Edit Disfluency Detection and Correction Using a Cleanup Language Model and an Alignment Model

Jui-Feng Yeh and Chung-Hsien Wu, *Senior Member, IEEE*

***Abstract*—This investigation presents a novel approach to detecting and correcting the edit disfluency in spontaneous speech. Hypothesis testing using acoustic features is first adopted to detect potential interruption points (IPs) in the input speech. The word order of the cleanup utterance is then cleaned up based on the potential IPs using a class-based cleanup language model, the deletable region and the correction are aligned using an alignment model. Finally, log linear weighting is applied to optimize the performance. Using the acoustic features, the IP detection rate is significantly improved especially in recall rate. Based on the positions of the potential IPs, the cleanup language model and the alignment model are able to detect and correct the edit disfluency efficiently. Experimental results demonstrate that the proposed approach has achieved error rates of 0.33 and 0.21 for IP detection and edit word deletion, respectively.**

***Index Terms*—Edit disfluency, language model, potential interruption point (IP) detection, rich transcription.**

I. Introduction

As information exchangeability and communication becomes increasingly human oriented, human–computer interfaces that provide interconnective services are increasingly important. Significant effort has been made in the last few decades to develop automatic speech recognition technology to enrich the output in the most informative manner using spontaneous speech. The transcription, that converts speech into written or typewritten form, symbolically represents an acoustic signal for indexing, retrieval, and analysis [1]. Rich transcription has emerged as an interdisciplinary field combining automatic speech recognition, speaker diarization, and natural language processing. Moreover, the concept of rich transcription comprises generating an orthographic transcription of a spontaneous speech utterance enriched with side information concerning the speaker, the presence of background sound, the topics, and any information related to the structure of the speech utterance [2], [3]. Shriberg defined and analyzed the disfluencies of spontaneous speech in 1994. The annotated information includes observable characteristics of the speech domain, speaker, sentence in which a disfluency occurs, and word-related and simple acoustic properties of the disfluency [4]. Recent research has attempted to enhance acoustic and language modeling for speech recognition and novel algorithms for extracting structural metadata sentence boundaries and disfluencies. Combined with the linguistic applications, rich transcription can be applied to dictation systems, speech-to-speech translation, spoken language understanding, and transcription system for meeting.

Two essential conceptual issues in rich transcription are speech-to-text (STT) and metadata extraction (MDE). Well-known research projects such as Effective, Affordable, Reusable Speech-to-text Program (EARS), which is supported by DARPA, have developed many tools and corpora for rich transcription especially in speech recognition related work. A famous workshop, The Rich Transcription evaluation series (RT), which opened a rich transcription track by NIST since 2002, provides a rich corpus and evaluation material. Speech recognition systems have been an important issue for rich transcription research in past RT workshops. The Disfluency in Spontaneous Speech (DiSS) workshop considers the challenge of research of spontaneous speech, particularly in disfluency. While these workshops have undoubtedly provided important research on rich transcription, there must be considerable doubt as to deal with the problems resulted from the disfluency without exception. One of the essential issues is how to model the edit disfluency. Conversational utterances have several problems, including interruption, correction, filled pauses, and grammatically incorrect sentences. The definitions of disfluencies have been discussed in SimpleMDE [5], where simple edit disfluencies are divided into three categories, namely repetitions, revisions or repairs, and restarts. Each category is illustrated as follow.

1) Repetition: The abandoned words repeated in the corrected portion of the utterance, as depicted in Fig. 1(b).
2) Revision or repair: Although similar to repetitions, the corrected portion that replaces the abandoned constituent modifies its meaning, rather than repeating it. Fig. 1(c) shows an example.
3) Restarts or false starts: A speaker abandons an utterance, and neither corrects it nor repeats it partially or wholly, but instead restructures the utterance and starts over, as illustrated in Fig. 1(d).

To circumvent the difficulties caused by speech repair, a reliable model must be adopted to detect and correct conversation utterances with disfluency [6]. Several cues exist to suggest when an edit disfluency may occur in the spontaneous speech. These cues can be observed from language features, acoustic features [7] and integrated knowledge sources [8]. Shriberg *et al.* [9] outlined the phonetic consequences of disfluency to improve disfluency processing methods in speech applications. Savova and

Manuscript received October 1, 2005; revised May 19, 2006. This work was supported by the National Science Council, Taiwan, R.O.C., under Contract NSC 94-2213-E-006-018. The associate editor coordinating the review of this paper and approving it for publication was Dr. Geoffrey Zweig.

The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Tainan, Taiwan, R.O.C (e-mail: jfyeh@csie.ncku.edu.tw; chwu@csie.ncku.edu.tw).

Digital Object Identifier 10.1109/TASL.2006.878267

Chinese sentence: 我(I) 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)
English translation: (I want to go to Taipei tomorrow.)

(a) An example of fluent utterance

Chinese sentence: 我(I) 明天(tomorrow) * 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)
English translation: (I want to go to Taipei tomorrow * tomorrow.)
Correction : 我(I) ~~明天(tomorrow)~~ * 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)

(b) An example of repetition

Chinese sentence: 我(I) 今天(today) * 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)
English translation: (I want to go to Taipei today * tomorrow.)
Correction : 我(I) ~~今天(today) *~~ 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)

(c) An example of revision

Chinese sentence: 我(I) * 我(I) 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)
English translation: (I * I want to go to Taipei tomorrow.)
Correction : ~~我(I) *~~ 我(I) 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei)

(d) An example of restart

Chinese sentence: 我(I) * 我(I) 明天(tomorrow) 想要(want to) 去(go to) 台北(Taipei) * 高雄(Kaohsiung)
English translation: (I * I want to go to Taipei * Kaohsiung tomorrow.)
Correction : ~~我(I)~~ * 我(I) 明天(tomorrow) 想要(want to) 去(go to) ~~台北(Taipei) *~~ 高雄(Kaohsiung)

(e) An example of complex edit disfluency

Fig. 1. Some instances of edit disfluency for each category.

Bachenko presented four disfluency rules using intonation, segment duration and pause duration [10]. IBM has adopted a discriminatively trained full covariance Gaussian system [11]. Kim *et al.* utilized a decision tree to detect the structural metadata [12]. Furui *et al.* [13] presented the corpus collection, analysis and annotation for conversational speech processing.

Charniak and Johnson [14] proposed the architecture for parsing transcribed speech using an edit word detector to remove edit words or fillers from the sentence string, and adopted a standard statistical parser to parse the remaining words. The statistical parser and the parameters estimated by boosting are employed to detect and correct the disfluency. Lease *et al.* later presented a novel parser to model the structural information of spoken language by tree adjoining grammars (TAGs) [15], [16]. They proposed a method comprising a TAG transducer and a reranker based on maximum entropy for three specific metadata extraction (MDE) tasks: edit word detection, filler word detection, and interruption point (IP) detection. The TAG transducer in the channel model is responsible for generating the words in the reparandum and the interregnum of a speech repair from the corresponding repair. The maximum entropy reranker chooses the best edit word hypothesis from the list of the 25 best analyses for each sentence-like unit generated by the TAG channel model [17]. Heeman presented a statistical language model that can identify POS tags, discourse markers, speech repairs, and intonational phrases [18], [19]. By solving these simultaneously, the detection of edit disfluency was addressed separately. The noisy channel model was proposed to model the edit disfluency [15], [20], [21]. Snover *et al.* [22] integrated the lexical information and rules generated from 33 rule templates for disfluency detection. Hain *et al.* [23] presented techniques in front-end processing, acoustic modeling, language, and pronunciation modeling for transcribing conversational telephone speech automatically. Soltau *et al.* transcribed telephone speech using LDA [24]. Harper *et al.* utilized parsing approaches to rich transcription [25]. Liu *et al.* not only detected the boundaries of sentence-like units using the conditional random fields [26], but also compared the performances of HMM, maximum entropy [27], and conditional random fields on disfluency detection [28].

This paper focuses on the detection and correction of the edit disfluency based on the word order information. The first process attempts to detect the IPs based on hypothesis testing. Acoustic features, including duration, pitch, and energy features, are adopted in hypothesis testing. To circumvent the problems resulting from disfluency, particularly in edit disfluency, a reliable and robust language model must be adopted to correct speech recognition errors. To handle language-related phenomena in edit disfluency, a cleanup language model characterizing the structure of the cleanup sentences and an alignment model for aligning words between deletable region and correction part are proposed for detecting and correcting edit disfluency.

The rest of this paper is organized as follows. Section II briefly introduces the framework of the edit disfluency transcription system. Section III then presents the use of hypothesis testing to detect potential IPs according to acoustic features. Next, Section IV defines the cleanup language model to deal with the phenomena of edit disfluency. Section V summarizes experimental results, along with a discussion made of those results. Conclusions are finally made in Section VI, along with directions for future research.

II. Framework of Transcription System for Edit Disfluency

Edit disfluencies refer to portions of speech in which a speaker's utterance is incomplete or disfluent; instead, the speaker corrects or alters the utterance, or abandons it entirely and starts over [5]. An edit disfluency has a complex internal structure comprising the deletable region (delreg), IP, and correction. Editing terms, such as fillers, particles, and markers, are optional and follow the IP in edit disfluency. In spontaneous speech, acoustic features such as short pause (silence and filler),

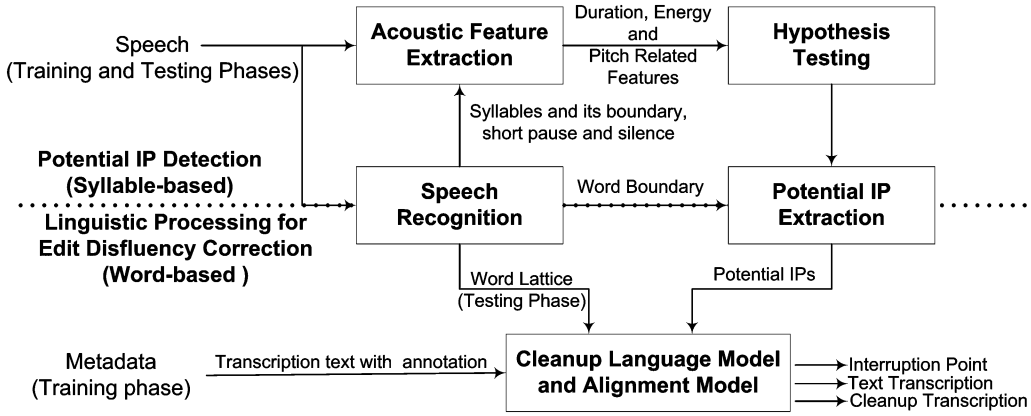

Fig. 2. Framework of transcription system for spontaneous speech with edit disfluencies.

energy, and pitch reset generally appear along with edit disfluency. The possible IPs can be detected from these features. Moreover, since the IPs generally appear at the boundary of two successive words, the unlikely IPs whose positions are within a word can be excluded. Additionally, since the structural patterns between deletable and corrected word sequences are very similar, a deletable word sequence in edit disfluency can be replaced by the corresponding correction word sequence.

According to Fig. 2, the overall transcription task in conversational speech with edit disfluency in the proposed method consists of two main mechanisms: IP detection module and edit disfluency correction module. The IP detection module first predicts the potential IPs. The edit disfluency correction module generates the rich transcription that contains information of interruption, text transcription from the speaker's utterances, and the cleaned-up text transcription without disfluencies.

The speech signal is fed into both the acoustic feature extraction module and the speech recognition engine in IP detection module. Information concerning durations of syllables and silences from speech recognition is provided for acoustic feature extraction. Features related to duration, pitch, and energy, along with side information from speech recognition, are extracted and used to model the IPs by a Gaussian mixture model (GMM). Additionally, in order to perform hypothesis testing on IP detection, an anti-IP GMM is constructed based on the extracted features from non-IP regions. The hypothesis testing verifies whether the posterior probability of the acoustic features of a syllable boundary exceeds a threshold and, therefore, determines whether the syllable boundary is an IP. Since IP is an event that arises in an interword location [29], the detected IPs that do not appear in the word boundary can be removed.

The edit disfluency correction module has two processing stages, cleanup and alignment. The cleanup process divides a word string into three components, the deletable region (delreg), the editing term and correction, according to the locations of potential IPs detected by the IP detection module. The cleanup process is performed by shifting the correction part and replacing the deletable region to form a new cleanup transcription. The edit disfluency correction module comprises an n-gram language model and the alignment model. The n-gram model considers the cleanup transcriptions as fluent utterances, and models their word order information. The alignment model finds the optimal correspondence between the deletable region and the correction in edit disfluency.

III. Potential Interruption Point Detection

Since precise IP detection using acoustic features is still an open issue, this study considers whether linguistic features could be integrated to provide a more reliable solution. Rather than detecting exact IPs, additional candidates called potential IPs are selected for further linguistic processing. The location information of a potential IP can help the language model verify an IP reliably and, therefore, correct the edit disfluency. Since the IP is the point at which the speaker breaks off the deletable region and the correction, it has some characteristic acoustic events. In a syllabic language like Chinese, every character is pronounced as a monosyllable, while a word comprises one or more syllables. The speech input of a syllabic language with n syllables can be described as a sequence,

$$\mathrm{Seq}_{\mathrm{syllable\_silence}} \equiv \{\mathrm{syllable}_1, \mathrm{silence}_1, \mathrm{syllable}_2, \mathrm{silence}_2, \ldots, \mathrm{silence}_{n-1}, \mathrm{syllable}_n\}.$$

This sequence can then be separated into a syllable sequence

$$\mathrm{Seq}_{\mathrm{syllable}} \equiv \{\mathrm{syllable}_1, \mathrm{syllable}_2, \ldots, \mathrm{syllable}_n\}$$

and a silence sequence

$$\mathrm{Seq}_{\mathrm{silence}} \equiv \{\mathrm{silence}_1, \mathrm{silence}_2, \ldots, \mathrm{silence}_{n-1}\}.$$

The interruption detection problem is modeled as a choice between H_0, which is the IP that is not embedded in the silence hypothesis, and H_1, which is the IP that is embedded in the silence hypothesis. The likelihood ratio test is employed to detect the potential IPs. The function $L(\mathrm{Seq}_{\mathrm{syllable\_silence}})$ is called the likelihood ratio, since it indicates the likelihood of H_1 versus the likelihood of H_0 for each value of $\mathrm{Seq}_{\mathrm{syllable\_silence}}$

$$L\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}\right) = \frac{P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}; H_1\right)}{P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}; H_0\right)}. \quad (1)$$

By introducing the threshold γ to adjust the precision and recall rates, $H_1 : L\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}\right) \geq \gamma$ means the IP is embedded in $\mathrm{silence}_k$, which is conceptually a potential IP. Under the assumption of independence, the probability of IP appearing in $\mathrm{silence}_k$ can be considered as the product of probabilities obtained from $\mathrm{silence}_k$ and the syllables around it. The probability density functions (pdf's) under each hypothesis are denoted and estimated as

$$\begin{aligned} &P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}; H_1\right) \\ &\quad = P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}} \mid E_{ip}\right) \\ &\quad = P\left(\mathrm{Seq}_{\mathrm{silence}} \mid E_{ip}\right) \times P\left(\mathrm{Seq}_{\mathrm{syllable}} \mid E_{ip}\right) \end{aligned} \tag{2}$$

and

$$\begin{aligned} &P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}}; H0\right) \\ &\quad = P\left(\mathrm{Seq}_{\mathrm{syllable\_silence}} \mid \neg E_{ip}\right) \\ &\quad = P\left(\mathrm{Seq}_{\mathrm{silence}} \mid \neg E_{ip}\right) \times P\left(\mathrm{Seq}_{\mathrm{syllable}} \mid \neg E_{ip}\right) \end{aligned} \tag{3}$$

where E_{ip} denotes that IP is embedded in $\mathrm{silence}_k$, and $\neg E_{ip}$ means that IP does not appear in $\mathrm{silence}_k$.

A. Probability of Interruption Point Using Posterior Probability of Silence Duration

Since the IP always appears at the intersyllable position, the $n-1$ silences between n syllables are considered as the $n-1$ IP candidates. Thus, IP detection means verifying whether each of the $n-1$ silences contains the IP. In conversation, speakers may hesitate to speak the correct words and, therefore, generate edit disfluency. Hesitation is generally realized as a pause. Since the length of silence is very sensitive to disfluency, two normal distributions are adopted to model the posterior probabilities of IP appearing and not appearing in $\mathrm{silence}_k$

$$\begin{aligned} P\left(\mathrm{Seq}_{\mathrm{silence}} \mid E_{ip}\right) &= \frac{2}{\sqrt{2\pi}\sigma_{ip}} \\ &\quad \times \exp\left(-\frac{\left(\mathrm{Seq}_{\mathrm{silence}} - \mu_{ip}\right)^2}{2\sigma_{ip}^2}\right) \end{aligned} \tag{4}$$

$$\begin{aligned} P\left(\mathrm{Seq}_{\mathrm{silence}} \mid \neg E_{ip}\right) &= \frac{2}{\sqrt{2\pi}\sigma_{nip}} \\ &\quad \times \exp\left(-\frac{\left(\mathrm{Seq}_{\mathrm{silence}} - \mu_{nip}\right)^2}{2\sigma_{nip}^2}\right) \end{aligned} \tag{5}$$

where μ_{ip}, μ_{nip}, σ_{nip}^2, and σ_{ip}^2 denote the means and variances of the silence duration containing and not containing the IP, respectively.

B. Syllable-Based Acoustic Feature Extraction

Acoustic features such as duration, pitch, and energy for each syllable are useful for spontaneous speech recognition [13]. Stolcke *et al.* have proven the effectiveness of these features and have significantly improved automatic detection of sentence boundaries and disfluencies [30]. Accordingly, this study adopts these features for IP detection. Because silence is treated as a potential IP, a feature vector of the syllables within an observation window around the silence is constructed as the input of the GMM for IP detection. A window of $2w$ syllables with w syllables after and before $\mathrm{silence}_k$ is adopted. Fig. 3 illustrates the flow for acoustic feature vectorization.

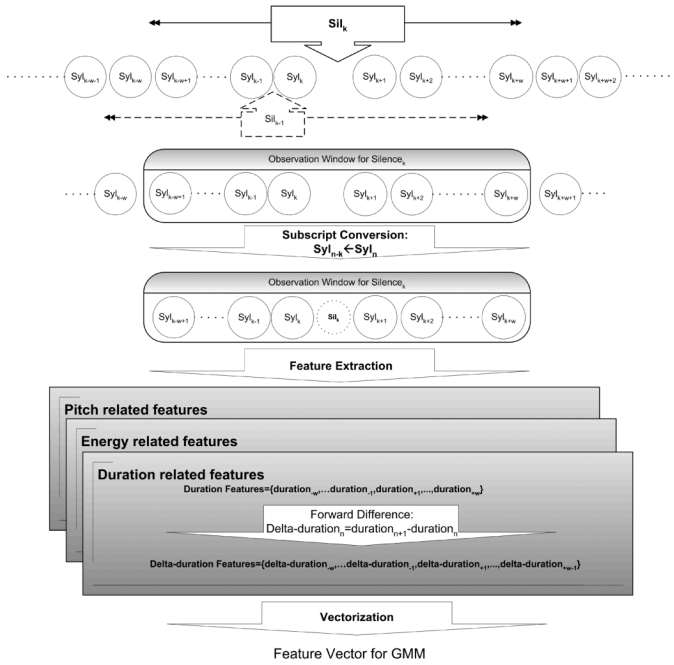

Fig. 3. Flow of acoustic feature extraction.

Since the durations of syllables are not the same even for the same syllable, the duration ratio is defined as the average duration of the syllable normalized by the average duration over all syllables

$$nf_{\mathrm{duration}_i} \equiv \frac{\sum_{j=1}^{n_i} \mathrm{duration}\left(\mathrm{syllable}_{i,j}\right)}{\sum_{i=1}^{|\mathrm{syllable}|} \sum_{j=1}^{n_i} \mathrm{duration}\left(\mathrm{syllable}_{i,j}\right)} \tag{6}$$

where $\mathrm{syllable}_{i,j}$ denotes sample j of syllable i in the corpus; $|syllable|$ is the total number of the syllables in the corpus; n_i represents the number of the syllable i in the corpus. Similarly, frame-based statistics are adopted used to calculate the normalized energy and pitch for each syllable. The estimation is defined as follows:

$$nf_{\mathrm{pitch}_i} \equiv \frac{\sum_{j=1}^{n_i} \sum_{k=1}^{|\mathrm{frame}_{i,j}|} \log\left(\mathrm{ptich}\left(\mathrm{frame}_{i,j,k}\right)\right)}{\sum_{i=1}^{|\mathrm{syllable}|} \sum_{j=1}^{n_i} \sum_{k=1}^{|\mathrm{frame}_{i,j}|} \log\left(\mathrm{ptich}\left(\mathrm{frame}_{i,j,k}\right)\right)} \tag{7}$$

$$nf_{\mathrm{energy}_i} \equiv \frac{\sum_{j=1}^{n_i} \mathrm{energy}\left(\mathrm{syllable}_{i.j}\right)}{\sum_{i=1}^{|\mathrm{syllable}|} \sum_{j=1}^{n_i} \mathrm{energy}\left(\mathrm{syllable}_{i.j}\right)}. \tag{8}$$

Considering the result of speech recognition, the features are normalized to be the first order features. To model the speaking rate and variation in the energy and pitch during the utterance, second-order features delta-duration, delta-energy, and delta-pitch are obtained from the forward difference of the first-order

features, as shown in (9)–(11) at the bottom of the page. In the (9)-(11), w is half the observation window size. Three two-order features are available after feature extraction. These features are combined to form a vector with $24w-6$ features as the input vector of the GMM. The acoustic features are denoted as the syllable-based observation sequence corresponding to the potential IP silence_k by

$$\left\{O=[O_D,O_P,O_E]\in R^{\dim}\right\} \tag{12}$$

where $O_s \in R^{\dim_s}$, $S \in \{D,P,E\}$ represents the vectors of single features, and *dim* denotes the number of dimensions of the feature vector that consist of features related to duration, pitch and energy, as shown in (13)–(15) at the bottom of the page.

C. Gaussian Mixture Model for Interruption Point Detection

The GMM is adopted to detect IP using the acoustic features

$$\begin{aligned} P\left(\text{Seq}_{\text{syllable}} \mid C_j\right) &\equiv P\left(O_t \mid \lambda_j\right) \\ &= \sum_{i=1}^{W} w_i N\left(O_t;\mu_i,\sum_i\right). \end{aligned} \tag{16}$$

In (16), $C_j = \{E_{ip}, \neg E_{ip}\}$ denotes the hypothesis set for silence_k containing and not containing the IP; λ_j is the GMM for class C_j, and w_i represents a mixture weight that must satisfy the constraint $\sum_{i=1}^{W} w_i = 1$, where W is the number of mixture components, and $N(\cdot)$ denotes the Gaussian density function

$$\begin{aligned} N\left(O_t;\mu_i,\sum_i\right) = &\frac{1}{(2\pi)^{\dim/2}\left|\sum_i\right|^{1/2}} \\ &\exp\left(-\frac{1}{2}\left(O_t-\mu_i\right)^T\sum_i^{-1}\left(O_t-\mu_i\right)\right) \end{aligned} \tag{17}$$

where μ_i and $\sum_i$ are the mean vector and covariance matrix of component i, and O_t denotes observation t in the training corpus. The parameters $\theta = [w_i, \mu_i, \sum_i]$, $i = 1 \ldots M$ can be estimated iteratively using the EM algorithm [31] for mixture i

$$\hat{w}_i = \frac{1}{N}\sum_{t=1}^{N} P\left(i \mid O_t,\lambda\right) \tag{18}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{N} P\left(i \mid O_t,\lambda\right) O_t}{\sum_{t=1}^{N} P\left(i \mid O_t,\lambda\right)} \tag{19}$$

$$\hat{\sum_i} = \frac{\sum_{t=1}^{N} P\left(i \mid O_t,\lambda\right)\left(O_t-\hat{\mu}_i\right)\left(O_t-\hat{\mu}_i\right)^T}{\sum_{t=1}^{N} P\left(i \mid O_t,\lambda\right)} \tag{20}$$

where $P(i \mid O_t,\lambda) = \left(P(O_t \mid \lambda)\, w_i / \sum_{j=1}^{W} P(O_t \mid \lambda)\, w_j\right)$ and N denote the total numbers of feature observations.

D. Potential Interruption Point Extraction

Based on the assumption the detected IPs that do not appear in the word boundary can be removed by assuming that an IP usually occurs at the boundary of two successive words. After eliminating unlikely IPs, the remaining IPs are kept for linguistic processing. Since the word graph or word lattice is obtained from the speech recognition module, every path in the word graph or word lattice forms its potential IP set for an input utterance.

IV. Linguistic Processing for Edit Disfluency Correction

The previous section indicates that potential IPs are identified by acoustic features. However, correcting edit disfluency using linguistic features is a key for rich transcription. This study detects the edit disfluency by maximizing the likehood of the language model for the cleaned-up utterances and the word correspondence between the deletable region and the correction given the position of the IP. The log linear model has been proven

$$\Delta nf_{\text{duration}_i} = \begin{cases} nf_{\text{duration}_{i+1}} - nf_{\text{duration}_i}, & \text{if } -w < i < w \text{ and } i \neq -1 \\ nf_{\text{duration}_{+1}} - nf_{\text{duration}_{-1}}, & i = -1 \\ 0, & \text{others} \end{cases} \tag{9}$$

$$\Delta nf_{\text{pitch}_i} = \begin{cases} nf_{\text{pitch}_{i+1}} - nf_{\text{pitch}_i}, & \text{if } -w < i < w \text{ and } i \neq -1 \\ nf_{\text{pitch}_{+1}} - nf_{\text{pitch}_{-1}}, & i = -1 \\ 0, & \text{others} \end{cases} \tag{10}$$

$$\Delta nf_{\text{energy}_i} = \begin{cases} nf_{\text{energy}_{i+1}} - nf_{\text{energy}_i}, & \text{if } -w < i < w \text{ and } i \neq -1 \\ nf_{\text{energy}_{+1}} - nf_{\text{energy}_{-1}}, & i = -1 \\ 0, & \text{others.} \end{cases} \tag{11}$$

$$O_D \equiv \begin{bmatrix} nf_{\text{duration}_{-w+1}},\ldots,nf_{\text{duration}_{-1}},nf_{\text{duration}_0},nf_{\text{duration}_{+1}},nf_{\text{duration}_{+2}},\ldots,nf_{\text{duration}_{+w}} \\ \Delta nf_{\text{duration}_{-w+1}},\ldots,\Delta nf_{\text{duration}_{-1}},\Delta nf_{\text{duration}_0},\Delta nf_{\text{duration}_{+1}},\Delta nf_{\text{duration}_{+2}},\ldots,\Delta nf_{\text{duration}_{+w-1}} \end{bmatrix}^T \tag{13}$$

$$O_P \equiv \begin{bmatrix} nf_{\text{pitch}_{-w+1}},\ldots,nf_{\text{pitch}_{-1}},nf_{\text{pitch}_0},nf_{\text{pitch}_{+1}},nf_{\text{pitch}_{+2}},\ldots,nf_{\text{pitch}_{+w}} \\ \Delta nf_{\text{pitch}_{-w+1}},\ldots,\Delta nf_{\text{pitch}_{-1}},\Delta nf_{\text{pitch}_0},\Delta nf_{\text{pitch}_{+1}},\Delta nf_{\text{pitch}_{+2}},\ldots,\Delta nf_{\text{pitch}_{+w-1}} \end{bmatrix}^T \tag{14}$$

$$O_E \equiv \begin{bmatrix} nf_{\text{energy}_{-w+1}},\ldots,nf_{\text{energy}_{-1}},nf_{\text{energy}_0},nf_{\text{energy}_{+1}},nf_{\text{energy}_{+2}},\ldots,nf_{\text{energy}_{+w}} \\ \Delta nf_{\text{energy}_{-w+1}},\ldots,\Delta nf_{\text{energy}_{-1}},\Delta nf_{\text{energy}_0},\Delta nf_{\text{energy}_{+1}},\Delta nf_{\text{energy}_{+2}},\ldots,\Delta nf_{\text{energy}_{+w-1}} \end{bmatrix}^T. \tag{15}$$

to achieve a promising performance [32] and was, therefore, adopted in this investigation. Considering a word sequence W generated by a speech recognition engine, this study finds the cleanup word string W^* using a log linear mixture model, in which the language and alignment models are both included as follows. Fig. 4 shows the proposed model. See (21) at the bottom of the page. Here, α is the combination weight for cleanup language model and alignment model; IP denotes the interruption point obtained from the IP detection module, and n represents the position of the potential IP.

A. Language Model of Cleanup Utterance

Statistical language models have previously been applied to speech recognition, and have significantly improved the recognition results. Stolcke and Shriberg first proposed speech disfluency detection using a statistical language model [33]. Speech disfluencies are represented by the probabilistic events occurring within the word stream. However, estimating the probability of word sequences can be expensive and always suffers from the problem of data sparseness. In practice, the statistical language model is frequently approximated by the class-based N-gram model with modified Kneser–Ney discounting probabilities [34] for further smoothing

$$\begin{aligned} &P\left(w_1, w_2, \ldots w_t, w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N\right) \\ &\quad = \prod_{i=1}^{t} P\left(w_i \mid \text{Class}\left(w_1^{i-1}\right)\right) P\left(w_{n+1} \mid \text{Class}\left(w_1^{t}\right)\right) \\ &\qquad \times \prod_{j=n+2}^{N} P\left(w_j \mid \text{Class}\left(w_1^{t} w_{n+1}^{j-1}\right)\right). \end{aligned} \tag{22}$$

Here, $\text{Class}(\cdot)$ is the conversion function that translates a word sequence into a word class sequence. This study employs two word classes, semantic and part-of-speech (POS) class. A semantic class, such as the synset in WordNet (http://wordnet.princeton.edu/), contains words that share a semantic property based on semantic relations, for example, hyponym and hypernym. A POS is a syntactic or grammatical category, and is defined as the role played by a word in a sentence such as noun, verb, or adjective.

The other essential issue of n-gram models for correcting edit disfluency is the number of orders in the Markov model. Since an IP is the point at which the speaker breaks off the deletable region and the correction that consists of the portion of the utterance repaired by the speaker, the cleanup utterance can be considered fluent by removing the words in the deletable region. However, removing part of the word string shortens the string, and increases the probability of a shorter word string. Consequently, short word strings are favored. To deal with this problem, we can increase the order to constrain the perplexity and normalize the word length by aligning the deletable region and the correction.

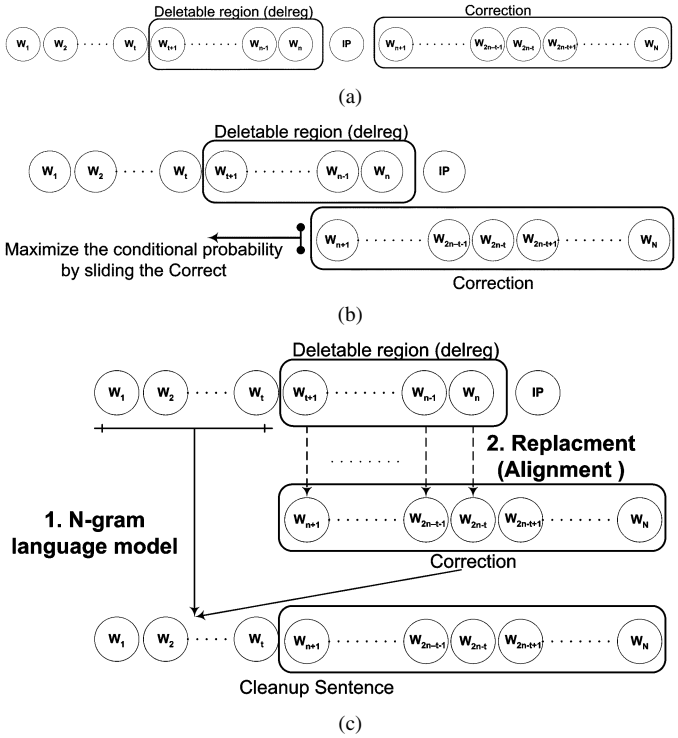

Fig. 4. (a) Structure of edit disfluencies (b) Correction process for edit disfluency. (c) Linguistic processing for edit disfluency correction.

B. Alignment Model Between Deletable Region and Correction

The structural pattern of a deletable region is generally similar to that of the correction in conversational speech. The deletable region sometimes arises as a substring of the

$$\begin{aligned} W^* &= \arg\max_{W,IP} P(W; IP) \\ &= \arg\max_{W,IP} P(w_1^N; IP) \\ &= \arg\max_{W,IP} P(w_1, w_2, \ldots w_t, w_{t+1}, \ldots w_n, w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N; IP) \\ &= \arg\max_{W,n,t} \Big(P(w_1, w_2, \ldots w_t, w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N)^{\alpha} \\ &\qquad \times P(w_{t+1}, \ldots w_n \mid w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N)^{(1-\alpha)} \Big) \\ &= \arg\max_{W,n,t} \Big(\alpha \log(P(w_1, w_2, \ldots w_t, w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N)) \\ &\qquad + (1-\alpha) \log(P(w_{t+1}, \ldots w_n \mid w_{n+1}, \ldots w_{2n-t}, w_{2n-t+1}, \ldots w_N)) \Big). \end{aligned} \tag{21}$$

correction. Accordingly, the structural pattern can be found in the starting point of the correction, which generally follows the IP. The potential IP can then be taken as the center, and the word string can be aligned before and after it. Since the correction is adopted to replace the deletable region and end the utterance, a correspondence exists between the words in the deletable region and the correction. The alignment can therefore be modeled by assuming the conditional probability of the correction given the possible deletable region. According to this observation, class-based alignment [35] is presented to clean up edit disfluency. The alignment model can be described as

$$\begin{aligned} &P\left(w_{n+1},\ldots w_{2n-t},w_{2n-t+1},\ldots w_N \mid w_{t+1},\ldots w_n\right) \\ &\quad=\prod_{k=t+1}^{n}\Bigg(P\left(f_k \mid \text{Class}\left(w_k\right)\right) \\ &\qquad\qquad\times\prod_{l=1}^{f_k}P\left(\text{Class}\left(w_l\right) \mid \text{Class}\left(w_k\right)\right)\Bigg) \\ &\quad\times\prod_{k,l,m}P\left(l \mid k,m\right) \end{aligned} \tag{23}$$

where fertility f_k denotes the number of words in the correction corresponding to the word w_k in the deletable region; k and l are positions of the w_k and w_l in the deletable region and the correction, respectively, and m denotes the number of words in the deletable region. The alignment model for cleanup comprises three components: fertility probability, translation or corresponding probability and distortion probability. The fertility probability of word w_k is defined as

$$P\left(f_k \mid \text{Class}\left(w_k\right)\right)=\frac{\sum_{w_i\in\text{Class}(w_k)}\delta\left(f_i=f_k\right)}{\sum_{p=0}^{N}\sum_{w_j\in\text{Class}(w_k)}\delta\left(f_j=p\right)} \tag{24}$$

where $\delta\left(\cdot\right)$ is an indicator function, and N denotes the maximum fertility value. The translation or corresponding probability is measured according to Wu and Palmer's definition [36]

$$\begin{aligned} &P\left(\text{Class}\left(w_l\right) \mid \text{Class}\left(w_k\right)\right) \\ &\quad=\frac{2\times\text{Depth}\left(LCS\left(\text{Class}\left(w_l\right),\text{Class}\left(w_k\right)\right)\right)}{\text{Depth}\left(\text{Class}\left(w_l\right)\right)+\text{Depth}\left(\text{Class}\left(w_k\right)\right)} \end{aligned} \tag{25}$$

where $\text{Depth}\left(\cdot\right)$ denotes the depth of the word class, and $LCN\left(\cdot\right)$ represents the lowest common subsumer of the words. The distortion probability $P\left(l \mid k,m\right)$ is the mapping probability of the word sequence between the deletable region and the correction.

V. Experimental Results and Discussion

A. Data Preparation and Speech Recognition

The Mandarin Conversational Dialogue Corpus (MCDC) [37], gathered from 2000 to 2001 at the Institute of Linguistics of Academia Sinica, Taiwan, R.O.C., comprises 30 digitized conversational dialogues numbered from 01 to 30 of a total length of 27 h. Sixty subjects living in Taiwan were randomly chosen. The annotation in [37] gives concise explanations and detailed operational definitions of each tag. Corresponding to SimpleMDE, direct repetitions, partial repetitions, overt repairs, and abandoned utterances are considered as the edit disfluency in MCDC. Spontaneous speech is usually divided into formal and informal speech. Herein, the speech sounds defined in the original acoustic syllable models are called formal speech, and the informal spontaneous speech consists of particle, paralinguistic phenomena, and other uncertainty units. To illustrate speech characteristics of MCDC, eight dialogs numbered 01, 02, 03, 05, 09, 10, 25, and 26 were randomly selected for further detailed analysis. Table I illustrates the statistics of the MCDC. The total number of syllable-like units was 147,548 of which about 20% syllable-like units were categorized as informal spontaneous speech. The informal spontaneous speech is modeled by six extra acoustic filler models, similar to the words tagged as "@reject@" in [38]. These six acoustic filler models were defined for training using the Hidden Markov Model Toolkit (HTK) [39], and the recognized results were considered in language modeling. Dialogs 01, 02, 03, and 05 were randomly selected as the testing corpus with 4,148 sub-turn-like units consisting of 7,715 syllable-like units, and the others were used as the training data.

TABLE I
Statistics of Speech Characteristics of MCDC

	Formal speech	Informal spontaneous speech		
	Syllable	Particle	Paralinguistic phenomena	Others
Syllable-like unit	116,657	10,386	12,199	8,324
Ratio	79.1%	7%	8.3%	5.6%
Total	147,548			

A transcription system for spontaneous speech with edit disfluencies in Chinese was developed to evaluate the performance of the proposed approach. A speech recognition engine using HTK was built as the syllable recognizer using eight states (three states for the Initial part, and five states for Final part in Mandarin syllable). The input speech was pre-emphasized with a coefficient of 0.97. A frame size of 32 ms (512 samples) with a frame shift of 10.625 ms (170 samples) was used. The MAT Speech Database, TCC-300 [40] and MCDC were used to train the parameters in the speech recognizer. A TCC-300 with 15 896 sentences was adopted to train the syllable models. Table II lists the recognition accuracy for fluent speech (TCC-300) and spontaneous speech (MCDC) without language model, revealing that the recognition performance of spontaneous speech degrades significantly compared to that of fluent speech.

B. Potential IP Detection

An observation window over several syllables is adopted because IP detection can be treated as a position determination problem. In this observation window, the values of pitch and energy of the syllables just before an IP are usually larger than that after the IP. This phenomenon indicates that the pitch reset and energy reset co-occur with IP in the edit disfluency. This phenomenon generally arises in the syllables of the first word immediately after the IP. The pitch reset event is very obvious when the disfluency type is repair. The energy plays the same role as the pitch when the edit disfluency appears, but its effect is not as obvious as that of the pitch. The filler words or phrase after the IP are lengthened to strive for the time for the speaker to

TABLE II
SPEECH RECOGNITION PERFORMANCE FOR FLUENT SPEECH AND SPONTANEOUS SPEECH

	Acc.	Del.	Sub.	Ins.
TCC-300 (Top 1)	89.51	0.15	9.55	0.79
TCC-300 (Top 3)	89.76	0.15	9.32	0.77
TCC-300 (Top 5)	90.38	0.15	8.82	0.65
MCDC (Top 1)	52.83	7.79	32.35	16.36
MCDC (Top 3)	53.27	7.75	32.32	16.32
MCDC (Top 5)	53.92	7.75	31.42	16.26

TABLE III
ERROR ANALYSIS OF INTERRUPTION POINT DETECTION VERSUS THE VALUES OF γ WITH/WITHOUT WORD BOUNDARY CONSTRAINTS

Metrics \ γ =	8	4	2	1	0.5	0.25	0.125
$Error_{IP}$	0.71	0.64	0.67	0.60	0.62	0.64	1.52
n_{M-IP}/n_{IP}	0.71	0.62	0.49	0.33	0.20	0.04	0.04
n_{FA-IP}/n_{IP}	0.00	0.02	0.18	0.27	0.42	0.60	1.48
$Error_{IP}^{*}$	0.71	0.64	0.59	0.49	0.40	0.33	0.82
$(n_{M-IP}/n_{IP})^{*}$	0.71	0.62	0.48	0.33	0.20	0.04	0.04
$(n_{FA-IP}/n_{IP})^{*}$	0.00	0.02	0.11	0.16	0.20	0.29	0.78

construct the correction and attract the listener's attention. This factor can significantly improve the IP detection rate.

The hypothesis testing, combined with the GMM model with four mixture components using the syllable features, determines if the silence contains the IP. The parameter γ should be determined to improve the result. The overall IP detection error rate defined in RT'04F [41] is the sum of the missed detection and false alarm rates

$$\mathrm{Error_{IP}} = \frac{n_{\mathrm{M-IP}} + n_{\mathrm{FA-IP}}}{n_{\mathrm{IP}}} \tag{26}$$

where $n_{\mathrm{M-IP}}$ and $n_{\mathrm{FA-IP}}$ denote the numbers of missed detections and false alarms, respectively, and n_{IP} represents the number of reference IPs. The threshold γ in (1) can be adjusted for $n_{\mathrm{M-IP}}$ and $n_{\mathrm{FA-IP}}$, as shown in Table III. Since the goal of the IP detection module is to detect the potential IPs, false alarms for IP detection is a less serious problem than missing detection errors. That is, the aim of this work is to achieve a high recall rate without significantly increasing the false alarm rate. Finally, the threshold γ was set to 0.25.

Since the IP always appears in a word boundary, this constraint can be adopted to remove unlikely IPs. The constraint obviously significantly reduces the number of false alarms in IP detection.

C. Clean-Up Disfluency Using Linguistic Information

Two different transcriptions were used to evaluate the edit disfluency correction model: human generated transcription (REF) and speech-to-text recognition output (STT). The reference transcriptions provide the best case for the evaluation of the edit disfluency correction module because they have no word errors. For practicability, the syllable lattice from speech recognition was fed into the edit disfluency correction module to assess the performance.

For class-based approach, POS, and semantic class were employed as the word class. Herein, semantic class was obtained based on Hownet (http://www.keenage.com/), which defines the relation "IS-A" as the primary feature. Therefore, the words can be categorized according to their hypernyms or concepts, and every word can map to its own semantic class.

TABLE IV
RESULTS (%) OF LINGUISTIC MODULE WITH EQUAL WEIGHT $\alpha = 0.2$ FOR EDIT WORD DETECTION ON REF AND STT CONDITIONS

	Human generated transcription (REF)			Speech-to-text recognition output (STT)		
	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$
1-gram[1]	0.18	0.14	0.32	0.49	0.48	0.97
1-gram[2]	0.14	0.18	0.32	0.46	0.48	0.94
2-gram[1]	0.12	0.16	0.28	0.39	0.44	0.83
2-gram[2]	0.14	0.17	0.32	0.41	0.43	0.84
3-gram[1]	0.13	0.16	0.29	0.37	0.41	0.78
3-gram[2]	0.29	0.12	0.41	0.43	0.39	0.82
DF-gram (3-gram)[1]	0.18	0.11	0.29	0.46	0.28	0.74
DF-gram (3-gram)[2]	0.12	0.14	0.26	0.35	0.34	0.69
alignment[1]	0.15	0.16	0.31	0.37	0.36	0.73
2-gram+ alignment	0.17	0.12	0.29	0.40	0.32	0.72
2-gram+ alignment[1]	0.09	0.15	0.24	0.36	0.32	0.68
2-gram+ alignment[2]	0.10	0.21	0.31	0.34	0.54	0.88
3-gram+ alignment	0.16	0.12	0.28	0.31	0.35	0.66
3-gram+ alignment[1]	0.09	0.12	0.21	0.32	0.32	0.64
3-gram+ alignment[2]	0.07	0.16	0.23	0.28	0.30	0.58

[1]: word class based on part of speech (POS) [2]: word class based on the semantic class

The edit word detection (EWD) detects input speech containing the words in the deletable regions. One of the primary metrics for edit disfluency correction is to adopt the edit word detection method defined in RT'04F [40]. This method is simply the average number of missed edit word detections and falsely detected edit words per reference IP

$$\mathrm{Error_{EWD}} = \frac{n_{\mathrm{M-EWD}} + n_{\mathrm{FA-EWD}}}{n_{\mathrm{EWD}}} \tag{27}$$

where $n_{\mathrm{M-EWD}}$ is the number of deletable edit words in the reference transcription that are not covered by the deletable regions of the system-detected edits; $n_{\mathrm{FA-EWD}}$ denotes the number of reference words that are not deletable, yet are covered by deletable regions of the system-detected edits, and n_{EWD} represents the number of deletable reference edit words. Table IV presents the result of the proposed method using linguistic information.

Due to the lack of structural information, 1-gram does not obtain any improvement, while 2-gram provides more significant improvement combined with POS class-based alignment than semantic class-based alignment. Using 3-gram and semantic class-based alignment outperforms other combinations, because 3-gram with more strict constraints can reduce the false alarm rate for edit word detection. A 4-gram model was also used to attempt to gain further improvement over 3-gram, but the excessive computation meant that the improvement was less significant than expected. Additionally, the 4-gram statistics were sparser than those of the 3-gram model. These findings indicate that the best combination in edit disfluency correction module is 3-gram and semantic class.

According to the analytical results shown in Table IV, the values of the probabilities of the n-gram model are much smaller than those of the alignment model. Since the alignment can be taken as the penalty for edit words, the effects of the 3-gram and the alignment with semantic class should be balanced by a log linear combination weight α. To optimize the performance, α

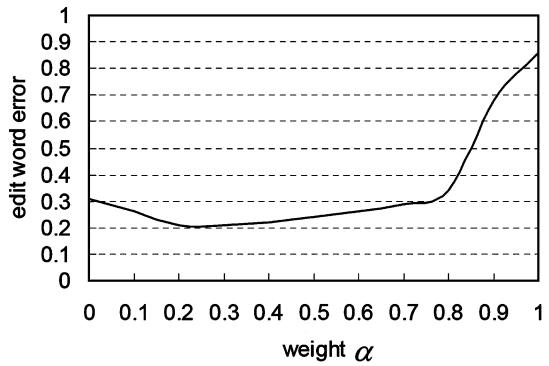

Fig. 5. Edit word error rate as a function of the values of the log linear combination weight.

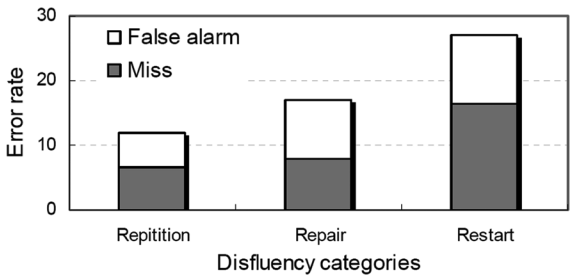

Fig. 6. Experimental results for three disfluency types.

was empirically estimated by minimizing the edit word errors. The result is shown in Fig. 5, in which the weight α was specified as 0.2 to optimize performance.

Fig. 6 illustrates the experimental results for the three disfluency types. These results indicate that the alignment model can handle repetition and revision disfluencies very well. However, the alignment model did not perform as well as expected for the restart disfluency detection, where the improvement was less pronounced than that for other edit disfluency categories. Some restart disfluencies called abandoned utterances in [42] and discontinuities in [43] do not exhibit the parallel structure between the deleteable region and the correction part. This class of restart disfluency results from the speaker abandoning a current utterance completely and starting a new utterance, and is not solvable using the proposed alignment model. Since the complexity of speech-to-text recognition output is higher than that of human-generated transcription, the performance of the STT task is significantly worse than that of the REF task, especially in recognizing acoustic syllable models. For practicability, the robustness of speech recognition should be enhanced in the future.

D. Interruption Pont Detection

Fig. 7 shows the receiver operator characteristic (ROC) curves of interruption point detection using four different criteria: maximum entropy using acoustic features, acoustic GMM models with and without word boundary constraint, and the proposed model shown in Fig. 2.

The maximum entropy model [28] was performed to compare the performance of the interruption point detection model. The acoustic features, shown in Sections III-A and B, are also adopted to train the model. Acoustic GMM models with and without word boundary constraints have been described in Section V-B. Instead of the first best hypothesis to detect the interruption point, the word lattice generated by the speech recognition engine is considered to postpone the disfluency detection decision to the language model, rather than annotating the first best hypothesis in the proposed model.

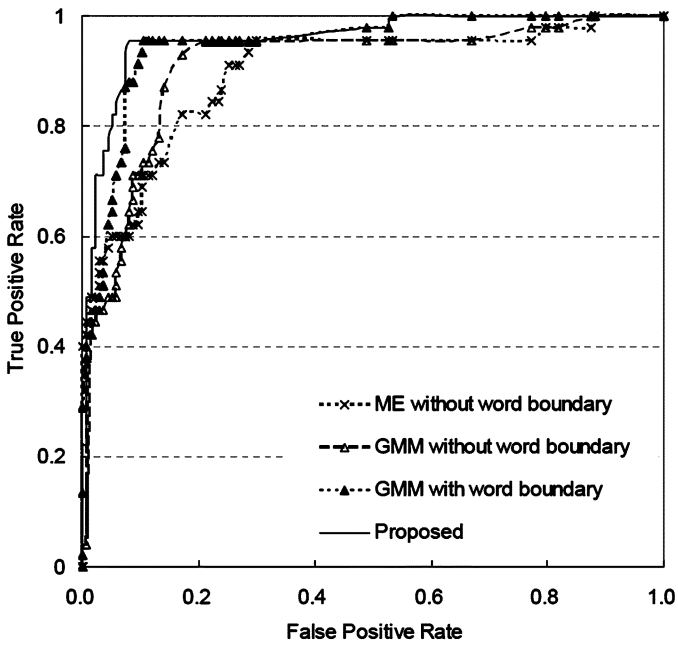

Fig. 7. ROC curves for interruption point detection by four criteria.

The results in Fig. 6 indicate a significant rise in the true positive rate of the maximum entropy model when the false positive rate is less than 0.1. In this case, the curve of the maximum entropy model intersects that of the GMM model without word boundaries when the values of true and false positive rates are 0.6 and 0.073, respectively. The maximum entropy model clearly outperforms the GMM model when the false positive rate is small. The observations indicate that the precision rate of maximum entropy is better than that of GMM model, but the recall rate is not. However, the recall rate of the maximum entropy model can clearly be refined by adding additional features. As revealed in Section V-B, the performance can be improved by introducing the word boundaries. The proposed approach achieves the best performance by integrating the language and alignment models.

VI. Conclusion and Future Work

This investigation has presented an approach to edit disfluency detection and correction for rich transcription. The proposed method, based on a two-stage process, aims to model the behavior of edit disfluency and cleanup the disfluency. An IP detection module using hypothesis testing from the acoustic features is adopted to detect the potential IPs. A word-based linguistic module comprising a cleanup language model and an alignment model is adopted to verify the position of the IP and thus correct the edit disfluency. Experimental results indicate that the IP detection mechanism can recall IPs by adjusting the threshold in hypothesis testing. To investigate the linguistic properties of edit disfluency, the linguistic module was explored to correct the disfluency based on the potential IPs. The experimental results indicate that performance was significantly enhanced. In the future, this framework will be extended to deal

with problems resulting from subwords, and improve the performance of restart disfluency to improve the transcription performance. Additionally, robust speech recognition with tolerance to versatile spontaneous speech is essential in future methods.

REFERENCES

[1] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 420–435, Jul. 2004.

[2] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynods, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. Eurospeech*, 2003, pp. 1585–1588.

[3] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue Univ., West Lafayette, IN, 2004.

[4] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D., Dept. Psychol., Univ. California, Berkeley, 1994.

[5] S. Strassel. (2004) Simple Metadata Annotation Specification Version 6.2. Linguistic Data Consortium. [Online]. Available: http://www.ldc.upenn.edu/Projects/MDE

[6] C.-H. Wu and G.-L. Yan, "Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 330–344, May 2005.

[7] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154, 2000.

[8] J. Bear, J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detecting and correction of repairs in human computer dialog," in *Proc. ACL*, 1992, pp. 56–63.

[9] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, pp. 455–472, 2005.

[10] G. Savova and J. Bachenko, "Prosodic features of four types of disfluencies," in *Proc. DiSS*, 2003, pp. 91–94.

[11] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2005, pp. 205–208.

[12] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning," in *Proc. HLT/NAACL*, 2004, pp. 137–144.

[13] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous japanese," *Speech Commun.*, vol. 47, pp. 208–219, 2005.

[14] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proc. NAACL*, 2001, pp. 118–126.

[15] M. Johnson and E. Charniak, "A tag-based noisy channel model of speech repairs," in *Proc. ACL*, 2004, pp. 33–39.

[16] M. Lease, E. Charniak, and M. Johnson, "Parsing and its applications for conversational speech," in *Proc. ICASSP*, 2005, pp. 961–964.

[17] M. Johnson, E. Charniak, and M. Lease, "An improved model for recogizing disfluencies in conversational speech," in *Proc. Rich Transcription 2004 Fall Workshop*, 2004.

[18] P. Heeman and J. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue," *Comput. Ling.*, vol. 25, pp. 527–571, 1999.

[19] P. A. Heeman, K. Loken-Kim, and J. F. Allen, "Combining the detection and correction of speech repairs," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Oct. 1996, pp. 358–361.

[20] M. Honal and T. Schultz, "Automatic disfluency removal on recognized spontaneous speech—Rapid adaptation to speaker dependent dislfuencies," in *Proc. ICASSP*, 2005, pp. 969–972.

[21] ——, "Corrections of disfluencies in spontaneous speech using a noisy-channel approach," in *Proc. Eurospeech*, 2003, pp. 2781–2784.

[22] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection," in *Proc. HLT/NAACL*, 2004, pp. 157–160.

[23] T. Hain, P. C. Woodland, G. Evermann, M. J. F. Gales, X. Liu, G. L. Moore, D. Povey, and L. Wang, "Automatic transcription of conversational telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1173–1185, Nov. 2005.

[24] H. Soltau, H. Yu, F. Metze, C. Fugen, J. Qin, and S.-C. Jou, "The 2003 ISL rich transcription system for conversational telephony speech," in *Proc. Acoust., Speech, Signal Process.*, 2004, pp. 17–21.

[25] M. Harper, B. J. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L. Yung, A. Krasnyanskaya, and R. Stewart, "Final report on parsing and spoken structural event detection," in *Proc. Johns Hopkins Summer Workshop*, 2005, pp. 1–116.

[26] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proc. 43nd Annu. Meeting Assoc. Computat. Ling.*, 2005, pp. 451–458.

[27] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, pp. 917–920.

[28] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing HMM, maximum entropy, and conditional random fields for disfluency detection," in *Proc. Eurospeech*, 2005, pp. 3313–3316.

[29] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. l. Woodland, and M. Harper, "Structural metadata research in the ears program," presented at the *ICASSP, invited paper*, 2005, pp. 957–960.

[30] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 2247–2250.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Statist Soc. B*, pp. 1–39, 1977.

[32] Y. Liu, Q. Liu, and S. Lin, "Log-linear models for word alignment," in *Proc. 43rd Annu. Meeting Assoc. Comput. Ling.*, 2005, pp. 459–466.

[33] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP*, vol. 1, 1996, pp. 405–408.

[34] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Center Res. Comput. Technol., Harvard Univ., Cambridge, MA, Tech. Rep. TR-10-98, 1998.

[35] H. Ney, S. Niessen, F. J. Och, H. Sawaf, C. Tilhnmm, and S. Vogel, "Algorithms for statistical translation of spoken language," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 24–36, Jan. 2000.

[36] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proc. 32nd ACL*, 1994, pp. 133–138.

[37] S.-C. Tseng and Y.-F. Liu, "Annotation of Mandarin Conversational Dialogue Corpus," Academia Sinica, CKIP Tech. Rep.-01, 2002.

[38] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, Oct. 2004, pp. 1961–1964.

[39] S. J. Young, G. Evermann, T. Hain, D. Kershaw, G. L. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book*. Cambridge, U.K: Cambridge Univ. Press, 2003.

[40] MAT Speech Database—TCC-300 [Online]. Available: http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm

[41] Rich Transcription (RT-04F) Evaluation Plan (2004). [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.doc

[42] C.-K. Lin, S.-C. Tseng, and L.-S. Lee, "Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous mandarin speech," in *Proc. DiSS*, 2005, pp. 117–121.

[43] M. Snover, R. Schwartz, B. Dorr, and J. Makhoul, "RT-S: Surface rich transcription scoring, methodology, and initial results," in *Proc. DARPA Rich Transcription Workshop*, 2004.

Jui-Feng Yeh received the B.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, R.O.C, in 1993 and the M.S. degree in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 1995, respectively. He is currently pursuing the Ph.D. degree in computer science and information engineering from National Cheng Kung University.

His research interests include information retrieval, spoken language processing, and ontology construction.

Chung-Hsien Wu (SM'03) received the Ph.D. degree in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, R.O.C., in 1991.

Since August 1991, he has been with the Department of Computer Science and Information Engineering, NCKU. He became a Professor in August 1997. He is currently the Editor-in-Chief for the *International Journal of Computational Linguistics and Chinese Language Processing*. His research interests include speech recognition, text-to-speech, and multimedia information retrieval.

Dr. Wu is a member of the International Speech Communication Association (ISCA) and ROCLING.