ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

LANGUAGE IDENTIFICATION THROUGH LARGE VOCABULARY CONTINOUS SPEECH RECOGNITION

Boon Pang LIM , Haizhou LI and Yu CHEN

Institute for Infocomm Research
Speech and Dialogue Processing Lab
21 Heng Mui Keng Terrace
Singapore 119613

ABSTRACT

In recent years, automatic language identification has become an increasingly important component in practical spoken language systems, and much attention has been devoted to various competing approaches. In this paper, we are concerned with the automatic identification of languages that may be highly similar in nature, such as the various dialects of Chinese. Our approach differs from many recent successful systems by exploiting a fusion of feature scores readily available from a large vocabulary speech recognition system. We show that such features are able to distinguish among the similar sounding dialects of Chinese, and experiments on a nine language corpus show promising performance on a three way identification task.

1. INTRODUCTION

Automatic Language Identification has gathered a lot of attention in recent years, and many systems have been tested and built. Among the more successful are those based on phone recognition followed by language modelling (PRLM) [1]. Most recently proposed approaches either follow a similar technique of fusing acoustic and phonotactic (language model) information [2] use larger subword units beyond phonemes [3], or integrate other features of spoken language such as rhythmic stress [4] and articulatory features [5].

Meanwhile, language identification (LID) systems which perform identification as a by-product of parallel large vocabulary continuous speech recognition (LVCSR) have been constructed, but to date there are few comprehensive studies involving many languages, and none to our knowledge have based a study on a set of dialectal languages. In [6], good LID performance was obtained using a large vocabulary recognizer, and showed that it is feasible to train an accurate system from unrelated speech corpora. A medium-sized vocabulary recognizer using an LVCSR approach was also tested in [7] on the OGI MLTS database, but did not yield better performance compared to PRLM approaches [1]. This runs counter-intuitive to a comparative study of LVCSR-based identification systems in [8]. This work compared large vocabulary systems employing various degrees of phonetic and lingustic knowledge, and found strong indications that applying higher level sources of knowledge improved upon techniques that merely incorporate acoustical and phonotactical information.

In this paper, we revisit the LVCSR approach, and study the ability of various feature scores produced by LVCSR to distinguish among similar sounding languages. We propose the use of acoustic likelihood ratios (acoustic confidence) in the language identification task. A comparative study of feature scores over a 9-language corpus, containing 4 related Chinese dialects, demonstrate good performance for this approach for similar sounding languages.

2. LANGUAGE IDENTIFICATION OVERVIEW

We constructed a three-way language identification system to distinguish between English, Mandarin and other languages. Two acoustic-linguistic models for English and Mandarin were trained on its respective speech data and run in parallel on the same decoder engine (Abacus). The engine provided acoustically-based confidence as well as language model scores, which were then used as inputs to a neural network classifier. Fig.1 shows the overall architecture of our system.

2.1. Language Identification Corpus

We assembled a speech corpus comprised of the following nine languages: 4 Chinese dialects (Mandarin, Minnan dialect, Shanghai dialect and Cantonese), and 5 foreign languages (English, Korean, Japanese, Spanish and German). Utterances for English and the Asian languages were obtained from our collected training and test database, whilst the Spanish and German sessions were extracted from the Linguistic Data Consortium's CallHome corpora. Only utterances with more than three words were used, and these

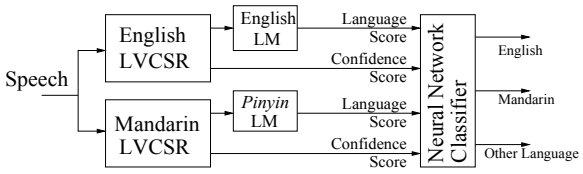


Fig. 1. Language identification architecture

were checked to ensure that no foreign words were present. These utterances were then randomly concatenated to constitute longer fixed duration sessions of 5, 10, 15, 20, 25, and 30 seconds. This produced about 1000 sessions for each language and time duration, yielding a total of 6000 sessions per language for a total of 262.5 hours of speech. One fifth of the corpus was reserved for training the neural network, and the remainder for open tests.

2.2. Acoustical Confidence Measure

Abacus is a frame-synchronous HMM-based large vocabulary continuous speech recognizer, capable of employing class-based n-gram language models. A conceptual overview of its architecture is provided in Fig. 2. The signal processing front-end uses liftered mel-scale cepstral coefficients to extract feature vectors in accordance with the ETSI standard.

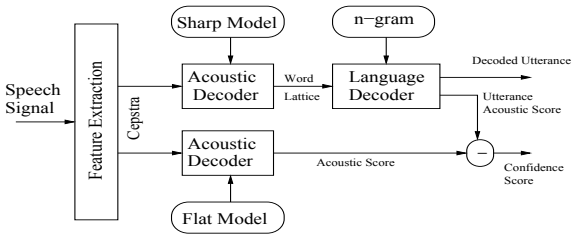


Fig. 2. Speech recognizer architecture

The cepstral features are fed into two separate acoustic decoders employing acoustic models of different granularity; a sharp acoustic model employing three-state HMMs for a set of tied-state context-dependant triphones, and a corresponding flat acoustic model employing broad classes of context independant phones, in which similar properties of sounds (such as nasals, fricatives or plosives) are grouped under the same class. Also, a phonetic lexicon is integrated into the sharp model to yield a word lattice at the acoustic decoder output, which is then scored by a language model decoder. The likelihood ratio for an utterance with a set of observation vectors O, as matched against the sharp and flat models is given by

$$r(O) = \frac{P(O|\lambda)}{P(O|\bar{\lambda})} = \frac{P(o_1, ..., o_K|\lambda_1, ..., \lambda_M)}{P(o_1, ..., o_K|\bar{\lambda}_1, ...\bar{\lambda}_M)}. \quad (1)$$

Here, $O = \{o_1, ...o_K\}$ is the set of the feature vectors observed for an incoming utterance with K non-silent speech frames. $\lambda = \{\lambda_1, ..\lambda_M\}$ corresponds to the best matching phoneme sequence from the integrated acoustic-linguistic decoder after language modelling, and $\bar{\lambda} = \{\bar{\lambda}_1, ...\bar{\lambda}_M\}$ corresponds to the best matching phoneme sequence from the flat acoustic model decoder.

Taking the log and normalizing across K non-silent frames yields the acoustic confidence measure

$$C = \frac{\sum_{k=1}^{K} \left[\log P(o_k|\lambda) - \log P(o_k|\bar{\lambda})\right]}{K}. \quad (2)$$

The acoustic confidence measure reflects how likely that a session is in a given language, by using the flat model as a reference point; a larger score indicates a better acoustical match. This approach helps remove channel effects, such as noise in the recording environment, that occur within the original training corpus. It also permits us to more effectively use disjoint corpora of speech in training without prior knowledge of the acoustic environment of the test sessions. Our studies on the distribution of feature scores from different languages also indicate that confidence scores are better indicators of the correct language as compared with just acoustic log probability scores.

2.3. Language Model Scoring

Given a sentence W with the sequence of N words $w_1...w_N$, we can compute its probability within a n-gram language model as

$$p(W) = \prod_{i=1}^{N} p(w_i|w_1...w_{i-1}),$$

where p's are n-gram probabilities from the language model. Cross-entropy is evaluated using

$$H_p(W) = -\frac{1}{N} \log p(W). \quad (3)$$

Cross-entropy tells how well a sentence matches a given language model where a smaller value indicates its better fit. These cross-entropies are taken to be the language feature scores for our classifier.

The best hypothesis output from the recognizer is evaluated against two backoff language models trained using the SRILM Toolkit [9] through Eqn. 3. A word-bigram language model for English was trained from a corpus of Wall Street Journal transcriptions from the years '87 through '94 that contains 5 million sentences. A syllable-based or *Pinyin* trigram model was used for Mandarin. This language model was trained from a corpus of standard newspaper transcriptions containing 30 million sentences. To convert Chinese characters to its *Pinyin* equivalent, the text is segmented into a series of words, where a word is a sequence of up to 8

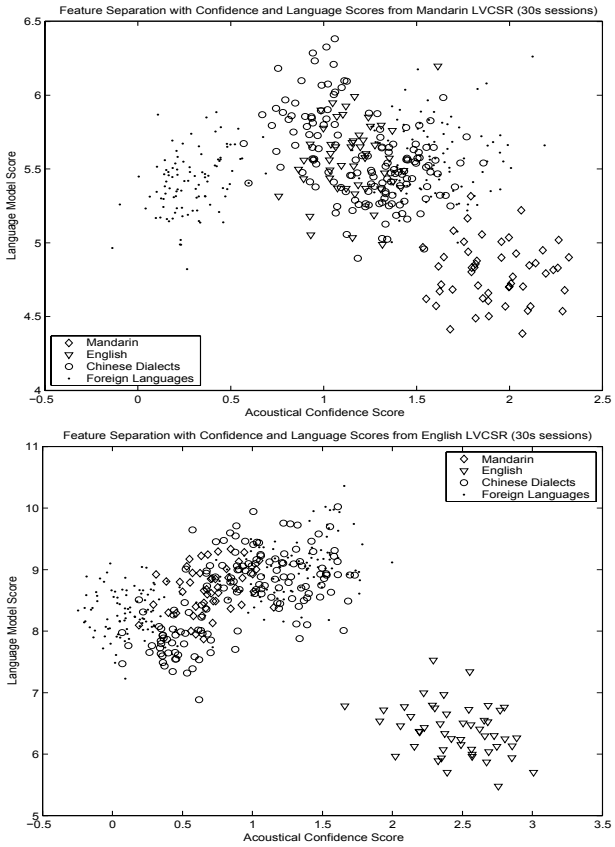


Fig. 3. Feature separation for each language (30s)

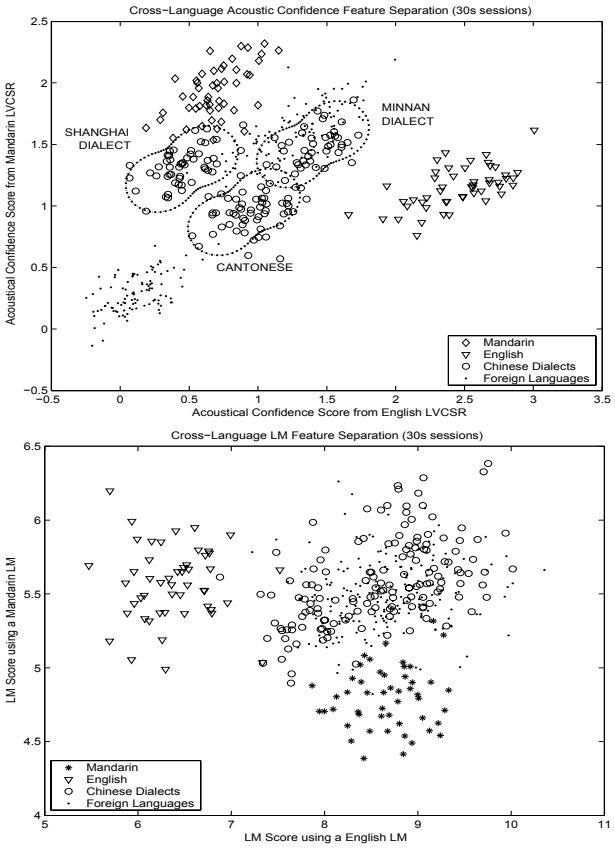


Fig. 4. Cross language feature separation (30s)

Chinese characters with a predetermined meaning and pronunciation. This method helps to reduce errors stemming from the pronunciation overloading that frequently occurs with Chinese characters. In cases where a single character cannot be grouped into a word, it is converted to its most common pronunciation.

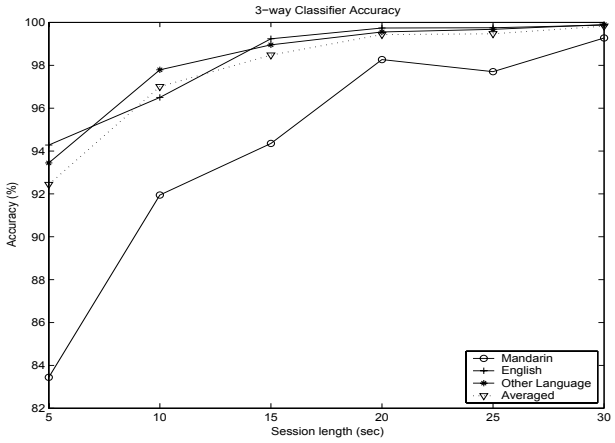
Finally, the acoustic and language feature scores are fed into a 3-layer perceptron with a 4-node input layer (for each feature), and a 3-node output layer (for English, Mandarin or others). The language whose neuron has the greatest activation is chosen as the final result.

3. EXPERIMENTS

We first studied the ability of the various features to distinguish between different sets of languages for sessions 30s in length. Fig. 3 shows a scatter plot of confidence and language scores per session. The plots show that confidence scores range from -0.5 to 3, with a higher score indicating a better match, while language scores are generally positive with a lower value indicating a better match. We observe that there is good feature separation between English, Mandarin and other languages. In addition, feature scores from any LVCSR system are sufficient to distinguish that same language from all other languages, and both LM and Confidence scores are equally good as classification measures by themselves.

Fig. 4 presents the same data of Fig. 3 from a different perspective. It plots scores for the same type (acoustic or language) of features across the LVCSRs. In general, the feature scores for sessions in the same languages as the LVCSR will be better, as illustrated by the two well separated clusters corresponding to English and Mandarin. In the case of language scores, there is no clear separation for each language since only English or Mandarin will match well with its own language model. Confidence scores, however, are not only able to distinguish between English, Mandarin and others, but can also partially distinguish between the Chinese dialects. The three clusters attributed to the Chinese dialects appear well separated enough to permit good identification, although we have yet to validate this through a classification experiment.

Next, we evaluate the performance of our system in the open test using a 3-way neural network classifier. Fig. 5 plots the accuracy of our classifier, as the length of the sessions are varied from 5 to 30s. On average, the system



length	5s	10s	15s	20s	25s	30s
Av Acc %	93.1	97.1	98.3	99.4	99.4	99.8

Fig. 5. Three-way classifier performance.

achieves very good accuracy and is comparable to existing language identification systems.

Fig. 6 shows a breakdown of the errors that occur for each source language in our corpus for sessions 15s in length. We observe that Chinese dialects tend to be misclassified among themselves, which is consistent with the general observation that dialects of Chinese tend to sound very similar to each other. One interesting observation is that Korean and Japanese tended to be misidentified as Mandarin, which suggests that these languages are sound more similar to Mandarin than we originally suspected. For the worst performing language, the Shanghai dialect, we suspect that there is a correlation between the closeness of its cluster position to Mandarin in the cross-language acoustic feature space and high misclassification rate. Nevertheless, a consistently high accuracy is achieved across the board on all languages, even for similar sounding Chinese dialects.

4. CONCLUSION

In this paper, we propose to use acoustic confidence and language model scores from an LVCSR as inputs of a multi-classifier for language identification. Experiments show that these scores provide two effective dimensions for identifying Chinese dialects, and that this approach performs well in a nine language test set containing four Chinese dialects.

In Chinese spoken language, there are more than a dozen popular dialects that share similar writing systems, phoneme sets and phonic rules. Driven by commercial needs, LVCSR systems are becoming available in more and more languages, making an approach to language identification based on LVCSRs more viable. Given the promising performance of this approach, we suggest that the LVCSR-based approach is worthwhile pursuing further, especially for the identification of Chinese dialects.

Language	Man	Eng	Others	Acc. %
Mandarin	752	0	45	94.35
English	0	772	6	99.23
Minnan Dialect	3	0	796	99.62
Shanghai Dialect	41	0	772	94.96
Cantonese	0	0	802	100.00
Spanish	0	0	787	100.00
German	0	0	822	100.00
Japanese	4	0	797	99.50
Korean	10	1	781	98.61

Fig. 6. Confusion across various languages (15s).

5. REFERENCES

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

[2] J. Gutierrez, J.-L. Rouas, and R. Andre-Obrecht, "Fusing language identification systems using performance confidence indexes," in *procs. ICASSP*, vol. 1, May 2004.

[3] A. K. V. S. Jayram, V. Ramasubramanium, and T. V. Sreenivas, "Language identification using parallel subword recognition," in *procs. ICASSP*, vol. 1, 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 *procs. ICASSP*, vol. 1, May 2002.

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

[6] 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.

[7] J. L. Hieronymus and S. Kadambe, "Spoken language identification using large vocabulary speech recognition," in *proc. ICSLP 96*, Philadelphia, USA, 1996.

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

[9] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *proc. ICSLP*, 2002, pp. 901–904.