtures for m^- and 64 mixtures for m^+ because m^- has much more training data than m^+ has in our experiments. For a system of 15 target languages, the pairwise SVM ensemble classifiers reduce the high dimensional *bag-of-sounds* vectors to $15 \times (15-1)/2 = 105$ dimensional output vectors, from which we further train GMM models $\{m^+, m^-\}$ for each language.

In the query expansion by relevant utterances, we consider top 40 utterances from the reference corpus as the relevant feedback to augment the initial query. In the query expansion by relevant clusters, we use top 2 clusters from the reference corpus. In the query expansion by reweighting query terms, we use top 20 utterances as the relevant feedbacks and the bottom 20 utterances as the non-relevant feedbacks, with $\beta = 0.75$ and $\gamma = 0.25$ which are set empirically as suggested in [6]. The experiment results are reported in Table 1.

Table 1. Equal error rates (EER%) for 3-second test queries of NIST 1996, 2003, 2005 LRE tasks.

	1996	2003	2005
Baseline	21.16	21.25	24.23
By relevant utterances	19.35	20.54	22.77
By relevant clusters	17.91	20.12	22.18
By reweighting query terms	17.72	20.20	21.77

The results suggest that query expansion by reweighting the query terms works the best among the three strategies with an EER reduction of 16.2%, 4.9% and 10.2% relative to the baseline results on NIST 1996, 2003 and 2005 LRE tasks respectively. Without surprise, cluster-based feedbacks provide slightly more reliable statistics than utterance-based feedbacks, resulting in improved performance. The results are also consistent with the observations in text-based information retrieval literature.

If we consider the traditional query expansion in text information retrieval as having supervised relevance feedbacks, then the proposed speech query expansion in this paper can be seen as having unsupervised relevance feedbacks.

5. CONCLUSION

We have studied three speech query expansion methods for spoken language recognition. The experiment results show that all the three methods are effective in LID task. The query expansion approach is especially useful when test query is short. Although we only experiment on 3 second queries in this paper, the same methods are applicable to longer queries as well. In this paper, we discuss query expansion in the context of phonotactical vector-based LID. Without loss of generality, it can be extended to any other common LID frameworks, such as acoustic GMM, Spectral SVM, with minor modification of the query expansion procedure.

The query expansion provides a way to improve LID performance at a low cost. Nonetheless, the quality of the reference corpus may affect the relevance feedbacks. It is important that we choose a reference corpus that is close to the test task in terms of acoustic or channel conditions. We will extend the proposed methods towards other LID frameworks and study the effects of reference corpus in the future work.

6. REFERENCES

[1] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, 2007

[2] A. Martin, M. Przybocki, "NIST 2003 Language Recognition Evaluation", *Proc. Eurospeech*, 2003

[3] E. N. Efthimiadis, "A User-centered Evaluation of Ranking Algorithms for Interactive Query Expansion", *Proc. 16th ACM SIGIR* pp.146-159, 1993

[4] S. E. Robertson, "On Term Selection for Query Expansion," *Journal of Documentation*, vol 46, no. 4, pp. 359-364, 1990

[5] D. Harman, "Relevance Feedback Revisited", *Proc. 15th ACM SIGIR*, 1992

[6] J. J. Rocchio, "Relevance Feedback in Information Retrieval", in G. Salton (Ed.), *The SMART Retrieval System*, pp.313-323, Englewood Cliffs, N.J.: Prentice-Hall, Inc.

[7] H. Li, B. Ma, and R. Tong, "Spoken Language Recognition with Output Coding", *Interspeech* 2006, Sept 17-21, 2006