A Phonotactic Language Model for Spoken Language Identification

Haizhou Li *and* **Bin Ma**
Institute for Infocomm Research
Singapore 119613
{hli,mabin}@i2r.a-star.edu.sg

Abstract

We have established a phonotactic language model as the solution to spoken language identification (LID). In this framework, we define a single set of acoustic tokens to represent the acoustic activities in the world's spoken languages. A voice tokenizer converts a spoken document into a text-like document of acoustic tokens. Thus a spoken document can be represented by a count vector of acoustic tokens and token *n*-grams in the vector space. We apply *latent semantic analysis* to the vectors, in the same way that it is applied in information retrieval, in order to capture salient phonotactics present in spoken documents. The vector space modeling of spoken utterances constitutes a paradigm shift in LID technology and has proven to be very successful. It presents a 12.4% error rate reduction over one of the best reported results on the 1996 NIST Language Recognition Evaluation database.

1 Introduction

Spoken language and written language are similar in many ways. Therefore, much of the research in spoken language identification, LID, has been inspired by text-categorization methodology. Both text and voice are generated from language dependent vocabulary. For example, both can be seen as stochastic time-sequences corrupted by a channel noise. The *n*-gram language model has achieved equal amounts of success in both tasks, e.g. *n*-character slice for text categorization by language (Cavnar and Trenkle, 1994) and Phone Recognition followed by *n*-gram Language Modeling, or PRLM (Zissman, 1996) .

Orthographic forms of language, ranging from Latin alphabet to Cyrillic script to Chinese characters, are far more unique to the language than their phonetic counterparts. From the speech production point of view, thousands of spoken languages from all over the world are phonetically articulated using only a few hundred distinctive sounds or phonemes (Hieronymus, 1994). In other words, common sounds are shared considerably across different spoken languages. In addition, spoken documents[1], in the form of digitized wave files, are far less structured than written documents and need to be treated with techniques that go beyond the bounds of written language. All of this makes the identification of spoken language based on phonetic units much more challenging than the identification of written language. In fact, the challenge of LID is inter-disciplinary, involving digital signal processing, speech recognition and natural language processing.

In general, a LID system usually has three fundamental components as follows:

1) A voice tokenizer which segments incoming voice feature frames and associates the segments with acoustic or phonetic labels, called tokens;
2) A statistical language model which captures language dependent phonetic and phonotactic information from the sequences of tokens;
3) A language classifier which identifies the language based on discriminatory characteristics of acoustic score from the voice tokenizer and phonotactic score from the language model.

In this paper, we present a novel solution to the three problems, focusing on the second and third problems from a computational linguistic perspective. The paper is organized as follows: In Section 2, we summarize relevant existing approaches to the LID task. We highlight the shortcomings of existing approaches and our attempts to address the

[1] A spoken utterance is regarded as a spoken document in this paper.

Proceedings of the 43rd Annual Meeting of the ACL, pages 515–522,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

issues. In Section 3 we propose the *bag-of-sounds* paradigm to turn the LID task into a typical text categorization problem. In Section 4, we study the effects of different settings in experiments on the 1996 NIST Language Recognition Evaluation (LRE) database[2]. In Section 5, we conclude our study and discuss future work.

2 Related Work

Formal evaluations conducted by the National Institute of Science and Technology (NIST) in recent years demonstrated that the most successful approach to LID used the phonotactic content of the voice signal to discriminate between a set of languages (Singer *et al.*, 2003). We briefly discuss previous work cast in the formalism mentioned above: tokenization, statistical language modeling, and language identification. A typical LID system is illustrated in Figure 1 (Zissman, 1996), where language dependent voice tokenizers (VT) and language models (LM) are deployed in the Parallel PRLM architecture, or P-PRLM.

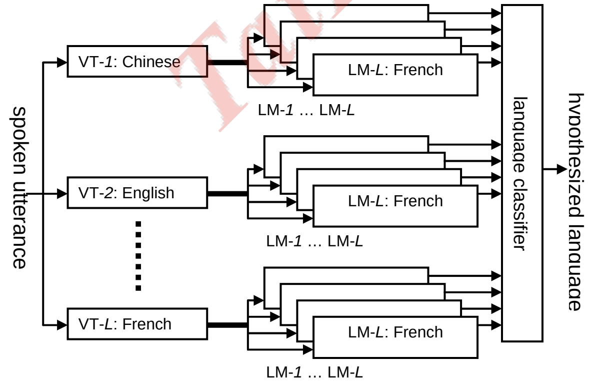

Figure 1. *L* monolingual phoneme recognition front-ends are used in parallel to tokenize the input utterance, which is analyzed by LMs to predict the spoken language

2.1 Voice Tokenization

A voice tokenizer is a speech recognizer that converts a spoken document into a sequence of tokens. As illustrated in Figure 2, a token can be of different sizes, ranging from a speech feature frame, to a phoneme, to a lexical word. A token is defined to describe a distinct acoustic/phonetic activity. In early research, low level spectral frames, which are assumed to be independent of each other, were used as a set of prototypical spectra for each language (Sugiyama, 1991). By adopting hidden Markov models, people moved beyond low-level spectral analysis towards modeling a frame sequence into a larger unit such as a phoneme and even a lexical word.

Since the lexical word is language specific, the phoneme becomes the natural choice when building a language-independent voice tokenization front-end. Previous studies show that parallel language-dependent phoneme tokenizers effectively serve as the tokenization front-ends with P-PRLM being the typical example. However, a language-independent phoneme set has not been explored yet experimentally. In this paper, we would like to explore the potential of voice tokenization using a unified phoneme set.

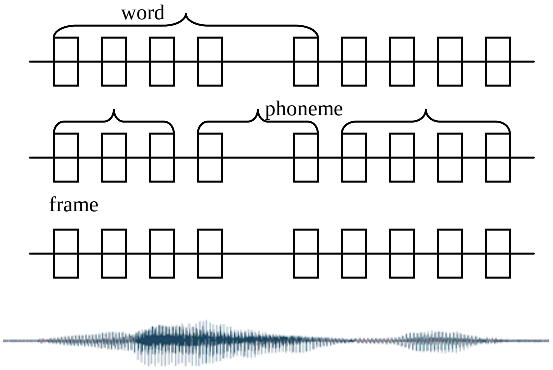

Figure 2 Tokenization at different resolutions

2.2 *n*-gram Language Model

With the sequence of tokens, we are able to estimate an *n*-gram language model (LM) from the statistics. It is generally agreed that phonotactics, i.e. the rules governing the phone/phonemes sequences admissible in a language, carry more language discriminative information than the phonemes themselves. An *n*-gram LM over the tokens describes well *n*-local phonotactics among neighboring tokens. While some systems model the phonotactics at the frame level (Torres-Carrasquillo *et al.*, 2002), others have proposed P-PRLM. The latter has become one of the most promising solutions so far (Zissman, 1996).

A variety of cues can be used by humans and machines to distinguish one language from another. These cues include phonology, prosody, morphology, and syntax in the context of an utterance.

[2] http://www.nist.gov/speech/tests/

However, global phonotactic cues at the level of utterance or spoken document remains unexplored in previous work. In this paper, we pay special attention to it. A spoken language always contains a set of high frequency function words, prefixes, and suffixes, which are realized as phonetic token substrings in the spoken document. Individually, those substrings may be shared across languages. However, the pattern of their co-occurrences discriminates one language from another.

Perceptual experiments have shown (Muthusamy, 1994) that with adequate training, human listeners' language identification ability increases when given longer excerpts of speech. Experiments have also shown that increased exposure to each language and longer training sessions improve listeners' language identification performance. Although it is not entirely clear how human listeners make use of the high-order phonotactic/prosodic cues present in longer spans of a spoken document, strong evidence shows that phonotactics over larger context provides valuable LID cues beyond *n*-gram, which will be further attested by our experiments in Section 4.

2.3 Language Classifier

The task of a language classifier is to make good use of the LID cues that are encoded in the model λ_l to hypothesize $\hat{l}$ from among *L* languages, Λ, as the one that is actually spoken in a spoken document *O*. The LID model λ_l in P-PRLM refers to extracted information from acoustic model and *n*-gram LM for language *l*. We have $\lambda_l = \{\lambda_l^{AM}, \lambda_l^{LM}\}$ and $\lambda_l \in \Lambda \ (l = 1,...,L)$. A maximum-likelihood classifier can be formulated as follows:

$$\begin{aligned}\hat{l} &= \arg\max_{l\in\Lambda} P(O/\lambda_l) \\ &\approx \arg\max_{l\in\Lambda} \sum_{T\in\Gamma} P\left(O/T,\lambda_l^{AM}\right)P\left(T/\lambda_l^{LM}\right)\end{aligned} \quad (1)$$

The exact computation in Eq.(1) involves summing over all possible decoding of token sequences $T \in \Gamma$ given *O*. In many implementations, it is approximated by the maximum over all sequences in the sum by finding the most likely token sequence, $\hat{T}_l$, for each language *l*, using the Viterbi algorithm:

$$\hat{l} \approx \arg\max_{l\in\Lambda} [P\left(O/\hat{T}_l,\lambda_l^{AM}\right)P\left(\hat{T}_l/\lambda_l^{LM}\right)] \quad (2)$$

Intuitively, individual sounds are heavily shared among different spoken languages due to the common speech production mechanism of humans. Thus, the acoustic score has little language discriminative ability. Many experiments (Yan and Barnard, 1995; Zissman, 1996) have further attested that the *n*-gram LM score provides more language discriminative information than their acoustic counterparts. In Figure 1, the decoding of voice tokenization is governed by the acoustic model λ_l^{AM} to arrive at an acoustic score $P\left(O/\hat{T}_l,\lambda_l^{AM}\right)$ and a token sequence $\hat{T}_l$. The *n*-gram LM derives the *n*-local phonotactic score $P\left(\hat{T}_l/\lambda_l^{LM}\right)$ from the language model λ_l^{LM}.

Clearly, the *n*-gram LM suffers the major shortcoming of having not exploited the global phonotactics in the larger context of a spoken utterance. Speech recognition researchers have so far chosen to only use *n*-gram local statistics for primarily pragmatic reasons, as this *n*-gram is easier to attain. In this work, a language independent voice tokenization front-end is proposed, that uses a unified acoustic model λ^{AM} instead of multiple language dependent acoustic models λ_l^{AM}. The *n*-gram LM λ_l^{LM} is generalized to model both local and global phonotactics.

3 *Bag-of-Sounds* Paradigm

The *bag-of-sounds* concept is analogous to the *bag-of-words* paradigm originally formulated in the context of information retrieval (IR) and text categorization (TC) (Salton 1971; Berry *et al.*, 1995; Chu-Caroll and Carpenter, 1999). One focus of IR is to extract informative features for document representation. The *bag-of-words* paradigm represents a document as a vector of counts. It is believed that it is not just the words, but also the co-occurrence of words that distinguish semantic domains of text documents.

Similarly, it is generally believed in LID that, although the sounds of different spoken languages overlap considerably, the phonotactics differentiates one language from another. Therefore, one can easily draw the analogy between an acoustic token in *bag-of-sounds* and a word in *bag-of-words*. Unlike words in a text document, the phonotactic information that distinguishes spoken languages is concealed in the sound waves of spoken languages. After transcribing a spoken document into a text like document of tokens, many IR or TC techniques can then be readily applied.

It is beyond the scope of this paper to discuss what would be a good voice tokenizer. We adopt phoneme size language-independent acoustic tokens to form a unified acoustic vocabulary in our voice tokenizer. Readers are referred to (Ma *et al.*, 2005) for details of acoustic modeling.

3.1 Vector Space Modeling

In human languages, some words invariably occur more frequently than others. One of the most common ways of expressing this idea is known as Zipf's Law (Zipf, 1949). This law states that there is always a set of words which dominates most of the other words of the language in terms of their frequency of use. This is true both of written words and of spoken words. The short-term, or *local phonotactics*, is devised to describe Zipf's Law.

The local phonotactic constraints can be typically described by the token *n*-grams, or phoneme *n*-grams as in (Ng *et al*., 2000), which represents short-term statistics such as lexical constraints. Suppose that we have a token sequence, *t1 t2 t3 t4*. We derive the unigram statistics from the token sequence itself. We derive the bigram statistics from *t1(t2) t2(t3) t3(t4) t4(#)* where the token vocabulary is expanded over the token's right context. Similarly, we derive the trigram statistics from the *t1(#,t2) t2(t1,t3) t3(t2,t4) t4(t3,#)* to account for left and right contexts. The # sign is a place holder for free context. In the interest of manageability, we propose to use up to token trigram. In this way, for an acoustic system of Y tokens, we have potentially Y^2 bigram and Y^3 trigram in the vocabulary.

Meanwhile, motivated by the ideas of having both short-term and long-term phonotactic statistics, we propose to derive *global phonotactics* information to account for long-term phonotactics:

The global phonotactic constraint is the high-order statistics of *n*-grams. It represents document level long-term phonotactics such as co-occurrences of *n*-grams. By representing a spoken document as a count vector of *n*-grams, also called *bag-of-sounds* vector, it is possible to explore the relations and higher-order statistics among the diverse *n*-grams through *latent semantic analysis* (LSA).

It is often advantageous to weight the raw counts to refine the contribution of each *n*-gram to LID. We begin by normalizing the vectors representing the spoken document by making each vector of unit length. Our second weighting is based on the notion that an *n*-gram that only occurs in a few languages is more discriminative than an *n*-gram that occurs in nearly every document. We use the *inverse-document frequency* (*idf*) weighting scheme (Spark Jones, 1972), in which a word is weighted inversely to the number of documents in which it occurs, by means of $idf(w) = \log D / d(w)$, where *w* is a word in the vocabulary of *W* token *n*-grams. *D* is the total number of documents in the training corpus from *L* languages. Since each language has at least one document in the training corpus, we have $D \geq L$. $d(w)$ is the number of documents containing the word *w*. Letting $c_{w,d}$ be the count of word *w* in document *d*, we have the weighted count as

$$c'_{w,d} = c_{w,d} \times idf(w) / (\sum_{1 \le w' \le W} c^2_{w',d})^{1/2} \qquad (3)$$

and a vector $c_d = \{c'_{1,d}, c'_{2,d}, ..., c'_{W,d}\}^T$ to represent document *d*. A corpus is then represented by a *term-document* matrix $H = \{c_1, c_2, ..., c_D\}$ of $W \times D$.

3.2 Latent Semantic Analysis

The fundamental idea in LSA is to reduce the dimension of a document vector, *W* to *Q*, where $Q << W$ and $Q << D$, by projecting the problem into the space spanned by the rows of the closest rank-*Q* matrix to *H* in the Frobenius norm (Deerwester *et al*, 1990). Through singular value decomposition (SVD) of *H*, we construct a modified matrix H_Q from the *Q*-largest singular values:

$$H_Q = U_Q S_Q V_Q^T \qquad (4)$$

U_Q is a $W \times Q$ left singular matrix with rows $u_w, 1 \le w \le W$; S_Q is a $Q \times Q$ diagonal matrix of *Q*-largest singular values of *H*; V_Q is $D \times Q$ right singular matrix with rows v_d, $1 \le d \le D$.

With the SVD, we project the *D* document vectors in *H* into a reduced space V_Q, referred to as *Q*-space in the rest of this paper. A test document c_p of unknown language ID is mapped to a pseudo-document v_p in the *Q*-space by matrix U_Q

$$c_p \rightarrow v_p = c_p^T U_Q S_Q^{-1} \quad (5)$$

After SVD, it is straightforward to arrive at a natural metric for the closeness between two spoken documents v_i and v_j in *Q*-space instead of their original *W*-dimensional space c_i and c_j.

$$g(c_i, c_j) \approx \cos(v_i, v_j) = \frac{v_i \cdot v_j^T}{\| v_i \| \cdot \| v_j \|} \quad (6)$$

$g(c_i, c_j)$ indicates the similarity between two vectors, which can be transformed to a distance measure $k(c_i, c_j) = \cos^{-1} g(c_i, c_j)$.

In the forced-choice classification, a test document, supposedly monolingual, is classified into one of the *L* languages. Note that the test document is unknown to the *H* matrix. We assume consistency between the test document's intrinsic phonotactic pattern and one of the *D* patterns, that is extracted from the training data and is presented in the *H* matrix, so that the SVD matrices still apply to the test document, and Eq.(5) still holds for dimension reduction.

3.3 *Bag-of-Sounds* Language Classifier

The *bag-of-sounds* phonotactic LM benefits from several properties of vector space modeling and LSA.

1) It allows for representing a spoken document as a vector of *n*-gram features, such as unigram, bigram, trigram, and the mixture of them;
2) It provides a well-defined distance metric for measurement of phonotactic distance between spoken documents;
3) It processes spoken documents in a lower dimensional *Q*-space, that makes the *bag-of-sounds* phonotactic language modeling, λ_l^{LM}, and classification computationally manageable.

Suppose we have only one prototypical vector c_l and its projection in the *Q*-space v_l to represent language *l*. Applying LSA to the *term-document* matrix $H: W \times L$, a minimum distance classifier is formulated:

$$\hat{l} = \arg\min_{l \in \Lambda} k(v_p, v_l) \quad (7)$$

In Eq.(7), v_p is the *Q*-space projection of c_p, a test document.

Apparently, it is very restrictive for each language to have just one prototypical vector, also referred to as a centroid. The pattern of language distribution is inherently multi-modal, so it is unlikely well fitted by a single vector. One solution to this problem is to span the language space with multiple vectors. Applying LSA to a *term-document* matrix $H: W \times L'$, where $L' = L \times M$ assuming each language *l* is represented by a set of *M* vectors, Φ_l, a new classifier, using *k*-nearest neighboring rule (Duda and Hart, 1973) , is formulated, named *k*-nearest classifier (KNC):

$$\hat{l} = \arg\min_{l \in \Lambda} \sum_{l' \in \phi_l} k(v_p, v_{l'}) \quad (8)$$

where ϕ_l is the set of *k*-nearest-neighbor to v_p and $\phi_l \subset \Phi_l$.

Among many ways to derive the *M* centroid vectors, here is one option. Suppose that we have a set of training documents D_l for language l, as subset of corpus Ω, $D_l \subset \Omega$ and $\cup_{l=1}^{L} D_l = \Omega$. To derive the *M* vectors, we choose to carry out vector quantization (VQ) to partition D_l into *M* cells $D_{l,m}$ in the *Q*-space such that $\cup_{m=1}^{M} D_{l,m} = D_l$ using similarity metric Eq.(6). All the documents in each cell $D_{l,m}$ can then be merged to form a super-document, which is further projected into a *Q*-space vector $v_{l,m}$. This results in *M* prototypical centroids $v_{l,m} \in \Phi_l$ $(m = 1,...M)$. Using KNC, a test vector is compared with *M* vectors to arrive at the *k*-nearest neighbors for each language, which can be computationally expensive when *M* is large.

Alternatively, one can account for multi-modal distribution through finite mixture model. A mixture model is to represent the *M* discrete components with soft combination. To extend the KNC into a statistical framework, it is necessary to map our distance metric Eq.(6) into a probability measure. One way is for the distance measure to induce a family of exponential distributions with pertinent marginality constraints. In practice, what we need is a reasonable probability distribution, which sums to one, to act as a lookup table for the distance measure. We here choose to use the empirical multivariate distribution constructed by allocating the total probability mass in proportion to the distances observed with the training data. In short, this reduces the task to a *histogram normalization*. In this way, we map the distance $k(c_i, c_j)$ to a conditional probability distribution $p(v_i | v_j)$

subject to $\sum_{i=1}^{|\Omega|} p(v_i \mid v_j) = 1$. Now that we are in the probability domain, techniques such as mixture smoothing can be readily applied to model a language class with finer fitting.

Let's re-visit the task of L language forced-choice classification. Similar to KNC, suppose we have M centroids $v_{l,m} \in \Phi_l$ $(m = 1,...M)$ in the Q-space for each language l. Each centroid represents a class. The class conditional probability can be described as a linear combination of $p(v_i \mid v_{l,m})$:

$$p(v_i \mid \lambda_l^{LM}) = \sum_{m=1}^{M} p(v_{l,m}) p(v_i \mid v_{l,m}) \qquad (9)$$

the probability $p(v_{l,m})$, functionally serves as a mixture weight of $p(v_i \mid v_{l,m})$. Together with a set of centroids $v_{l,m} \in \Phi_l$ $(m = 1,...M)$, $p(v_i \mid v_{l,m})$ and $p(v_{l,m})$ define a mixture model λ_l^{LM}. $p(v_i \mid v_{l,m})$ is estimated by *histogram normalization* and $p(v_{l,m})$ is estimated under the maximum likelihood criteria, $p(v_{l,m}) = C_{m,l} / C_l$, where C_l is total number of documents in D_l, of which $C_{m,l}$ documents fall into the cell m.

An *Expectation-Maximization* iterative process can be devised for training of λ_l^{LM} to maximize the likelihood Eq.(9) over the entire training corpus:

$$p(\Omega \mid \Lambda) = \prod_{l=1}^{L} \prod_{d=1}^{|D_l|} p(v_d \mid \lambda_l^{LM}) \qquad (10)$$

Using the phonotactic LM score $P\left(\hat{T}_l / \lambda_l^{LM}\right)$ for classification, with $\hat{T}_l$ being represented by the *bag-of-sounds* vector v_p, Eq.(2) can be reformulated as Eq.(11), named mixture-model classifier (MMC):

$$\begin{aligned} \hat{l} &= \arg\max_{l \in \Lambda} p(v_p \mid \lambda_l^{LM}) \\ &= \arg\max_{l \in \Lambda} \sum_{m=1}^{M} p(v_{l,m}) p(v_p \mid v_{l,m}) \end{aligned} \qquad (11)$$

To establish fair comparison with P-PRLM, as shown in Figure 3, we devise our *bag-of-sounds* classifier to solely use the LM score $P\left(\hat{T}_l / \lambda_l^{LM}\right)$ for classification decision whereas the acoustic score $P\left(O / \hat{T}_l, \lambda_l^{AM}\right)$ may potentially help as reported in (Singer *et al.*, 2003).

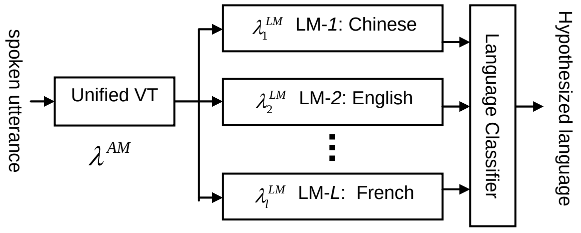
spoken utterance
Unified VT
λ^{AM}
λ_1^{LM} LM-*1*: Chinese
λ_2^{LM} LM-*2*: English
λ_l^{LM} LM-*L*: French
Language Classifier
Hypothesized language

Figure 3. A *bag-of-sounds* classifier. A unified front-end followed by L parallel *bag-of-sounds* phonotactic LMs.

4 Experiments

This section will experimentally analyze the performance of the proposed *bag-of-sounds* framework using the 1996 NIST Language Recognition Evaluation (LRE) data. The database was intended to establish a baseline of performance capability for language recognition of conversational telephone speech. The database contains recorded speech of 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. We use the training set and development set from LDC *CallFriend* corpus[3] as the training data. Each conversation is segmented into overlapping sessions of about 30 seconds each, resulting in about 12,000 sessions for each language. The evaluation set consists of 1,492 30-sec sessions, each distributed among the various languages of interest. We treat a 30-sec session as a spoken document in both training and testing. We report error rates (ER) of the 1,492 test trials.

4.1 Effect of Acoustic Vocabulary

The choice of *n*-gram affects the performance of LID systems. Here we would like to see how a better choice of acoustic vocabulary can help convert a spoken document into a phonotactically discriminative space. There are two parameters that determine the acoustic vocabulary: the choice of acoustic token, and the choice of *n*-grams. In this paper, the former concerns the size of an acoustic system Y in the unified front-end. It is studied in more details in (Ma *et al.*, 2005). We set Y to 32 in

[3] See http://www.ldc.upenn.edu/. The overlap between 1996 NIST evaluation data and *CallFriend* database has been removed from training data as suggested in the 2003 NIST LRE website http://www.nist.gov/speech/tests/index.htm

this experiment; the latter decides what features to be included in the vector space. The vector space modeling allows for multiple heterogeneous features in one vector. We introduce three types of acoustic vocabulary (AV) with mixture of token unigram, bigram, and trigram:

a) AV1: 32 broad class phonemes as unigram, selected from 12 languages, also referred to as P-ASM as detailed in (Ma *et al*., 2005)
b) AV2: AV1 augmented by 32 × 32 bigrams of AV1, amounting to 1,056 tokens
c) AV3: AV2 augmented by 32 × 32 × 32 trigrams of AV1, amounting to 33,824 tokens

	AV1	AV2	AV3
ER %	46.1	32.8	28.3

Table 1. Effect of acoustic vocabulary (KNC)

We carry out experiments with KNC classifier of 4,800 centroids. Applying *k*-nearest-neighboring rule, *k* is empirically set to 3. The error rates are reported in Table 1 for the experiments over the three AV types. It is found that high-order token *n*-grams improve LID performance. This reaffirms many previous findings that *n*-gram phonotactics serves as a valuable cue in LID.

4.2 Effect of Model Size

As discussed in KNC, one would expect to improve the phonotactic model by using more centroids. Let's examine how the number of centroid vectors *M* affects the performance of KNC. We set the acoustic system size *Y* to 128, *k*-nearest to 3, and only use token bigrams in the *bag-of-sounds* vector. In Table 2, it is not surprising to find that the performance improves as *M* increases. However, it is not practical to have large *M* because $L' = L \times M$ comparisons need to take place in each test trial.

#*M*	1,200	2,400	4,800	12,000
ER %	17.0	15.7	15.4	14.8

Table 2. Effect of number of centroids (KNC)

To reduce computation, MMC attempts to use less number of mixtures *M* to represent the phonotactic space. With the smoothing effect of the mixture model, we expect to use less computation to achieve similar performance as KNC. In the experiment reported in Table 3, we find that MMC (*M*=1,024) achieves 14.9% error rate, which almost equalizes the best result in the KNC experiment (*M*=12,000) with much less computation.

#*M*	4	16	64	256	1,024
ER %	29.6	26.4	19.7	16.0	14.9

Table 3. Effect of number of mixtures (MMC)

4.3 Discussion

The *bag-of-sounds* approach has achieved equal success in both 1996 and 2003 NIST LRE databases. As more results are published on the 1996 NIST LRE database, we choose it as the platform of comparison. In Table 4, we report the performance across different approaches in terms of error rate for a quick comparison. MMC presents a 12.4% ER reduction over the best reported result[4] (Torres-Carrasquillo *et al*., 2002).

It is interesting to note that the *bag-of-sounds* classifier outperforms its P-PRLM counterpart by a wide margin (14.9% vs 22.0%). This is attributed to the global phonotactic features in λ_l^{LM}. The performance gain in (Torres-Carrasquillo *et al*., 2002; Singer *et al*., 2003) was obtained mainly by fusing scores from several classifiers, namely GMM, P-PRLM and SVM, to benefit from both acoustic and language model scores. Noting that the *bag-of-sounds* classifier in this work solely relies on the LM score, it is believed that fusing with scores from other classifiers will further boost the LID performance.

	ER %
P-PRLM[5]	22.0
P-PRLM + GMM acoustic[5]	19.5
P-PRLM + GMM acoustic + GMM tokenizer[5]	17.0
Bag-of-sounds classifier (MMC)	14.9

Table 4. Benchmark of different approaches

Besides the error rate reduction, the *bag-of-sounds* approach also simplifies the on-line computing procedure over its P-PRLM counterpart. It would be interesting to estimate the on-line computational need of MMC. The cost incurred has two main components: 1) the construction of the

[4] Previous results are also reported in DCF, DET, and equal error rate (EER). Comprehensive benchmarking for *bag-of-sounds* phonotactic LM will be reported soon.

[5] Results extracted from (Torres-Carrasquillo *et al*., 2002)

pseudo document vector, as done via Eq.(5); 2) $L' = L \times M$ vector comparisons. The computing cost is estimated to be $\mathcal{O}(Q^2)$ per test trial (Bellegarda, 2000). For typical values of Q, this amounts to less than 0.05 Mflops. While this is more expensive than the usual table look-up in conventional *n*-gram LM, the performance improvement is able to justify the relatively modest computing overhead.

5 Conclusion

We have proposed a phonotactic LM approach to LID problem. The concept of *bag-of-sounds* is introduced, for the first time, to model phonotactics present in a spoken language over a larger context. With *bag-of-sounds* phonotactic LM, a spoken document can be treated as a text-like document of acoustic tokens. This way, the well-established LSA technique can be readily applied. This novel approach not only suggests a paradigm shift in LID, but also brings 12.4% error rate reduction over one of the best reported results on the 1996 NIST LRE data. It has proven to be very successful.

We would like to extend this approach to other spoken document categorization tasks. In monolingual spoken document categorization, we suggest that the semantic domain can be characterized by latent phonotactic features. Thus it is straightforward to extend the proposed *bag-of-sounds* framework to spoken document categorization.

Acknowledgement

The authors are grateful to Dr. Alvin F. Martin of the NIST Speech Group for his advice when preparing the 1996 NIST LRE experiments, to Dr G. M. White and Ms Y. Chen of Institute for Infocomm Research for insightful discussions.

References

Jerome R. Bellegarda. 2000. *Exploiting latent semantic information in statistical language modeling*, In Proc. of the IEEE, 88(8):1279-1296.

M. W. Berry, S.T. Dumais and G.W. O'Brien. 1995. *Using Linear Algebra for intelligent information retrieval*, SIAM Review, 37(4):573-595.

William B. Cavnar, and John M. Trenkle. 1994. *N-Gram-Based Text Categorization*, In Proc. of 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161-169.

Jennifer Chu-Carroll, and Bob Carpenter. 1999. *Vector-based Natural Language Call Routing*, Computational Linguistics, 25(3):361-388.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, 1990, *Indexing by latent semantic analysis*, *Journal of the American Society for Informatin Science*, 41(6):391-407

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and scene analysis*. John Wiley & Sons

James L. Hieronymus. 1994. *ASCII Phonetic Symbols for the World's Languages: Worldbet*. Technical Report AT&T Bell Labs.

Spark Jones, K. 1972. *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, 28:11-20

Bin Ma, Haizhou Li and Chin-Hui Lee, 2005. *An Acoustic Segment Modeling Approach to Automatic Language Identification*, submitted to Interspeech 2005

Yeshwant K. Muthusamy, Neena Jain, and Ronald A. Cole. 1994. *Perceptual benchmarks for automatic language identification*, In Proc. of ICASSP

Corinna Ng , Ross Wilkinson , Justin Zobel, 2000. *Experiments in spoken document retrieval using phoneme n-grams*, Speech Communication, 32(1-2):61-77

G. Salton, 1971. T*he SMART Retrieval System*, Prentice-Hall, Englewood Cliffs, NJ, 1971

E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell and D.A. Reynolds. 2003. *Acoustic, Phonetic and Discriminative Approaches to Automatic language recognition*, In Proc. of Eurospeech

Masahide Sugiyama. 1991. *Automatic language recognition using acoustic features*, In Proc. of ICASSP.

Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, and J.R. Deller. Jr. 2002. *Language identification using Gaussian Mixture model tokenization*, in Proc. of ICASSP.

Yonghong Yan, and Etienne Barnard. 1995. *An approach to automatic language identification based on language dependent phone recognition*, In Proc. of ICASSP.

George K. Zipf. 1949. *Human Behavior and the Principal of Least effort, an introduction to human ecology*. Addison-Wesley, Reading, Mass.

Marc A. Zissman. 1996. *Comparison of four approaches to automatic language identification of telephone speech*, IEEE Trans. on Speech and Audio Processing, 4(1):31-44.