THE ISL RT04 MANDARIN BROADCAST NEWS EVALUATION SYSTEM

Hua Yu, Yik-Cheung Tam, Thomas Schaaf, Sebastian Stüker, Qin Jin, Mohamed Noamany, Tanja Scultz

Interactive Systems Laboratories
Carnegie Mellon University (USA)
University of Karlsruhe (Germany)
{hyu,yct,tschaaf,qjin,mfn,tanja}@cs.cmu.edu, {stueker}@ira.uka.de

ABSTRACT

This paper describes our effort in developing a Mandarin Broadcast News system for the RT-04f (Rich Transcription) evaluation. Starting from a legacy system, we revisited all the issues including partitioning, acoustic modeling, language modeling, decoding and system combination strategies. We have achieved a sizable improvement, from 21.2% to 5.2% on the development set, from 42.7% to 22.4% measured on the RT-04f evaluation set, over a period of three months.

1. INTRODUCTION

Recognition of Mandarin broadcast news audio has received increased attention over the past several years [1, 2, 3, 4]. The goal is to provide high quality transcripts for Mandarin radio or TV newscast without any human intervention. The challenge is two-fold. First, everyday broadcast news contains a variety of acoustic conditions. In addition to the typical anchor speech, there is also music, phone interviews, foreign language, to name a few. An ASR system must be able to effectively deal with all conditions. Second, Mandarin Chinese is very different from English. For example, Chinese text are not explicitly segmented at the word level; tones play an important role in distinguishing characters.

Our system architecture is shown in Figure 1. First, the audio feed is segmented, classified and clustered. Music segments are discarded; foreign language utterances are tagged and rejected later on. Then, multi-pass decoding and rescoring are carried out on the speech segments. Cross-adaptation is applied between two sets of acoustic models: one based on initial-finals (or demi-syllables), the other based on phones. Several sets of hypotheses are further combined to produce the final hypotheses through consensus network [5].

[1]Note that the RT03 eval set contains 5 shows, 3 of which from mainland and 2 from Taiwan. The mainland shows and Taiwanese shows are very different in terms of both language usage and acoustic conditions. To avoid building separate models for Taiwanese shows, we decided to focus on the mainland part only.

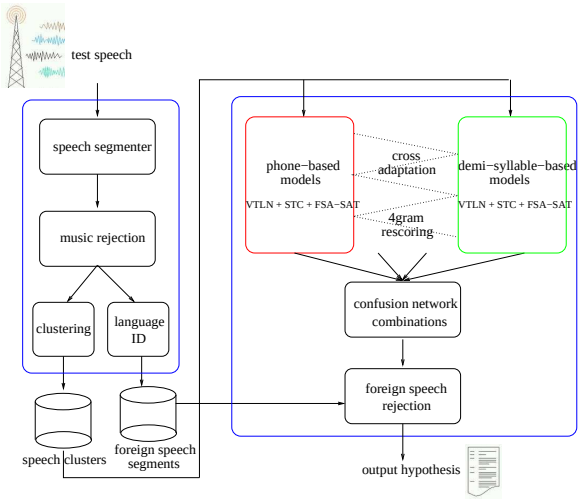

Fig. 1. System Architecture

We used several development sets during our system development (Table 1). For completeness, the RT04 eval set is also listed. We started from a legacy system, which had a CER (Character Error Rate) of 31.6% on the Hub4m97 set, significantly worse than the best system in the 1997 Broadcast News evaluation (19.8%). Over a period of three months, we have drastically improved our system performance. The final system achieves a 5.2% CER on the RT03 eval set, and 22.4% CER on the RT04 evaluation (20.9% without foreign language rejection).

This paper is organized as follows. First, we give a brief overview of Chinese specific issues. We then discuss partitioning, which includes segmentation, music/language classification and clustering. Next, we present issues in acoustic modeling, language modeling and pronunciation lexicon design. Finally, we give decoding results on RT03 and RT04, followed by a detailed analysis.

We remind the reader that since different setups were used during system development, results should be interpreted with respect to the corresponding baseline.

	description	sources	duration	best CER reported
Hub4m97	Hub4 1997 Mandarin eval set	CCTV,VOA,KAZN	60 min.	19.8% (1997)
RT03m	mainland shows of the RT03[1]eval set	CCTV,CNR,VOA	36 min.	6.6% (2003)
Dev04	RT04 dev set	CCTV	32 min.	
RT04	RT04 eval set	CCTV,NTDTV,RFA	60 min.	

Table 1. Various Mandarin broadcast news test sets. CNR stands for China National Radio, CCTV is the official TV station in mainland China, VOA=Voice of America, RFA=Radio Free Asia, KAZN is a chinese radio station in Los Angelos, NTDTV is a chinese TV station (New Tang Dynasty) based in New York.

2. CHINESE SPECIFIC ISSUES

Chinese text is not segmented at the word level. In other words, a sentence is simply a sequence of characters, with no spaces in between. It is not trivial to segment Chinese text into words. To make matters worse, since the distinction between words and phrases is weak, a sentence can have several acceptable segmentations with the same meaning. For language modeling purposes, it is important to have a good word list and to segment the training data properly.

While the number of words can be unlimited, there are only about 6.7K characters in simplified Chinese. Each character is pronounced as a syllable, hence Chinese is a mono-syllabic language. A syllable can have five different tones: flat, rising, dipping, falling, and neutral (unstressed). There are about 1300 unique tonal syllables, or 408 unique syllables disregarding tones. Studies have shown that the realization of tones is context sensitive, an effect known as tone *sandhi*. For example, when a word is comprised of two third-tone characters, the first character will be realized in a second tone.

Pinyin is the official romanization system for Mandarin Chinese. While most European languages are transcribed at the phone level, Pinyin is essentially a demi-syllable level representation, also known as initial-final: an initial is typically a consonant; a final can be either a monophthong, a diphthong or a triphthong. There are 23 initials and 37 finals in Mandarin. Since the Pinyin representation is standard, it is easy to find pronunciation lexicons in this format.

Alternatively, one can use a phonetic representation for pronunciations. The LDC 1997 Mandarin CallHome lexicon contains phonetic transcriptions for about 44K words, using a phone set of 38 phones. While phonemes are well studied and understood, they are not the most natural representation for Chinese. It also remains unclear whether there is a widely accepted phonetic transcription standard for Chinese.

3. PARTITIONING

3.1. Speaker Segmentation and Clustering

The CMU segmenter is used to produce the initial segmentation [6]. The classification and clustering components of the package are not used.

We developed our own GMM-based music classifier, which detects and rejects music segments before clustering. It uses the MFCC feature, its delta and double delta. To train the music classifier, 3 shows are manually annotated, giving 6.4 minutes worth of music and 68 minutes of non-music. The classification criterion is log-likelihood ratio between the two GMMs. The decision boundary is slightly biased towards non-music to avoid mistakenly rejecting speech segments. On the RT04 evaluation set, 59 seconds of music are correctly rejected while all speech segments are retained.

The resulted speech segments are then grouped into several clusters, with each cluster corresponding to an individual speaker ideally. A hierarchical, agglomerative clustering technique is used. It is based on TGMM-GLR distance measurement and the Bayesian Information Criterion (BIC) stopping criteria [7].

We first train a TGMM θ on all speech segments. Adapting θ to each segment generates a GMM (Gaussian mixture model) for that segment. The GLR distance between two segments Seg_a and Seg_b is defined as

$$D(\mathrm{Seg}_a, \mathrm{Seg}_b) = -\log \frac{P(X_a \cup X_b|\theta_c)}{P(X_a|\theta_a)\ P(X_b|\theta_b)}$$

where X_a, X_b are feature vectors in Seg_a and Seg_b, respectively. θ_{a}, θ_b, and θ_c are statistical models built on X_a, X_b, and $X_a \cup X_b$ respectively. A symmetric distance matrix is computed from the pairwise distances between any two segments. At each clustering step, the two segments with the smallest distance are merged, and the distance matrix is updated. We use the BIC stopping criterion. Details are given in the Appendix.

Table 2 shows the differences of speech recognition performance when comparing manual segmentation to automatically segmentation on RT03.

	CER
manual segmentation	6.8%
automatic segmentation	9.9%

Table 2. CERs with different segmentation schemes on RT03

3.2. Language Identification

We have observed a number of foreign language segments, mostly English, in several Chinese news shows. As they cause high insertion errors for our Mandarin ASR system, it is beneficial to detect and discard them. A phonetic language modeling approach [8, 9] is used for this purpose.

Figure 2 illustrates the phonetic language model training and language identification procedure:

Phonetic Language Model Training We use an open-loop Chinese phone recognizer from the GlobalPhone project [10], to decode the Broadcast News shows. The output phone sequences from the Chinese BN shows are used to train the Chinese phonetic language model and the output phone sequences from the English BN show are used to train the English phonetic language model. The Chinese phonetic language model is trained on a 2-hour subset of the 1997 Hub4 Mandarin training data. The English phonetic language model is trained on a 5-hour subset of the 1996 BN English training data. Bigram language model is used in both cases.

Language Identification on testing segments During testing, the speech segment in question is first decoded by the Chinese phone recognizer. Then, the output phone sequence is compared to both the Chinese phonetic language model and the English phonetic language model. The likelihood ratio is used to determine the language identity of the segment. Since any false rejection of a Chinese segment as a English segment translates directly into ASR deletion errors, the threshold is set to favor Chinese.

Table 3 shows the effect of language identification on speech recognition performance. One can clearly see big gains by rejecting English segments from the ASR output.

	RT03	Dev04
before LID	5.9%	18.4%
after LID	5.2%	16.6%

Table 3. CER on development data set

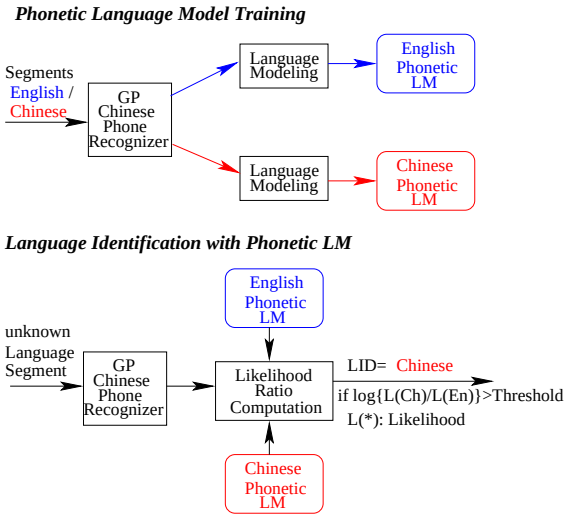

Fig. 2. Language Identification

	CER
baseline	20.6%
+ VTLN	19.6%
+ STC	18.4%

Table 4. Effect of VTLN and STC on Hub4m97

4. ACOUSTIC MODELING

For feature extraction, we use 13 Mel-Frequency cepstral coefficients (MFCC) per frame. Cepstral mean and variance normalization is performed on a speaker/cluster basis. Dynamic features are extracted by concatenating 15 adjacent frames, then using linear discriminant analysis (LDA) to produce the final feature vector of 42 dimensions [11]. Vocal tract length normalization (VTLN) is performed on a speaker/cluster basis.

As described before, the acoustic modeling units can be either initial-finals (IF) or phones. In both cases, context-dependent models are built and then clustered using decision trees. The IF system has 3000 clustered triphone states and a total of 168k Gaussians; the phone system has 3000 tied septaphone states with a total of 169k Gaussians. We find that both systems give comparable performance, with the IF-system slightly better than the phone-based system. Hence, both systems are retained so that we can take advantage of system combination during decoding.

We use maximum likelihood training for both sets of models. The Gaussian mixtures are grown incrementally over several iterations. A single global semi-tied covariance matrix (STC) is employed [12]. Furthermore, speaker-adaptive training is performed, using a single feature space transforms per speaker (FSA-SAT). Table 4 and 5 illustrate the effect of VTLN, STC and FSA-SAT.

	CER
VTLN,STC	11.4%
+ FSA-SAT	9.6%

Table 5. Effect of FSA-SAT on RT03

The acoustic training data consists of two parts: 27 hours of manually transcribed Mandarin Broadcast News data, and 85 hours of quickly transcribed TDT4 data. The TDT4 data does not have noise annotations and may include minor transcription errors. The TDT4 segments in the original transcripts are very long and often include more than one speakers per segment. Hence, we resegmented the TDT4 data at major silences located through forced alignment.

4.1. Handling of Tones

As discussed in Section 2, tones carry important information to disambiguate characters. It is natural to use tonal units in acoustic modeling. In practice, we observed that certain tonal variants of a final/vowel have very few instances during training. As suggested in [1], we adopted a better "soft-tone" approach where tonal information is used only in decision trees. A single decision tree is grown for all tonal variants of the same phone/final. Different tonal variants of the same phone/final can either have separate models or share the same model, determined completely in a data-driven fashion. This turns out to be a special case of single tree clustering [13]. It makes even more sense if we consider the tone sandhi effect.

Another issue is that MFCC coefficients were designed to capture spectral envelopes only, while suppressing tonal information. A popular solution is to extract pitch features in conjunction with the MFCC features. We have not yet explored this option due to time constraints.

4.2. Topology Experiments

For the phone-based system, we can extend the common practice in English: use three states per phone. Three states works for initials too, since they correspond to consonants. In contrast, different finals have very different durations and therefore warrant different numbers of states. Monophthongs are the shortest, where 3 states might be enough. Diphthongs and triphthongs are much longer and probably should have proportionally more states.

It is, however, not easy to determine the optimal number of states for different finals. There are two issues: durational constraints and temporal modeling resolution. In Table 6, we experimented with the durational constraints. Our baseline IF-model is trained using 3 states for initials and 5 states for finals, with 3 duplicate middle states. The baseline has a CER of 12.0%. Using the same model, but a 3-state topology during decoding, CER remains virtually the same, 12.2%. We then decoded with variable numbers of states (max=6) for each final, where the number of states is determined by statistics collected during training. CER remains unchanged: 12.1%. We also tried to use the simple 3-state topology for both training and decoding, which gives a CER of 12.0%. It appears that the performance is not sensitive to durational constraints at all. Later on, we switched to using 4 *different* states per final, instead of duplicating the middle state. This appears to give slightly better performance and is kept as the setup for our final IF-system.

	CER
5 states (bmmme)	12.0%
decoding with 3 states (bme)	12.2%
decoding with variable #states (max=6)	12.1%
3 states (both training and decoding)	12.0%

Table 6. Topology Experiments on RT03

5. LANGUAGE MODELING AND PRONUNCIATION LEXICON

5.1. Language Modeling

We used several corpora for our LM development: Mandarin Chinese News Text (LDC95T13), TDT{2,3,4}, Mandarin Gigaword corpus and the HUB4 1997 acoustic training transcript. Since the RT04 eval set contains two previously unseen sources, RFA and NTDTV, we also crawled the web to find relevant text material. Any text that falls into the excluded time frame (specified in the RT04-eval specification) was removed.

Before training a LM, we first processed the Chinese text data to normalize for ASCII numbers, ASCII strings and punctuations. We devised heuristic rules in combination with a Maximum Entropy (Maxent) classifier to normalize the numbers. The classifier classifies whether the input number is a digit string (e.g. telephone number) or a number quantity based on the surrounding word context. We mapped English words to a special token "+english+", human noises (such as breath and cough) to "+human_noise+". Non-human (environmental) noises were removed from the HUB4 training transcript. Since punctuations provide word boundary information which is useful for word segmentation, they were removed after word segmentation.

Word segmentation is based on a maximal substring matching approach which locates the longest possible word segment at each character position. Since proper names were often incorrectly segmented, we later on added the LDC Named-Entity (NE) list into the original wordlist (in the official LDC segmenter). The NE list contains different semantic categories, such as organization, company, person

and location names. Having them in the wordlist greatly improved segmentation quality, which translates to more accurate predictions in the ngram LM.

After word segmentation, we chose the vocabulary to be the top-N most frequent words. The commonly used Chinese characters (6.7k) is then added into the vocabulary. We trained a trigram LM as well as a 4-gram LM using the SRI LM toolkit with Kneser-Ney smoothing.

As shown in Table 7, several language models were used at different development stages. The corresponding perplexities and CERs are shown in Table 8. We observed nice gains by simply adding more and more text data. Interestingly, adding the Gigaword corpus only gave a marginal gain on the RT04 set; using the LDC NE list helps on the RT04 set, but not on the RT03-eval set.

As a reminder, since different LMs have different vocabulary sizes, we cannot compare perplexities across LMs. However, we can compare the perplexity on different data set for the same LM. From the table, it is clear that the perplexity on RT04 more than doubles that on RT03, which indicates significant mismatches between the two.

5.2. Pronunciation Lexicon

Our pronunciation lexicon was based on the LDC CallHome Mandarin lexicon, which contains about 44k words. Pronunciations for words not covered by the LDC lexicon were generated using a maximal matching method. The idea is similar to our word segmentation algorithm. We first compiled a list of all possible character segments for each covered vocabulary word. For each uncovered word, the algorithm repeatedly searches for the longest matching character segment from the beginning to the end of the word, producing a sequence of character segments. Pronunciations of these segments are then concatenated to produce the pronunciation for the new word.

We employed both demi-syllables (Initials/Finals) and phonemes as acoustic units and used them to train two separate acoustic models. There are 23 initials and 37 finals, and 38 phonemes defined by the CallHome lexicon. Eight additional phonemes were used to model human noises, environmental noises and silence. We used the demi-syllable-to-phoneme mappings provided by the lexicon to convert a demi-syllable lexicon into a phone-based lexicon.

6. DECODING

The IBIS single pass decoder is used to decode the evaluation data [14]. Since there are two sets of comparable acoustic models, we apply cross-adaptation between the two systems to progressively refine the hypotheses. Adaptation is carried out in both the model space (maximum likelihood linear regression, MLLR) and the feature space (FSA). A 4-gram language model is further used for lattice rescoring. We then apply confusion networks [5] to combine five different of hypotheses from earlier stages. Table 9 shows the decoding passes used in the RT04 evaluation. The total processing time is about 26 times real-time on a single 3.2GHz Pentium4 Linux box.

	RT03	RT04	comments
pass 1	8.7%	28.4%	IF-sys
pass 2	7.1%	23.2%	IF-sys
pass 3	6.8%	22.1%	phone-sys
pass 4	6.4%	21.5%	IF-sys, 4gram rescoring
pass 5	6.3%	21.7%	phone-sys, 4gram rescoring
pass 6	6.7%	21.4%	IF-sys, 8ms frame shift
	6.7%	21.9%	phone-sys, 8ms frame shift
pass 7	6.0%	20.9%	consensus network combination
pass 8	5.2%	22.4%	foreign language rejection

Table 9. Multi-pass Decoding on RT03 and RT04

7. ANALYSIS

As shown in Table 9, we found that foreign language rejection actually hurts us in the RT04 eval. Table 10 lists the character error rates for each show. We can see that language identification does help for CCTV and NTDTV. It unfortunately fails on the RFA show. Analysis indicates a significant amount of narrow-band speech in the RFA show, which causes some Chinese segments being misclassified as English and rejected. Table 10 also lists perplexities for each show in RT04. We can see that the perplexities on RFA and NTDTV are a lot higher than that on the CCTV show. Overall, the RT04 evaluation data is very different from our development sets, which renders some of our design decisions suboptimal.

show	CCTV	NTDTV	RFA	Overall
perplexity	302	584	702	497
before LID	12.4%	17.7%	34.1%	20.9%
after LID	12.3%	16.9%	40.4%	22.4%

Table 10. Perplexity and CER breakdown on RT04 shows

8. SUMMARY

We described the development of ISL's 2004 Mandarin Broadcast News evaluation system. As shown in Table 11, over a period of three months, we achieved a 76% relative improvement on the RT03 mainland set, 51% relative improvement on the RT04 evaluation set.

We have not thoroughly explored all the issues due to the tight schedule. In the future, we would like to investi-

LM	# characters	Vocab size	# 2-grams	# 3-grams
Small	26M (Hub4 transcripts, XH)	40k	1M	1.4M
Medium	247M (+TDT{2,3}, PD, CR)	51k	12M	15.8M
Big	621M (+Gigaword, TDT4, web)	51k	19M	13.6M
Big (resegmented)	621M	63k	24.9M	10M

Table 7. LM development by increasing the amount of training text data (XH, PD and CR refer to Xinhua news, People's daily and China Radio respectively contained in the Mandarin Chinese News Text Corpus).

LM	RT03			RT04		
	OOV rate	perplexity	CER	OOV rate	perplexity	CER
Small	0.2%	491	13.7%	2.0%	962	34.5%
Medium	0.2%	238	10.0%	0.4%	474	30.0%
Big	0.2%	170	8.7%	0.4%	432	29.8%
Big(resegmented)	0.6%	206	8.7%	1.3%	497	28.4%

Table 8. LM performances on RT03 and RT04. CERs are based on first-pass decoding using the demi-syllable system.

Test set	legacy system	final system
Hub4m97	31.6%	-
RT03 mainland	21.2%	5.2%
RT04	42.7%	22.4%

Table 11. Overall System Improvements

gate/revisit acoustic segmentation, lightly supervised training on the TDT4 data, as well as the use of pitch features.

9. REFERENCES

[1] P. Zhan, S. Wegmann, and S. Lowe, "Dragon systems' 1997 mandarin broadcast news system," in *DARPA Broadcast News Workshop*, 1997.

[2] D. Liu, J. Ma, D. Xu, A. Srivastava, and F. Kubala, "Real-time rich-content transcription of chinese broadcast news," in *Proc. ICSLP*, 2002.

[3] L. Nguyen, B. Xiang, and D. Xu, "The BBN RT03 BN mandarin system," in *DARPA RT-03 Workshop*, Boston, 2003.

[4] L. Lamel, L. Chen, and J. Gauvain, "The LIMSI RT03 mandarin broadcast news system," in *DARPA RT-03 Workshop*, Boston, 2003.

[5] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus among words: Lattice-based word error minimization," in *Proc. EuroSpeech*, 1999.

[6] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *DARPA Speech Recognition Workshop*, 1997.

[7] Q. Jin and T. Schultz, "Speaker segmentation and clustering in meetings," in *Proc. ICSLP*, 2004.

[8] M.A. Zissman, "Language identification using phone recognition and phonotactic language modeling," in *Proc. ICASSP*, 1995.

[9] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, "Speaker, accent, and language identification using multilingual phone strings," in *Proceedings of the Human Language Technology Conference (HLT)*, 2002.

[10] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, August 2001.

[11] H. Yu and A. Waibel, "Streamlining the front end of a speech recognizer," in *Proc. ICSLP*, 2000.

[12] M.J.F. Gales, "Semi-Tied Full-Covariance matrices for hidden markov models," Tech. Rep., Cambridge University, England, 1997.

[13] H. Yu and T. Schultz, "Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversation Speech Recognition," in *Proc. EuroSpeech*, 2003.

[14] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*, 2001.

[15] S.S. Chen and P.S. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, 1998.

Appendix: Speaker Change Detection using Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a model selection criterion widely used in statistics. It was introduced for speaker clustering in [15]. The Bayesian Information Criterion states that the quality of a model M to represent data $\{x_1, \ldots, x_N\}$ is given by

$$BIC(M) = \log L(x_1, \ldots, x_N | M) - \frac{\lambda}{2} V(M) \log N \quad (1)$$

with $L(x_1, \ldots, x_N | M)$ representing the likelihood of model M and $V(M)$ representing the complexity of model M, equal to the number of free model parameters. Theoretically, λ should equal to 1, but it is a tunable parameter in practice.

The problem of determining if there is a speaker change at point i in data $X = \{x_1, \ldots, x_N\}$ can be converted into a model selection problem. The two alternative models are: (1) model M_1 assumes that X is generated by a multi-Gaussian process, that is $\{x_1, \ldots, x_N\} \sim N(\mu, \Sigma)$, or (2) model M_2 assumes that X is generated by two multi-Gaussian processes, that is

$$\begin{aligned} \{x_1, \ldots, x_i\} &\sim N(\mu_1, \Sigma_1) \\ \{x_{i+1}, \ldots, x_N\} &\sim N(\mu_2, \Sigma_2) \end{aligned}$$

The BIC values for the two models are

$$\begin{aligned} BIC(M_1) &= \log L(x_1, \ldots, x_N | \mu, \Sigma) - \frac{\lambda}{2} V(M_1) \log N \\ BIC(M_2) &= \log L(x_1, \ldots, x_i | \mu_1, \Sigma_1) \\ &\quad + \log L(x_{i+1}, \ldots, x_N | \mu_2, \Sigma_2) \\ &\quad - \frac{\lambda}{2} V(M_2) \log N \end{aligned}$$

The difference between the two BIC values is

$$\begin{aligned} \Delta BIC &= BIC(M_1) - BIC(M_2) \\ &= \log \frac{L(x_1, \ldots, x_N | \mu, \Sigma)}{L(x_1, \ldots, x_i | \mu_1, \Sigma_1) L(x_{i+1}, \ldots, x_N | \mu_2, \Sigma_2)} \\ &\quad + \frac{\lambda}{2} [V(M_2) - V(M_1)] \log N \end{aligned}$$

If the value of ΔBIC is negative, it claims that model M_2 fits the data better, which means that there is a speaker change at point i. Therefore, we continue the segments merging until the value of ΔBIC for the two closest segments (candidates for merging) is negative.