USING LOCAL & GLOBAL PHONOTACTIC FEATURES IN CHINESE DIALECT IDENTIFICATION

Boon Pang LIM , Haizhou LI, Bin MA

{bplim,hli,mabin}@i2r.a-star.edu.sg
Institute for Infocomm Research, Republic of Singapore

ABSTRACT

Conventional techniques for spoken language identification use variants of phone similarity and language model scoring, which represent local phonetic constraints in spoken language. In this paper, we explore the identification of Chinese dialects which share the same written script and have similar sound systems and syllable structures. As such, local phonetic constraints do not provide enough discriminative information among the dialects. We propose to use Latent Semantic Analysis (LSA) to extract global features that represent the high-order statistics in the co-occurrence of sounds. The experiments show that we can achieve the best performance by combining acoustic, n-gram language modeling and LSA scores. An accuracy of 99.23% is achieved in 4-way classification tests using 20-second speech sessions.

1. INTRODUCTION

Among the more successful systems that have recently been built for language identification (LID) are those that perform phone recognition followed by n-gram language modelling (PRLM) [1]. Variations of this approach include using longer acoustic units [2] or integrating other types of information from stress or articulatory models [3, 4]. Meanwhile, comparable performance has also been achieved with LID systems based on large vocabulary continuous speech recognizers (LVCSR) [5, 6].

LVCSR-based systems and PRLM are similar in that they both generate feature scores by performing acoustic matching followed by n-gram language model scoring. We refer to such scores as local acoustic-phonotactic scores. Building an LVCSR is a labor intensive process that requires an extensive amount of language specific expertise. However, LVCSRs are more refined in that they apply lexical constraints during the decoding process. Thus, they produce more plausible phone decodings that can translate to better language identification performance. Our work explores the use of features easily derived from LVCSR-based syllable decoders to distinguish between three closely related dialects of Chinese (Mandarin, Cantonese and Shanghai dialect).

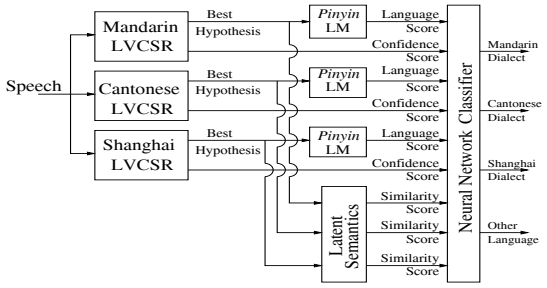


Fig. 1. Language Identification Architecture.

All Chinese dialects are syllabic languages. Thus it is natural to use the syllable as the basic acoustic unit in decoding. These dialects not only share a common written script, but also share similar vocabularies and word usage patterns. In some cases, a word may even have the same pronounciation in all three languages. These similarities make local scores less effective as discriminative features. In this paper, we introduce the use of Latent Semantic Analysis (LSA) [7] towards capturing large span phonotactic information within speech sessions in order to complement local acoustic-phonotactic scores. Our experiments with 4-way classification of 3 Chinese dialects and an *out-of-language* (OOL) set with 6 other languages show that this fusion of local and global features is extremely effective. We also investigate one method of improving the discrimination and robustness of linguistic feature scores.

2. LANGUAGE IDENTIFICATION OVERVIEW

Fig. 1 shows a 4-way language identification system based on an LVCSR (Abacus). Three LVCSRs are run in parallel as syllable decoders, one for each dialect. The Abacus speech recognizer, shown in Fig. 2, is a frame-synchronous HMM-based LVCSR engine that employs class-based n-gram language models. The signal processing front-end emits MFCC feature vectors in accordance with the ETSI standard. These cepstra are fed into two parallel acoustic decoders employing acoustic models of different granularity; a sharp acoustic model using three-state HMMs for a set

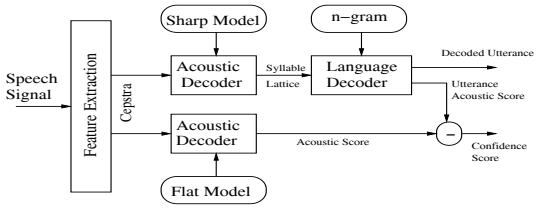


Fig. 2. Speech Recognizer Architecture.

of tied-state context-dependant triphones, and a flat acoustic model using broad classes of context independent phones, in which similar phones (such as nasals, fricatives or plosives) are put under the same class. A language decoder, assisted by a bigram language model, performs n-best decoding on the syllable lattice output by the acoustic decoder.

2.1. Features Representing Local Constraints

Two types of feature scores are readily available as intermediate outputs from the recognizer. Given an utterance with K non-silent frames of speech with cepstral features $o_1 ... o_K$, the acoustic confidence score is

$$C = \sum_{k=1}^{K} \left[\log P(o_k|\lambda) - \log P(o_k|\bar{\lambda}) \right] / K. \qquad (1)$$

which is a log likelihood ratio of acoustic observations between the sharp and flat acoustic models. Here, $P(o_k|\lambda)$ is the probability of observing cepstral features o_k in frame k given the best matching phone sequence λ output by the sharp acoustic model decoder and $P(o_k|\bar{\lambda})$ corresponds to the probability for the flat model. Acoustic confidence measures how likely sounds in an utterance belong to a specific language; a higher value indicates a better match.

A suitable linguistically-based score is the cross-entropy

$$H_P(W) = -\frac{1}{N} \log \prod_{l=1}^{N} P(w_l|w_1 ... w_{l-1}). \qquad (2)$$

Here, P's are n-gram probabilities from an appropriately trained trigram language model, and $W = (w_1 ... w_N)$ is the sequence of sounds for the best decoding of the utterance. Cross-entropy measures how well a decoding matches a language model, where a lower value indicates a better match.

2.2. Features Representing Global Constraints

Latent Semantic Analysis is a statistical technique which uses a *bag-of-words* approach to derive concepts from term co-occurrence data and perform classification [7, 8]. We consider an analogous *bag-of-sounds* approach by utilizing large-span phonotactic features for language identification, by treating each syllabic sound as a term and each utterance or set of utterances belonging to a dialect as a document. Singular Value Decomposition (SVD) is used to compute a set of basis vectors spanning the *global phonotactic* space that captures large span phonotactic information.

First, the M by N term-document matrix T, for M distinct sounds occurring in N dialects, is constructed from some training data. Each training utterance is decoded by all three dialect decoders. Note that the decodings may contain any of 3885 possible distinct tonal syllables from three disjoint sets: 1391 from Mandarin, 1532 for Cantonese, and 962 for Shanghai Dialect. The elements of T are given by

$$T_{i,j} = \frac{\mathrm{tf}(t_i, d_j)}{\sqrt{\sum_{k=1}^{N} \mathrm{tf}(t_i, d_k)^2}} \cdot idf'(t_i), \qquad (3)$$

where the term frequency $\mathrm{tf}(t_i, d_j)$ is the number of occurrences of sound t_i in the decodings of all training utterances belonging to dialect d_j. A modified version of inverse document frequency is $idf'(t_i) = log_2 \frac{N}{\sum_{k=1}^{N} n_k(t_i)}$, where $n_k(t_i)$ is the proportion of training utterances from dialect d_k that when decoded contain the sound t_i. Each element $T_{i,j}$ measures how much a particular sound t_i contributes towards the identification of a dialect d_j, while the column vectors of T represent each dialect as a collection of syllabic sounds. Multiplying each row of T by idf' emphasizes the contribution of less commonly decoded sounds which, intuitively, should be more discriminative when identifying the language. Rearranging the SVD of $T = U \cdot S \cdot V^T$ yields $R \cdot T = S^{-1} \cdot U^T \cdot T = V^T$. Since V^T is orthonormal, the linear operator R has the effect of projecting the column vectors T_j onto the orthogonal basis vectors V_j^T, such that V_j^Ts span the *global phonotactic* space and are as dissimilar as possible with respect to the cosine similarity measure.

To classify a speech utterance X, we compute the vector x of length M that represents its *bag-of-sounds*, and whose elements x_i are the number of times that sound t_i occurs in the three decodings of X. The cosine similarity measure $cos(R \cdot x, V_j^T)$, indicative of the similarity between X and dialect d_j, is computed for each dialect. Information fusion is achieved by combining the acoustic, cross-entropy and LSA-based cosine similarity scores from each dialect through a 3-layer multi-layer perceptron with 9 neurons in the input layer, and a 4 neurons in the output layer (for each dialect and OOL). The output neuron with the greatest activation represents the hypothesized language.

3. EXPERIMENTS

We trained three acoustic models using over 50 hours of speech data per dialect. As Chinese dialects share a common written script, language models for each dialect are trained from the same Chinese newspaper text corpus containing 30 million sentences. Using pronunciation dictionaries for each dialect, we converted each word in the text

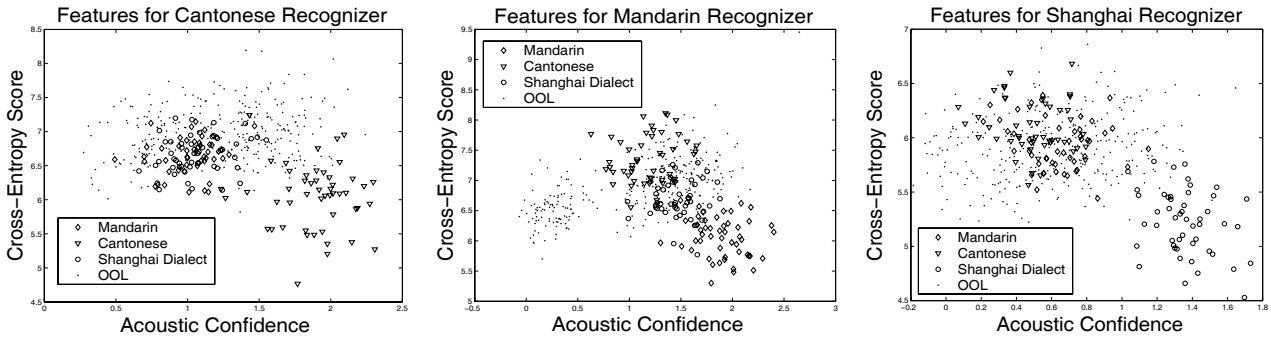


Fig. 3. Cross Examination of Local Acoustic-Phonotactic Scores for Three Parallel Syllable Decoders.

corpus to the standard romanized spelling of its dialect pronunciation, or native *Pinyin*. The converted corpus is used to train three back-off trigram language models, which are then used for cross-entropy evaluation. A similar process is used to train the language model that assists decoding.

We assembled a test speech database independent from the training database, with four categories: 3 Chinese dialects and an *out-of-language* (OOL) category with equal contributions from 6 other languages (English, Minnan Dialect, Korean, Japanese, Spanish and German). Reference transcripts of the test set include both newspaper and conversational style text. Short utterances from the same language were randomly concatenated to generate longer sessions with durations of 5, 10, 15 and 20 seconds. This yielded 1500 speech sessions per category for each time duration. Twenty percent of these speech sessions were withheld for training LSA term-document matrices and neural network classifiers, and the remaining sessions were used for evaluation. The training approach with LSA is rather robust; comparable performance can still be obtained when only 1% of the data is used for training.

3.1. Feature Score Analysis

We studied the ability of each type of feature score to contribute towards the identification task. Fig. 3 plots the separation of acoustic and cross-entropy scores from different language decoders for Chinese dialects. The plots illustrate that acoustic and linguistic scores vary independently, and thus provide complementary and orthogonal information. Each dialect decoder easily distinguishes its own dialect from other languages through a combination of low cross-entropy and high acoustic confidence.

Syllable decoding is not perfect, and in the presence of noise or speaker variation artifacts, increased decoding errors (reflected by higher word error rate) reduce the discriminative ability of cross-entropy scoring. This effect is illustrated in Fig. 4, which plots the classification error rate of Bayesian classifiers that use only cross-entropy scores. The dotted lines estimate a bound for the minimum classification error that can be achieved with the test data, by using cross-entropy scores evaluated over their reference transcripts. One way to alleviate this effect is to merge tonal syllables into toneless syllable classes for cross-entropy scoring. There are 1391, 1532 and 962 tonal syllables and 404, 612 and 562 toneless syllable classes in Mandarin, Cantonese and the Shanghai Dialect respectively. Since confusable tonal syllables now map to the same toneless class, discrimination with cross-entropy scores is improved. This improvement is shown in Fig. 5 for 4-way classifiers that use only the cross-entropy (LM) score.

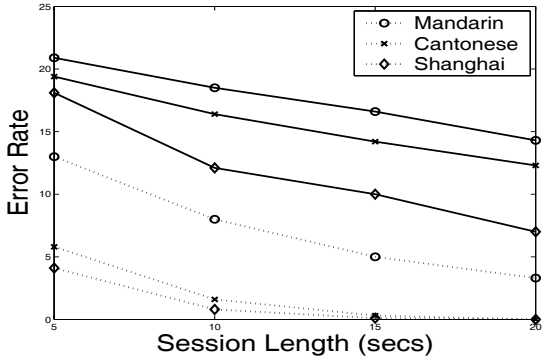


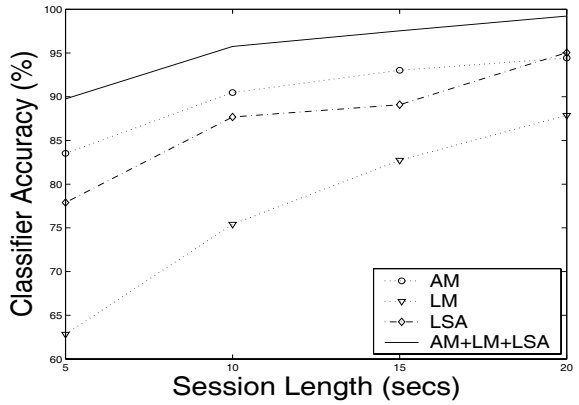
Fig. 4. Performance Bound on Cross-Entropy Scores.

LM only	# syl.	5s	10s	15s	20s
Tonal Syl.	3885	55.5	68.1	76.3	80.8
Toneless Syl.	1578	62.2	75.2	82.4	87.7

Fig. 5. Performance Improvement with Toneless Syllables.

3.2. Classification Experiments

We compared the performance of 4-way classifiers using any combination of acoustic (AM), cross-entropy (LM) and latent semantic (LSA) scores. In this experiment, cross-entropy and latent semantic feature scores were computed with toneless syllable sequences. Fig. 6 shows the average



Method	5s	10s	15s	20s
AM	83.52	90.48	93.02	94.44
LM	62.83	75.42	82.73	87.88
LSA	77.90	87.69	89.08	95.04
NB-KNN	73.31	85.11	92.18	95.40
AM+LM	86.85	93.17	96.15	98.17
AM+LSA	88.96	94.48	96.44	98.29
LM+LSA	82.35	91.71	93.25	97.58
AM+LM+LSA	89.75	95.75	97.54	99.23

Fig. 6. Classifier Performance for Different Feature Types.

accuracy for speech sessions of 5 to 20s durations.

A Naïve Bayes classifier [9], which assumes that each occurrence of a sound in the decoding is independent of other sounds, computes the probability of utterance X (decoded as K distinct sounds $\{x_1...x_K\}$) belonging to language d as $P(d|X) = \prod_{i=1}^{K} P(x_i|d)P(d)/\prod_{i=1}^{K} P(x_i)$. A hybrid Naïve Bayes classifier that uses K-Nearest Neighbours as a fallback (NB-KNN), produced similar results as LSA when trained and tested on the same *bag-of-sounds* features, thus verifying the LSA approach. The discriminative power of LSA is also testified by the fact that, among the two-feature fusion tests, AM+LSA provides the best performance. Furthermore, the combination of any two type of features outperform single feature type systems, and combining all three feature types give the best performance of all, demonstrating that local and global features capture complementary and orthogonal information with good discriminative ability for language identification.

4. CONCLUSION

We have presented an approach to Chinese dialect identification by fusing scores that represent local and global constraints of sound systems that can be extended to the identification of more diverse languages. It is shown that local and global features are complementary in providing language specific evidence, and they are appealing in their own ways. Furthermore, global features derived from *bag-of-sounds* separate the three dialects well in both LSA and Naïve Bayes classifier experiments. The accuracy of this approach can be further improved by merging confusable tonal syllables into toneless classes.

Looking forward, it will be interesting to compare LSA with different high-dimension vector classifiers such as support vector machines. It will also be worthwhile to use a unified syllable decoder that has syllable sounds from all three dialects instead of parallel decoding.

5. ACKNOWLEDGEMENT

We like to thank our colleagues Mr Zhang Jie for providing NB-KNN software, as well as Ms Chen Yu for comments.

6. REFERENCES

[1] T. P. Gleason and M. A. Zissman, "Composite background models and score standardization for language identification systems," in *proc. ICASSP*, vol. 1, May 2001.

[2] A. K. V. S. Jayram, V. Ramasubramanium, and T. V. Sreenivas, "Language identification using parallel subword recognition," in *proc. ICASSP*, vol. 1, May 2003.

[3] S. Parandekar and K. Kirchhoff, "Multi-stream language identification using data driven dependency selection," in *proc. ICASSP*, May 2003.

[4] J. Farinas, F. cois Pellegrino, J.-L. Rouas, and R. Andre-Obrecht, "Merging segmental and rhythmic features for automatic language identification," in *proc. ICASSP*, vol. 1, May 2002.

[5] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," in *proc. ICASSP 96*, Atlanta , USA, April 1996.

[6] T. Schultz, I. Rogina, and A. Waibel, "Experiments with LVCSR based language identification," in *proc. ICASSP 96*, Atlanta , USA, April 1996.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, 1990.

[8] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361–388, 1999.

[9] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *22nd Annual International SIGIR*, Berkley, August 1999, pp. 42–49.