LANGUAGE BOUNDARY DETECTION AND INDENTIFICATION OF MIXED-LANGUAGE SPEECH BASED ON MAP ESTIMATION

Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsin Hsieh and Chung-Hsien Wu

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{shiacj, chiuyh, ngsnail, chwu}@csie.ncku.edu.tw

ABSTRACT

This paper proposes a Maximum *a Posteriori* (MAP) based approach to jointly segment and identify an utterance with mixed languages. A statistical framework for language boundary detection and language identification is proposed. First, the MAP estimation is used to determine the boundary number and positions. Further, an LSA-based GMM and a VQ-based bi-gram language model are proposed to characterize a language and used for language identification. Finally, a likelihood ratio test approach is used to determine the optimal number of language boundaries. Experimental results show that the proposed approach exhibits encouraging potential in mixed-language segmentation and identification.

1. INTRODUCTION

In recent years, multi-lingual spoken language processing becomes increasingly necessary and provides the applications of human machine interaction in globalizing economy, communication and information exchangeability. Most of today's computerized spoken dialog systems identify a language using dedicated modules. During the past years, a wide range of approaches to automatic language identification (LID) was applied, such as language-dependent Gaussian mixture models (GMM), GMM tokenization and single/parallel phone recognizer followed by language modeling (single/parallel PRLM) [1][2]. Of these methods, GMM was herein the most efficient in terms of time complexity but yielded the lowest language ID rate. In contrast, parallel PRLM yielded the highest language ID rate but was the most complex system for identifying languages. The use of linguistic property contributes to distinguish languages of each other. Recent work emphasizes on integrating high level linguistic structure and extracting robust acoustic features to improve the LID rate.

However, these approaches focused on identifying an utterance with a single language. Identifying mixed languages in an utterance challenges the present LID systems. The mixed language ID applications arise very commonly in Asia. In Taiwan, three languages —Mandarin, Taiwanese and English— are frequently mixed and spoken in daily conversations. For example, the sentence {最近的"Starbucks"在哪裡?} (Where is the nearest "Starbucks"?) is spoken in Mandarin-English. These mixed-language sentences are generally used in applications like car navigation systems and information service dialog systems. In this task, segmenting such an utterance into several language segments is crucial to the development of a LID system and the integration of speech recognizers. A language segment with a short length of utterance cannot obtain a promising performance of LID rate due to insufficient information in the segment. The present LID systems were hard to handle this kind of utterances. Moreover, detecting the language boundary is also the key issue for this task. Recently, many methods of audio segmentation have been developed, such as Akaike's information criterion (AIC), the Bayesian information criterion (BIC), the delta-BIC [3] and the minimum description length (MDL). A detailed comparison was addressed in [4]. In these approaches, boundary segmentation is performed by detecting acoustic changes. But, the difficulty is how to select appropriate model parameters and the penalty weights. Conventionally, these values are determined empirically and therefore limit the segmentation performance.

In this paper, we propose a MAP-based framework for boundary detection and language identification of mixed-language utterances. A statistical approach, including a MAP-based probability model, an LSA-based GMM and a VQ-based bi-gram language model, is proposed. In this framework, an utterance with mixed languages is segmented into several language segments. To perform the language boundary detection, the probabilities of occurrences of the boundary positions, occurrences of hypothesized languages, the language probability of each segment and the number of boundaries are optimized using a likelihood ratio test approach. To identify a language in each segment, a VQ-based bi-gram model is adopted. Each codeword sequence is further converted into a codeword occurrence vector and all the vectors are used to form a segment-codeword vector matrix. A latent semantic analysis is then used to transform this matrix into a reduced space with a small set of discriminative features. Then an LSA-based GMM approach is proposed to model this reduced matrix. Finally, a dynamic programming algorithm is adopted to obtain the optimal hypothesized language sequence.

2. THE FRAMEWORK

Consider a speech utterance S with N_S feature vectors (frames) and q language boundaries, and the utterance is therefore segmented into $q+1$ speech segments to form a segment sequence $S=(S_1, S_2, \cdots, S_{q+1})$. Suppose the positions of the language boundaries are denoted as $R=(r_1, r_2, \cdots, r_q)$, $1<r_1<r_2<\cdots<r_q<N_S$, and the corresponding language sequence is $\tilde{L}=(L_{S_1}, L_{S_2}, \cdots, L_{S_{q+1}})$, $\forall L_{S_i} \in L=\{L_1, \cdots, L_K\}$, which L_{S_i} is a language associated with speech segment S_i.

The proposed MAP-based framework of detecting language boundary and identifying languages is given as follows.

$$\begin{aligned}\max P(\tilde{L},R,q\mid S) &= \max\frac{P(S,\tilde{L},R\mid q)P(q)}{P(S)}\\ &\approx \max P(S,\tilde{L},R\mid q)\\ &= \max P(S\mid\tilde{L},R,q)P(\tilde{L}\mid R,q)P(R\mid q)\end{aligned} \quad (1)$$

where $P(R\mid q)$ is the conditional probability of boundary positions R given the number of language boundary q; $P(\tilde{L}\mid R,q)$ is the conditional probability of language sequence $\tilde{L}$ respective to the boundary number and positions. And $P(S\mid\tilde{L},R,q)$ represents the conditional probability of a segment sequence with boundary number q and positions R respective to the language sequence $\tilde{L}$.

2.1 Probability Estimation of Boundary Position

To estimation the conditional probability $P(R\mid q)$, the occurrence probability of the boundary positions is assumed to be equal. There is no permutation of boundary positions by the constraint, $1<r_1<r_2<\cdots<r_q<N_S$. And a uniform distribution is considered and given as follows.

$$P(R\mid q)=1\Big/\binom{N_S-2}{q} \quad (2)$$

where N_S-2 possible positions are taken for boundary r_i.

2.2 Probability Estimation of Language Sequence

The conditional probability $P(\tilde{L}\mid R,q)$ represents the probability of language sequence $\tilde{L}$ associated with the number of language boundary and the boundary positions. Furthermore, the relation between boundary positions R and language sequence $\tilde{L}$ is independent, then a multi-nominal distribution estimation of language sequence as $\tilde{L}$ is follows.

$$\begin{aligned}&P(\tilde{L}\mid R,q)=P(L_{S_1},\cdots,L_{S_{q+1}}\mid q)\\ &=\binom{N_S}{n_{S_1},n_{S_2},\cdots,n_{S_{q+1}}}\left(P(L_{S_1})\right)^{n_{S_1}}\left(P(L_{S_2})\right)^{n_{S_2}}\cdots\left(P(L_{S_{q+1}})\right)^{n_{S_{q+1}}}\end{aligned} \quad (3)$$

where $n_{S_i}=r_i-r_{i-1}$ in the length of segment S_i; $P(L_{S_i})$ is the probability that the language of speech segment S_i is $L_{S_i}\in L=\{L_1,\cdots,L_K\}$ and is estimated from the training corpus to represent the *a priori* probability of language L_{S_i} for a segment.

$$P(L_k)=\begin{cases}\dfrac{\#\_of\_segments\_belong\_to\_L_k}{total\_\#\_of\_segments} & ,k=1,\cdots,K\\ 0 & ,\text{otherwise}\end{cases} \quad (4)$$

2.3 Segment-based Language Identification

Given the boundary number q and boundary positions $R=(r_1,r_2,\cdots,r_q)$, an utterance is divided into a segment sequence $S=(S_1,S_2,\cdots,S_{q+1})$. Assume the probabilities of each speech segment S_i associated with L_{S_i} are independent. The probability of utterance S given q, $R=(r_1,r_2,\cdots,r_q)$ and the corresponding language sequence $\tilde{L}=(L_{S_1},L_{S_2},\cdots,L_{S_{q+1}})$ is equal to the product of all probabilities of segment S_i given the corresponding language L_{S_i}, described as follows.

$$P(S\mid\tilde{L},R,q)=\prod_{i=1}^{q+1}P(S_i\mid L_{S_i}) \quad (5)$$

Consider a speech segment S_i with length n_{S_i} represented as a vector sequence $X_{S_i}=(x_{S_i}^1,x_{S_i}^2,\cdots,x_{S_i}^{n_{S_i}})$. The feature vector sequence is then vector quantized using a language-dependent codebook derived from K-means algorithm [5], yielding a codeword sequence $C_{S_i}=(c_{S_i}^1,c_{S_i}^2,\cdots,c_{S_i}^{n_{S_i}})$. The codeword sequence of each segment is further represented by a codeword occurrence vector $Y_{S_i}=(y_{1,S_i},y_{2,S_i},\cdots,y_{U,S_i})$, where U is the size of the final codebook, which is the union of all language-dependent codebooks, and y_{u,S_i} represents the number of occurrences of codeword u in segment S_i, given $u=1\cdots U$. Assuming that the *a priori* probabilities associated with language $P(L_k)$ and speech segment $P(S_i)$ are equally-likely, each speech segment can be represented using the above three components.

$$\begin{aligned}&P(S\mid\tilde{L},R,q)=\prod_{i=1}^{q+1}P(S_i\mid L_{S_i})=\prod_{i=1}^{q+1}P(Y_{S_i},C_{S_i},X_{S_i}\mid L_{S_i})\\ &=\prod_{i=1}^{q+1}P(Y_{S_i}\mid C_{S_i},X_{S_i},L_{S_i})P(C_{S_i}\mid X_{S_i},L_{S_i})P(X_{S_i}\mid L_{S_i})\end{aligned} \quad (6)$$

where $P(X_{S_i}\mid L_{S_i})$ represents the conditional probability associated with the feature vector sequence X_{S_i} with respect to language L_{S_i} and reflect the acoustic property of language L_{S_i} and is modeled by a Gaussian Mixture Model in acoustic space; $P(C_{S_i}\mid X_{S_i},L_{S_i})$ represents the conditional probability associated with the codeword sequence C_{S_i} given the language L_{S_i} and the feature vector sequence X_{S_i} specifies the syntactical property of a sequence of phones in language L_{S_i}. An n-gram language model is commonly used to model the syntactical characteristic of a language. In this approach, a VQ-based bi-gram model is adopted as follows.

$$P(C_{S_i} \mid X_{S_i}, L_{S_i}) = P(c_{S_i}^1 c_{S_i}^2, , \cdots, c_{S_i}^{n_{S_i}} \mid X_{S_i}, L_{S_i})$$
$$= \prod_{j=1}^{n_{S_i}} P(c_{S_i}^j \mid c_{S_i}^1, \cdots, c_{S_i}^{j-1}, X_{S_i}, L_{S_i}) \quad \{\text{n - gram}\} \tag{7}$$
$$\approx \prod_{j=1}^{n_{S_i}} P(c_{S_i}^j \mid c_{S_i}^{j-1}, X_{S_i}, L_{S_i}) \quad \{\text{bi - gram}\}$$

where $P(Y_{S_i} \mid C_{S_i}, X_{S_i}, L_{S_i})$ represents the conditional probability associated with L_{S_i}, X_{S_i} and C_{S_i} in codeword occurrence vector space, specifying the lexical property of language L_k. The GMM is adopted to model the distributions of codeword occurrence vectors in a language.

2.3.1 GMM with Latent Semantic Analysis

The LSA formalism is adopted to reduce the number of dimensions of the space by constructing a codeword-by-segment matrix, whose entry $y_{u,S_i}^{L_k}$ appropriately reflects the number of occurrences of codeword u, which appears in segment S_i in language L_k. The matrix is called the language information matrix (LIM), $M_{U\times N}^{LIM}$. Latent semantic analysis is the application of a particular mathematical technique, called Singular Value Decomposition (SVD) [6]. The SVD projection is computed by decomposing the codeword-by-segment matrix $M_{U\times N}^{LIM}$ into the product of three matrices, $T_{U\times n}$, $S_{n\times n}$ and $D_{N\times n}$:

$$M_{U\times N}^{LIM} = T_{U\times n} S_{n\times n} (D_{N\times n})^T \quad \text{with } n = \min(U, N) \tag{8}$$

where $T_{U\times n}$ and $D_{N\times n}$ have orthonormal columns. LSA mapping the original space onto a discriminative space with the first axis is the direction with largest variation, and the second with second largest variation, and so on. Choosing $\rho < n$, a truncated SVD with matrices $T_{U\times\rho}$, $S_{\rho\times\rho}$ and $D_{N\times\rho}$ is derived and represented as follows.

$$\hat{M}_{U\times N}^{LIM} = T_{U\times\rho} S_{\rho\times\rho} (D_{N\times\rho})^T \tag{9}$$

Let $L_{S_i} = L_k$, the conditional probability $P(Y_{S_i} \mid C_{S_i}, X_{S_i}, L_{S_i})$ is then measured as follows.

$$P(Y_{S_i} \mid C_{S_i}, X_{S_i}, L_{S_i} = L_k) \approx P(Y_{S_i} \mid \Lambda_k) \approx P(T_{Q\times\rho}^T Y_{S_i} \mid \Lambda_k)$$
$$= P(\overline{Y}_{S_i} \mid \overline{\Lambda}_k) = \sum_{m=1}^{M_k} \overline{w}_{k,m} N(\overline{Y}_{S_i}, \overline{\mu}_{k,m}, \overline{\Sigma}_{k,m}) \tag{10}$$
$$= \sum_{m=1}^{M_k} \overline{w}_{k,m} \frac{\left|\overline{\Sigma}_{k,m}\right|^{-1/2}}{(2\pi)^{\rho}} \exp\left(-\frac{1}{2}(\overline{Y}_{S_i} - \overline{\mu}_{k,m})^T \overline{\Sigma}_{k,m}^{-1} (\overline{Y}_{S_i} - \overline{\mu}_{k,m})\right)$$

where $\overline{Y}_{S_i}$ is the code word occurrence vector of segment S_i in the low-dimensional space; $\overline{\Lambda}_k = \{\overline{w}_{k,m}, \overline{\mu}_{k,m}, \overline{\Sigma}_{k,m}, M_k\}$, in which $\{\overline{\mu}_{k,m}, \overline{\Sigma}_{k,m}\}$ represents the mean vector and covariance matrix of the m-th mixture of the GMM of language L_k and $\overline{w}_{k,m}$ represents the weight of the m-th mixture. M_k is the number of mixture in the GMM of language L_k. These model parameters are estimated using the expectation maximization (EM) algorithm [7].

2.4 Hypothesized Language Sequence and Boundary

We have derived an approach for mixed-language boundary detection and language ID. Here a maximum likelihood ratio scheme is adopted for mixed-language boundary detection to determine the best solution of q.

$$\frac{\max_{\tilde{L}_{q+1}} P(S, \tilde{L}_{q+1}, R \mid q+1)}{\max_{\tilde{L}_q} P(S, \tilde{L}_q, R \mid q)} < \eta \tag{11}$$

where η is the likelihood ratio threshold. If the likelihood ratio is less than η, there are q language boundaries in S.

The hypothesized language sequence $\hat{\tilde{L}}$ given boundary number $\hat{q}$ is determined by maximizing the log-likelihood with respect to speech utterance S and is estimated as follows.

$$\hat{\tilde{L}} = \arg\max_{\tilde{L}} \log P(S, \tilde{L}, R \mid \hat{q})$$
$$= \arg\max_{\tilde{L}} \{\log P(R \mid \hat{q}) + \log P(\tilde{L} \mid R, \hat{q}) + \log P(S \mid \tilde{L}, R, \hat{q})\} \tag{12}$$

Here, dynamic programming [8] is adopted to search the best boundary positions $\hat{R}$.

3. EXPERIMENTAL RESULTS

In previous work on language ID task, some well-established corpora with single-language utterances have been used, such as the CallFriend and the OGI-TS corpus [1][2]. NIST reported their recent evaluation on the performance of LID using the Call Friend corpus with 12 languages [9]. However, none contains mixed-language utterances. In this work, we collected a read speech corpus with mixed languages to evaluate the proposed approaches, in which Mandarin-English and Mandarin-Taiwanese are considered. The text sentences for recording mixed-language speech are generated by embedding an English phrase or a Taiwanese phrase into a Chinese carrier sentence.

For model development, 3750 English utterances and 2250 Taiwanese utterances, spoken by 27 male and 14 female speakers, and 1725 Mandarin utterances (from 21 male and 14 female speakers) extracted from the database TCC300 [10] were collected as the training databases. These databases were used to train language-dependent codebooks, VQ-based bi-grams and GMMs. For testing database collection, 11 male and 5 female speakers, different from the speakers who provided the training data, were asked to record 21760 mixed-language utterances,

1360 for each speaker, containing 918 Mandarin-Taiwanese mixed utterances and 442 Mandarin-English mixed utterances.

In order to evaluate the proposed approach, two experiments on language boundary detection rate and LID rate were conducted. In the first experiment, the precision rate, recall rate and harmonic mean F criterion [11] were adopted to evaluate the performance. The codeword size of 64 and the mixture number of 64 for each language were trained and used. Tables 1 & 2 give the precision rates, recall rates and LID rate as a function of different thresholds and penalty weights for the MAP-based approach and the delta-BIC approach. Experimental results show that the proposed approach outperformed the delta-BIC. Figure 1 shows the LID rate as a function of mixture number. The proposed MAP-based boundary detection approach is compared with delta-BIC approach followed by different LID methods. Experimental results show that our proposed approach achieved 76.2% LID rate and outperformed other approaches using traditional GMM and the delta-BIC.

4. CONCLUSION

This work proposed an MAP-based approach to detecting language boundary and identifying mixed languages. In this framework, joint estimation of language boundary and LID exhibits encouraging potential in mixed-language ID task. The LSA-based GMM provides a new approach for using probabilistic distributions to characterize a language. The experimental results show a promising performance on the boundary detection rate and the LID rate.

Table 1. The precision, recall and language ID rates for different threshold values

Threshold value η	*Harmonic Mean Measure*			*LANGUAGE ID RATE(%)*
	Precision Rate	*Recall Rate*	*F*	
1.24	0.58	0.59	0.58	74.5
1.22	0.55	0.62	0.58	74.9
1.20	0.52	0.73	0.61	75.5
1.18	0.50	0.81	0.62	75.8
1.16	0.49	0.85	0.62	75.9
1.14	***0.48***	***0.90***	***0.63***	***76.2***
1.12	0.42	0.93	0.58	76.1
1.10	0.39	0.95	0.55	76.1

Table 2. The language ID rates for different penalty weights in delta-BIC

Penalty Weight	Window Size (seconds)	HARMONIC MEAN MEASURE			LANGUAGE ID RATE (%)
		Precision Rate	*Recall Rate*	*F*	
0.5	0.3	0.32	0.85	0.46	68.5
0.6	0.3	0.35	0.84	0.49	68.4
0.7	0.3	0.39	0.82	0.53	68.3
0.5	0.4	0.34	0.85	0.46	68.6
0.6	0.4	0.38	0.83	0.52	68.4
0.7	***0.4***	***0.41***	***0.82***	***0.55***	***68.5***
0.5	0.5	0.40	0.77	0.53	67.1
0.6	0.5	0.43	0.69	0.53	67.0
0.7	0.5	0.45	0.61	0.52	66.7

5. ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council, ROC, for its financial support of this work, under Contract No. NSC90-2213-E-006-088.

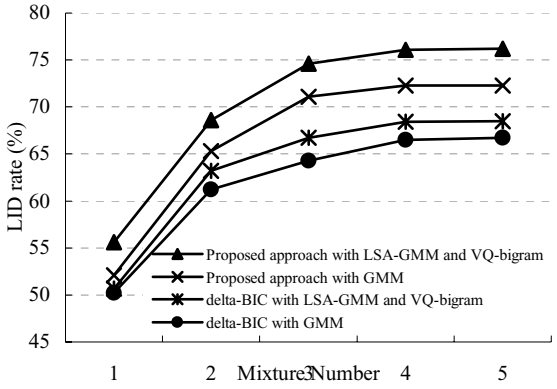

Fig. 1. Comparison of mixed-language ID rates

6. REFERENCES

[1] Marc A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. On Speech and Audio Proc.*, Vol. 4, NO. 1, pp. 31-44, January 1996.

[2] Y. K. Muthusamy, Etienne Barnard, and Ronald A. Cole, "Reviewing Automatic Language Identification," *IEEE Signal Processing Magazine*, Vol. 11, Issue 4, pp. 33-41, Oct 1994.

[3] Alain Tritschler and Ramesh Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," *in Proc. of EUROSPEECH99*, 1999, vol. 2, pp. 679-682.

[4] Mauro Cettolo and arcello Federico, "Model Selection Criteria for Acoustic Segmentation," *in Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, Paris, France, 2000, pp. 221-227.

[5] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[6] C. D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, pp. 554-566, the MIT Press, Cambridge, Massachusetts, 1999

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1-38, 1977.

[8] Steven M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall PTR, 1998.

[9] Alvin F. Martin and Mark A. Przybocki, "NIST 2003 language recognition evaluation," *in Proc. of EUROSPEECH 2003*, pp. 1341-1344, 2003.

[10] Y. J. Chen, "A Study on Conversational Speech Recognition and Verification in Computer Telephony Integration" Ph.D. Thesis, National Cheng Kung University, Tainan, Taiwan, 2000.

[11] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.