Code-Switching Event Detection Based on Delta-BIC Using Phonetic Eigenvoice Models

Wei-Bin Liang, Chung-Hsien Wu, and Chun-Shan Hsu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
{liangnet, chunghsienwu, chunshan7699}@gmail.com

Abstract

This paper presents a new paradigm for code-switching event detection based on delta Bayesian Information Criterion (ΔBIC). First, an automatic speech recognizer (ASR) and an articulatory feature (AF) detector are constructed. The inter-syllable boundaries obtained from the ASR are regarded as the potential code-switching boundaries. To estimate the language likelihood, eigenvoice models (EVMs) are employed to model the relationship between the senones/articulatory attributes and their corresponding eigenvoices constructed from the training data for different languages. The Euclidean distance and the inner product-based direction between the eigenvoice vector of the input sentence and the eigenvoice vector of a senone or an articulatory attribute in the EVMs for different languages are calculated for ΔBIC-based language likelihood estimation. Then, an n syllable Bayesian mask centered at each potential boundary is then employed to output the likelihood of language change for the potential boundary. Finally, the dynamic programming algorithm is employed to search the best language sequence given the inter-syllable boundaries from the ASR. The proposed approach was evaluated on a Chinese-English code-switching speech database and the results show that 71.93% accuracy for code-switching event detection can be obtained.
Index Terms: Code-Switching event Detection, Senone, Articulatory attributes, Eigenvoice, Bayesian mask

1. Introduction

Due to the globalization, code-switching has become a common phenomenon in daily life. Code-switching for different language combinations could be easily found, such as English-Spanish in United States, Cantonese-English in Hong Kong, and Mandarin-English/Taiwanese in Taiwan. Code-switching event detection is becoming indispensable in the application of human-machine communication, such as spoken dialogue systems. There are four code-switching types [1][2][3]: inter-sentential switching, intra-sentential switching, tag switching, and intra-word switching. An example of the second type is illustrated as follows.

Code-Switching:	媽,	我想要買一台	iMac.
English:	Mom,	I want to buy one	iMac.

Generally, English words or clauses are frequently embedded in a spoken Mandarin sentence, in which Mandarin is the primary language, also known as the matrix language, and English is the secondary language, usually referred to as the embedded language [4].

Automatic speech recognition (ASR), which transcribes the input utterance into a sequence of words, is one of the key technology in a spoken dialogue system. In recent years, code-switching has become a challenging issue for developing ASR-based speech applications in a multilingual society. Therefore, language identification (LID) of each speech segment is desirable to help recognize code-switched utterances.

Many research efforts have been made in the LID tasks and the modeling approaches are divided into two broad categories – acoustic-level and linguistic-level approaches. The first category of the LID systems is to train the model for each language using acoustic features extracted from speech signals. In this category, GMM-based approaches [5][6][7], SVM-based approaches [8][9][10], and HMM-based approaches are commonly used. Moreover, MFCCs [11], delta and delta-delta MFCCs [12], and shifted delta cepstra (SDC) [13] are widely used speech features for identification model training. In conventional LID systems, one GMM is trained for each language. The GMM-based LID systems used one single GMM to cover all languages, called the universal background model (UBM). The UBM is then adapted to generate a new GMM for each language. Although GMM is a simple and efficient modeling approach, performances of LID systems are often degraded due to the highly confusing units in the languages for identification. SVM-based classifiers are designed through an optimization process, which is discriminative in nature. In SVM classifier design, the kernel plays a central role but not easily chosen for unknown data distribution. For speech features, MFCC is one of the most commonly used speech features for speech recognition. While static feature vectors like the MFCCs provide a good estimation of the local spectra, they failed to capture the dynamic aspects of human speech which are very important for distinguishing between two different languages. Since the dynamic features such as delta and delta-delta MFCCs are used to model the slope of the speech features of the current speech frame, their abilities are limited to model higher level temporal aspects of speech signals. The SDC has been proposed to model the temporal information in the speech signal across a longer window. SDCs are obtained by concatenating the sampled future delta cepstra with the current feature vector. Generally, the size of the sampling window is value optimized empirically. For the secondary category, the most important differences between each of the different languages are that they have their own word roots and lexicons. The linguistic level features could be the syntactic and phonetic information. Phonetic information can be obtained from the speech data using a phone recognizer followed by an n-gram language model (LM) to estimate the occurrence probability of a particular phone sequence. The mechanism of phone recognition followed by language modeling is called PRLM [14][15][16]. Therefore, individual phone recognizers are trained on different languages to capture dif-

ferent acoustic characteristics. One particular way is to combine these phone recognizers to form a parallel phone recognizer (PPR) which is then followed by a language model (LM) (i.e. PPRLM) [15][14]. Nevertheless, a well-trained multilingual LM is not easy to obtain, especially for the code-switching language. The LM is seriously biased because the embedded language in a sentence is often the minority. Besides, LID systems may be improved by means of appropriate speech segmentation. Given two languages in a spoken sentence, an inter-boundary in a code-switched utterance could have four transition cases. However, classifying a speech segment to either the primary language or the secondary language is simplified as a binary classification problem. Recently, speech segmentation approaches such as ΔBIC [17] and MDL [18] are commonly employed for speech segmentation. A detailed comparison was addressed in [19] and the delta-BIC is a widely used approach. In these approaches, boundary segmentation is performed by detecting the acoustic change. But, the difficulty lies on how to select appropriate model parameters and the penalty weights. Conventionally, these values are determined empirically and therefore limit the segmentation performance.

However, these approaches cannot be directly applied to our task which has confusing acoustic units and seriously biased language models. This work investigates the techniques for code-switching event detection. The big challenge on this task is that the duration of the speech segment of the embedded language is very short. Speech segments with short duration in a code-switched utterances is very challenging for code-switching event detection. In fact, the length of code-switched utterances on average is only 3.65 seconds in the collected corpus, and the speech segments of the embedded language are even much shorter. Moreover, data collection is time consuming and code-switching does not occur all the time during conversation.

2. System Overview

Figure 1 illustrates the framework of the proposed code-switching event detection mechanism, including the training phase (solid-line arrow) and the test phase (dotted-line arrow). In the training phase, Hidden Markov Models (HMMs) are employed to construct a senone-based ASR and the artificial neural network (ANN) is adopted to implement an articulatory feature (AF) detector. The Eigenvoice model (EVM) is one of key component for further language likelihood estimation. Briefly, the EVMs are various matrices employed to model the relationship between senones and articulatory attributes and their corresponding eigenvoices obtained from recognized senones and the detected articulatory features (AFs). We will describe the details in the next section. In the test phase, the acoustic features (ACs) will be transcribed to the syllable sequence by the ASR. Next, ΔBIC-based code-switching (CS) event detection is used to detect the language change-points at the hypothesized inter-syllable boundaries. The concept of ΔBIC is utilized to develop a mask-alike strategy to measure the language likelihood. For an analogy example, the mask of an edge detector in image processing is often adopted to measure the vertical/horizontal intensity change of two sides of a target pixel. For language likelihood estimation, the Euclidean distance is usually used to measure the distance between input data and the given EVMs. However, the acoustic and articulatory features of the phones in one language may have similar identically independent distributions caused by linguistic characteristic. Therefore, the direction measure estimated from the eigenvoice vectors of senones

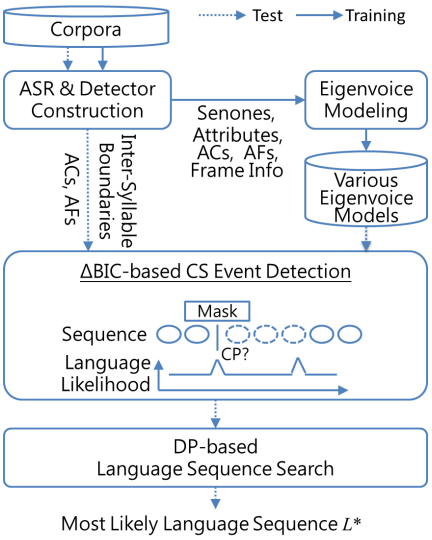


Figure 1: Framework of the proposed code-switching event detection

and articulatory attributes are also included to calculate the language likelihood. Finally, the dynamic programming (DP) algorithm is employed to search the most likely language sequence L^*.

3. Eigenvoice Modeling

Eigenvoice modeling is to model the relationship between eigenvoices and the phonetic units. In this paper, to model a wide range of acoustic-phonetic phenomena, the senones and the articulatory attributes are considered.

For the senones, the context-dependent subphonetic models have been used successfully in many ASR systems. Hence, we treat the states in phonetic HMMs as the basic subphonetic units - senone [20][21][22], which are the state-related modeling units. Senones are implemented by the state-tying technique described in the HTK [23]. For the articulatory attributes, linguists have tried to characterize what makes some phones similar to each other, and what makes them different. Furthermore, the sounds of any language can be further categorized by the attributes of the phones such as bilabial, dental, nasal, and voicing. These articulatory attributes can be classified as either place or manner of articulation [24] and applied successfully to the AF detector [25].

Figure 2 illustrates the diagram of an eigenvoice modeling procedure. All ACs from feature extraction component are used to construct the ASR and the ANN-based AF detector. Moreover, the AFs are the outputs from the ANN-based AF detector; each output corresponds to one articulatory attribute. The ACs and the AFs are used to extract the eigenvoices of senones and articulatory attributes based on principal component analysis (PCA), respectively. The forced alignment component is employed to decode the information of speech frame, including the recognized senone sequence. Finally, given one type of phonetic unit and the corresponding eigenvoices, the matrix $\mathbf{M}_{K\times P}$ is used to model the relationship between K eigenvoices and P phonetic units. Each entry $m_{p,k}$ is defined as

$$m_{p,k} = \frac{\sum_{f_j \to m_p} C(f_j \to E_k)}{N_{m_p}} \quad (1)$$

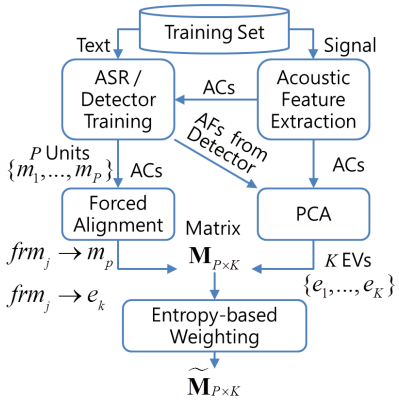


Figure 2: Diagram of Eigenvoice Modeling

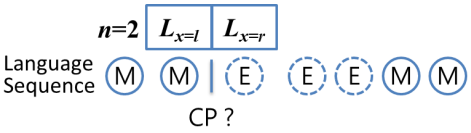


Figure 3: Illustration of mask-alike strategy for CS detection

where $f_j \rightarrow m_p$ is that the j-th frame f_j is decoded as the p-th phonetic unit m_p, $C(\bullet)$ is to count the number of frames assigned to the k-th eigenvoice e_k, and N_{m_P} denotes the total frame number of unit m_P. Although all eigenvectors are included in EVMs, the entropy-based weighting scheme [26] is employed to estimate the importance of each eigenvector for each senone or articulatory attribute. Finally, six EVMs are trained for senone/articulatory attributes and three types of languages (i.e. Mandarin, English, and their Union).

4. Code-Switching Event Detection

Figure 3 illustrate the proposed mask-alike strategy for CS event detection. A sliding window of n syllables is centered at the hypothesized inter-syllable boundary. Then, we define the general forms of similarity scores as

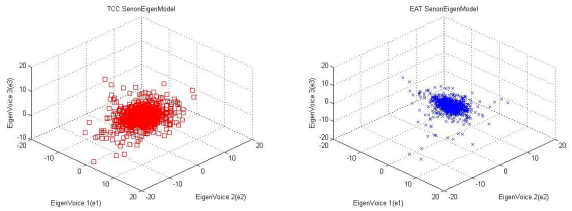
$$S_1 = N_l \log \mathrm{Sim}(L_l) + N_r \log \mathrm{Sim}(L_r) \quad (2)$$
$$S_0 = (N_l + N_r) \log \mathrm{Sim}(L_\cup)$$

where L_l, L_r and $L_\cup$ denote the languages of left, right and the entire segments, N_l and N_r are the number of frames of left and right segments, S_1 denotes the sum of similarity scores of L_l and L_r, S_0 denotes that both L_l and L_r belong to one language, and $\mathrm{Sim}(L_x)$ denotes the similarity score function of a specific language L given the segment x. Furthermore, a general form to estimate the language likelihood is proposed according to the concept of ΔBIC and defined as

$$\Delta BIC(L_r, L_l) = S_1 - S_0 - \frac{p}{2}\log(N_l + N_r) \quad (3)$$

where p is the penalty factor. Based on the idea of Eq.(2), a total of four cases should be considered, i.e.

$$\Delta BIC_1(L_l = M, L_r = E, L_\cup = M), \quad (4)$$
$$\Delta BIC_2(L_l = M, L_r = E, L_\cup = E),$$
$$\Delta BIC_3(L_l = E, L_r = M, L_\cup = M),$$
$$\Delta BIC_4(L_l = E, L_r = M, L_\cup = E).$$



(a) Mandarin /m/ (b) English /m/

Figure 4: Examples of distributions of Mandarin /m/ and English /m/

Figure 4 demonstrates the data distributions of Mandarin phone /m/ and English phone /m/ in top three eigenvoices. These two similar distributions may result in two close distances for the input speech, so, the direction measure is employed to discriminate two languages. For the function $\mathrm{Sim}(L)$, distance measure and direction measure are included and defined as

$$\mathrm{Sim}(L_x) = \omega_{Dist} \times \underset{Dist}{\mathrm{Sim}}(L_x) + \omega_{Dir} \times \underset{Dir}{\mathrm{Sim}}(L_x) \quad (5)$$

where ω_{Dir} and ω_{Dist} are the weights and $\omega_{Dir} + \omega_{Dist} = 1$. Given the weighted EVM $\widetilde{\mathbf{M}}^L$ for the specific language L and $\widetilde{\mathbf{M}}^\cup$ for the Union, assuming J senones and G AFs are within the segment x, the distance measure based on the Euclidean distance is defined as

$$\underset{Dist}{\mathrm{Sim}}(L_x) = \sum_i^{N_x}\Big(\sum_j^J \frac{\|e_j^S(f_i) - \widetilde{\mathbf{m}}_j^{\cup,S}\|}{\|e_j^S(f_i) - \widetilde{\mathbf{m}}_j^{L,S}\|} + \sum_g^G \frac{\|e_g^A(f_i) - \widetilde{\mathbf{m}}_g^{\cup,A}\|}{\|e_g^A(f_i) - \widetilde{\mathbf{m}}_g^{L,A}\|}\Big) \quad (6)$$

and the direction measure based on the inner product is defined as

$$\underset{Dir}{\mathrm{Sim}}(L_x) = \sum_i^{N_x}\Big(\sum_j^J \frac{\|e_j^S(f_i) \cdot \widetilde{\mathbf{m}}_j^{L,S}\|}{\|e_j^S(f_i) \cdot \widetilde{\mathbf{m}}_j^{\cup,S}\|} + \sum_g^G \frac{\|e_g^A(f_i) \cdot \widetilde{\mathbf{m}}_g^{L,A}\|}{\|e_g^A(f_i) \cdot \widetilde{\mathbf{m}}_g^{\cup,A}\|}\Big) \quad (7)$$

where $e_j^S(f_i)$ and $e_g^A(f_i)$ are the eigenvoice vectors of speech frame f_i aligned to the j-th senone through ACs and the g-th articulatory attribute through AFs, $\widetilde{\mathbf{m}}_j^S$ and $\widetilde{\mathbf{m}}_g^A$ with their language index (L and $\cup$) are the eigenvoice vector of the j-th senone and the g-th articulatory attribute in their corresponding weighted EVMS. Note that Eq.(6) and Eq.(7) are developed based on the concept of likelihood ratio. These scores will be used to find the language sequence L^*.

In this paper, given the input speech O and search space Ω_L, L^* is obtained as

$$L^* = \arg\max_{L\in\Omega_L} Pr(O|L)Pr(L) \quad (8)$$
$$\approx \arg\max_{L\in\Omega_L} Pr(S|L)Pr(A|L)Pr(L)$$

where $Pr(S|L)$ and $Pr(A|L)$ are the occurrence probabilities of the senone and articulatory attribute given a specific language in the training set. The $Pr(L)$ is the language transition probability used to restrict the phones near the change-points and can

be estimated according to Eq.(4). They are

$$M \to E \stackrel{def}{=} \Delta BIC_1 + \Delta BIC_2 \quad (9)$$
$$E \to M \stackrel{def}{=} \Delta BIC_3 + \Delta BIC_4$$
$$M \to M \stackrel{def}{=} -(\Delta BIC_1 + \Delta BIC_3)$$
$$E \to E \stackrel{def}{=} -(\Delta BIC_2 + \Delta BIC_4)$$

For the first case, if the language is really transited from Mandarin to English, the value of $Pr(L)$ for twice cross-language transition is larger than that for two transitions within monolanguage. Then, the transition score is mapped to a probability value by means of the sigmoid function. Thus, we can find the language sequence L^* by the DP algorithm.

5. Evaluation

To assess the detection performance of the proposed approach, the English Across Taiwan (EAT) corpus, the TCC-300 corpus and the Chinese-English COde-switching Speech(CECOS) corpus [27], spoken by native Mandarin speakers living in Taiwan were adopted to conduct the evaluation. For the training phase, $5,000$ English utterances from EAT and $5,000$ Mandarin utterances from TCC-300 were used to estimate the monolingual EVMs. Moreover, both of them were employed to estimate the union EVMs. For the test phase, a total of $1,000$ code-switched utterances, 100 Mandarin utterances and 100 English utterances in the CECOS were randomly selected to annotate the language labels.

For system implementation, the ASR was implemented by the HTK. A total of 53 acoustic phones are used to create $10,839$ context-dependent triphone HMMs. Each HMM comprises three left-to-right tied-states with two mixtures in each state. 12 MFCCs with energy and their Δ and $\Delta\Delta$ features were used to extract the AFs. A total of 22 attributes are employed to built the ANN-based AF detector according to [28].

The duration accuracy is the measure commonly used in the LID tasks and is defined as

$$Accuracy = 1 - \frac{Diff}{Ref} \quad (10)$$

where $Diff$ is the difference between the detected boundary and the reference boundary, and Ref is the length of the reference sentence. In addition, to avoid the mismatch between manually tagged boundaries and detection results, a 0.2 second tolerance window will be centered at the reference boundaries. If there is no tolerance window, an ignorable difference will be included to calculate the duration accuracy. In other words, the detected boundaries located within this window will be treated as correct detection. Moreover, silence segments are not included in the calculation of duration accuracy. Note that all reported accuracies are done under the results of forced alignment because of imperfect ASR.

The first evaluation is to find the best combination in the proposed subsystems to obtain a better performance.

Table 1 shows the analysis of the proposed system by fixed values of non-target terms in Eq.(8). At the beginning, the ASR is treated as the baseline and achieved 67.67% recognition accuracy. The Case (1) was to assess the effect on the $Pr(S|L)$ and achieved 66.93 accuracy. The Case (2) was to remove the effects of articulator features; hence, articulatory attribute-level EVMs were also not included in $Pr(L)$. The column captioned (3) was the result of Eq.(8) and the accuracy was degraded to 69.10%. Thus, case (4) was designed to find out the invalid factor. The result was improved while we removed the $Pr(A|L)$. This phenomena occurred because shared articulatory attributes of different phonetic alphabets (e.g., /b/and /p/) in monolingual share the same articulatory attributes result in confusing occurrence probability. Furthermore, we tried more combinations to obtain better results. Based on the previous case, the current best result achieved 71.86% if we removed the senone-level language likelihood estimation. It is reasonable that some phones (e.g., /er/ and /ch/) have different articulatory features in different languages.

Table 1: Accuracy(%) of progressive evaluation of our system

Case	ASR	(1)	(2)	(3)	(4)	(5)
Acc	67.67	66.93	71.69	69.10	71.70	71.86

Table 2: Evaluations of different weighted distance scores

weights	0.0	0.6	1.0
Acc(%)	71.84	71.89	71.61

Next, we will evaluate the effect on the different weighted distance measure in Eq.(5) from 0.0 to 1.0 shown in Table 2. One can observe that the best duration accuracy can be achieved at 71.89% with the weight 0.6. So, the direction score in language likelihood estimation should be taken into account.

The above experiments were conducted by the sliding window within two phones because each Mandarin syllable consists of two subsyllables. So, different window lengths are also evaluated and 71.93% accuracy can be obtained within three phones. After analysing the English word of the CECOS, most of them are comprised of about three phones.

Table 3 shows the results of comparisons among several approaches. To obtain a fair comparison, these approaches were used to replace the language likelihood. The ASR is directly employed to be the bilingual language tag recognizer (BLTR) [29]. The SVM-based approach is a binary identifier with the radial basis function and built by the libsvm. The GMM-based approach is constructed using the SDCs. The case SDC is that the original features, ACs, are replaced with the SDCs in our system. The final one, Propose is our proposed approach. The detection performances of those approaches using the SDCs are close to our system because the SDCs is a kind of feature-based approach.

6. Conclusions

A new methodology for code-switching event detection based on ΔBIC is proposed, implemented and evaluated. The main idea is that various EVMs were built to model the relationship between the senone/articulatory attributes and their corresponding eigenvoices extracted by the PCA using the ACs/AFs, respectively. Moreover, a sophisticated similarity score function based on the concept of ΔBIC is used to estimate the language likelihood. Evaluation results show that the proposed approach using phonetic EVMs outperforms the other LID systems based on SVM/GMM-based methods and language tag recognizer.

Table 3: Comparisons among several approaches

SVM	BLTR	GMM	SDC	Propose
55.63	66.16	70.89	71.91	**71.93**

7. References

[1] W. Li, Ed., *The Bilingualism Reader*. London: Routledge, 2000.

[2] H. Y. Su, "Code-switching between mandarin and taiwanese inthree telephone conversation: The negotiation of interpersonal relationships among bilingual speakers intaiwan," in *Proc. the Symposium about Language and Society*, Apr 2001.

[3] C. Chen, "Two types of code-switching in taiwan," in *Proc. Sociolinguistics Symposium 15 (SS15)*, Newcastle upon Tyne, United Kingdom., Apr 2004.

[4] H. Halmari, *Government and Code-Switching: Explaining American Finnish*. Amsterdam: John Benjamins, 1997.

[5] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. D. Jr., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. INTERSPEECH'2002*, 2002, pp. 1–1.

[6] W. Tsai and W. Chang, "Discriminative training of gaussian mixture bigram models with application to chinese dialect identification," *Speech Communication*, vol. 36, no. 3, pp. 317–326, Mar 2002.

[7] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. Deller, "Language identification using gaussian mixture model tokenization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002)*, vol. 1, May 2002, pp. 757–760.

[8] H. Suo, M. Li, P. Lu, and Y. Yan, "Automatic language identification with discriminative language characterization based on svm," *IEICE - Trans. Inf. Syst.*, vol. E91-D, no. 3, pp. 567–575, Mar 2008.

[9] Y. Deng and J. Liu, "Automatic language identification using support vector machines and phonetic n-gram," in *Proc. IEEE International Conference on Audio, Language and Image Processing (ICALIP'2008)*, Jul 2008, pp. 71–74.

[10] L. Lu, S. Z. Li, and H.-J. Zhang, "Content-based audio segmentation using support vector machines," in *IEEE International Conference on Multimedia and Expo (ICME'2001)*, Aug 2001, pp. 749–752.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[12] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *IEEE Transactions on Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

[13] M. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1994)*, vol. 1, Apr 1994, pp. 305–308.

[14] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, p. 31, Jan 1996.

[15] T. Nagarajan and H. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '2004)*, vol. 1, May 2004, pp. 401–404.

[16] C.-H. Wu and C.-H. Hsieh, "Story segmentation and topic classification of broadcast news via a topic-based segmental model and a genetic algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1612–1623, 2009.

[17] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH'1999)*, Sep 1999.

[18] C.-H. Wu and C.-H. Hsieh, "Multiple change-point audio segmentation and classification using an mdl-based gaussian model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 647–657, Mar 2006.

[19] M. Cettolo and M. Federico, "Model selection criteria for acoustic segmentation," in *in Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, 2000, pp. 221–227.

[20] X. Huang, F. Alleva, H. wuen Hon, M. yuh Hwang, and R. Rosenfeld, "The sphinx-ii speech recognition system: An overview," *Computer, Speech and Language*, vol. 7, pp. 137–148, 1992.

[21] M. Hwang and X. Huang, "Subphonetic modeling with markov states-senone," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1992)*, vol. 1, Mar 1992, pp. 33–36.

[22] H.-P. Shen, C.-H. Wu, and P.-S. Tsai, "Transformation-based accented speech modeling using articulatory attributes for non-native speech recognition," in *Proc. APSIPA Annual Summit and Conference 2011 (APSIPA ASC 2011)*, 2011.

[23] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[24] R. Jakobson, C. G. M. Fant, and M. Halle, "Preliminaries to speech analysis : the distinctive features and their correlates," MIT Press, Tech. Rep., 1952.

[25] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP'04)*, 2004.

[26] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug 2000.

[27] H.-P. Shen, C.-H. Wu, Y.-T. Yang, and C.-S. Hsu, "Cecos: A chinese-english code-switching speech database," in *Proc. International Conference on Speech Database and Assessments (Oriental COCOSDA'2011)*, Oct 2011, pp. 120–123.

[28] S. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2008)*, 2008, pp. 4261–4264.

[29] T. Schultz, I. Rogina, and A. Waibel, "Lvcsr-based language identification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1996)*, vol. 2, May 1996, pp. 781–784.