A Cross-language Study on Automatic Speech Disfluency Detection

Wen Wang
SRI International
Menlo Park, CA
wwang@speech.sri.com

Andreas Stolcke
Microsoft Research
Mountain View, CA
anstolcke@microsoft.com

Jiahong Yuan and Mark Liberman
University of Pennsylvania
Philadelphia, PA
jiahong.yuan@gmail.com
markyliberman@gmail.com

Abstract

We investigate two systems for automatic disfluency detection on English and Mandarin conversational speech data. The first system combines various lexical and prosodic features in a Conditional Random Field model for detecting edit disfluencies. The second system combines acoustic and language model scores for detecting filled pauses through constrained speech recognition. We compare the contributions of different knowledge sources to detection performance between these two languages.

1 Introduction

Speech disfluencies are common phenomena in spontaneous speech. They consist of spoken words and phrases that represent self-correction, hesitation, and floor-grabbing behaviors, but do not add semantic information; removing them yields the intended, fluent utterance. The presence of disfluencies in conversational speech data can cause problems for both downstream processing (parsing and other natural language processing tasks) and human readability of speech transcripts. There has been much research effort on automatic disfluency detection in recent years (Shriberg and Stolcke, 1997; Snover et al., 2004; Liu et al., 2006; Lin and Lee, 2009; Schuler et al., 2010; Georgila et al., 2010; Zwarts and Johnson, 2011), particularly from the DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) MDE (MetaData Extraction) (DARPA Information Processing Technology Office, 2003) program, which focused on the automatic transcription of sizable amounts of speech data and rendering such transcripts in readable form, for both conversational telephone speech (CTS) and broadcast news (BN).

However, the EARS MDE effort was focused on English only, and there hasn't been much research on the effectiveness of similar automatic disfluency detection approaches for multiple languages. This paper presents three main innovations. First, we extend the EARS MDE-style disfluency detection approach combining lexical and prosodic features using a Conditional Random Field (CRF) model, which was employed for detecting disfluency on English conversational speech data (Liu et al., 2005), to Mandarin conversational speech, as presented in Section 2. Second, we implement an automatic filled pause detection approach through constrained speech recognition, as presented in Section 3. Third, for both disfluency detection systems, we compare side-by-side contributions of different knowledge sources to detection performance for two languages, English and Mandarin, as presented in Section 4. Conclusions appear in Section 5.

2 EARS MDE Style Automatic Disfluency Detection

We focus on two types of disfluencies, *Fillers* and *Edit disfluencies*, following the EARS MDE disfluency types modeled in (Liu et al., 2006). Fillers include filled pauses (FP), discourse markers (DM), and explicit editing terms (ET). FPs are words used by the speakers as floor holders to maintain control of a conversation. They can also indicate hesitations of the speaker. In this work, English FPs

comprise *uh* and *um*, based on English CTS corpora. For Mandarin, Zhao and Jurafsky found that Mandarin speakers intensively used both demonstratives *zhege (literally 'this')* and *nage (literally 'that')* and *uh/mm* as FPs based on a large speech corpus of Mandarin telephone conversation (Zhao and Jurafsky, 2005). We study the same set of Chinese FPs in this study. DMs are words or phrases related to the structure of the discourse and help taking or keeping a turn, or serving as acknowledgment, for example, *I mean*, *you know*. An explicit ET is an editing term in an edit disfluency that is not an FP or a DM. For example, *we have two action items* $*$ *sorry three action items from the meeting*, where *sorry* is an explicit ET.

Edit disfluencies involve syntactically relevant content that is either repeated, revised, or abandoned. The basic pattern for edit disfluencies has the form **(reparandum)** $*$ <**editing term**> **correction**. The reparandum is the portion of the utterance that is corrected or abandoned entirely (in the case of restarts). An interruption point (IP), marked with '$*$' in the pattern, is the point at which the speaker breaks off the original utterance and then repeats, revises, or restarts the utterance. The editing term is optional and consists of one or more filler words. The correction is the portion of the utterance that corrects the original reparandum. Revisions denote the cases when a speaker modifies the original utterance with a similar syntactic structure, e.g., *we have two action items* $*$ *sorry three action items from the meeting*. Restarts denote the cases when a speaker abandons an utterance or a constituent and restarts all over again, e.g., *He* $*$ *I like this idea*.

We used a CRF model to combine lexical features, shallow syntactic features, and prosodic features for joint detection of edit words and IP words. A CRF defines a global log-linear distribution of the state (or label) sequence E conditioned on an observation sequence, in our case including the word sequence W and the features F, and optimized globally over the entire sequence considering the context event information for making decisions at each point. We used the Mallet package (McCallum, 2002) to implement the CRF model. We used a first-order model that includes only two sequential events in the feature set. The CRF model is trained to maximize the conditional log-likelihood of a given training set $P(E|W, F)$. During testing, the most likely sequence E is found using the Viterbi algorithm. To avoid over-fitting, a zero-mean Gaussian prior (McCallum and Li, 2003) was applied to the parameters, where the variance of the prior was optimized on the development test set. Each word is associated with a class label, representing whether it is an edit word or not. We included IP in the target classes and used five states, as *outside edit* (O), *begin edit with an IP* (B-E+IP), *begin edit* (B-E), *inside edit with an IP* (I-E+IP), and *inside edit* (I-E) (Liu et al., 2006). State transitions are also the same as in (Liu et al., 2006). We built a Hidden Markov Model (HMM) based part-of-speech (POS) taggers for English conversational speech and Mandarin broadcast conversation data. After employing the co-training approach described in (Wang et al., 2007), we achieved 94% POS tagging accuracy for both data sets. The features for CRF modeling include: n-grams from words and automatically generated POS tags, speaker turns, whether there is a repeated word sequence ending at a word boundary, whether a word is a fragment, whether there is a predefined filler phrase after the word boundary, and the prosody model posterior probabilities from a decision tree model (Shriberg and Stolcke, 1997) and discretized by cumulative binning (Liu et al., 2006). The prosodic features were computed for each interword boundary from words and phonetic alignments of the manual transcriptions. We extracted the same set of prosodic features for English and Mandarin data, based on duration, fundamental frequency (f0), energy, and pause information, and nonprosodic information such as speaker gender and speaker change, for training and applying the decision-tree-based prosody model (Liu et al., 2006).

We implemented a rule-based system for filler word detection. We defined a list of possible Chinese and English filler words, including filled pauses and discourse markers. The rules also explore POS tags assigned by our Chinese and English POS taggers.

3 Constrained Speech Recognition for Filled Pause Detection

We also propose an alternative approach for automatic detection of FPs given speech transcripts that omit FPs but are otherwise accurate. This approach is motivated by situations where only an edited, "cleaned-up" transcript is available, but where an accurate verbatim transcript is to be recovered automatically. We treat this task as a constrained speech recognition problem, and investigate how effectively it is solved by a state-of-the-art large vocabulary continuous speech recognition (LVCSR) system. Hence, this approach can be considered as combining LVCSR acoustic model (AM) and language model (LM) knowledge sources in a search framework for FP detection. Compared to the FP detection component in the disfluency detection systems described in Section 2, this alternative approach explores different knowledge sources. In particular, the AMs explore different front-end features compared to the lexical and prosodic features explored in those disfluency detection systems presented in Section 2. Details of the front-end features are illustrated below.

We evaluated this approach on both English and Mandarin conversational speech. For detecting FPs in English conversational speech, we used a modified and simplified form of the recognition system developed for the 2004 NIST Rich Transcription Conversational Telephone Speech (CTS) evaluations, described in (Stolcke et al., 2006). The first pass of the recognizer uses a within-word MFCC+MLP model (i.e, trained on Mel-frequency cepstral coefficient (MFCC) features augmented with Multi-Layer Perceptron (MLP) based phone-posterior features), while the second pass uses a cross-word model trained on Perceptual Linear Prediction (PLP) features adapted (by speaker) to the output of the first pass. For purposes of FP detection, the recognition is constrained to a word lattice formed by the manually transcribed non-FP reference words, with optional FP words inserted between any two words and at the beginning and end of each utterance. Both first and second pass decoding was constrained by the optional-FP lattices. In the second pass, HTK lattices were generated with bigram LM probabilities and rescored with a 4-gram LM. The consensus decoding output from the rescored lattices was used for scoring FP detection. The system thus evaluates the posterior probability of an FP at every word boundary using both acoustic model (AM) and language model (LM) evidence. The acoustic model for the English recognition system was trained on about 2300 hours of CTS data. The language models (which models FP like any other word) are bigram and 4-gram statistical word n-gram LMs estimated from the same data plus additional non-CTS data and web data.

For detecting FPs in Mandarin broadcast conversation speech, we used a modified form of the recognition system developed for the 2008 DARPA GALE (Global Autonomous Language Exploitation) Speech-to-Text evaluation, described in (Lei et al., 2009). The system conducted a constrained decoding on the optional-FP lattices, using a speaker-independent within-word triphone MPE-trained MFCC+pitch+MLP model and a pruned trigram LM. For the Mandarin ASR system, the MFCC+MLP front-end features were augmented with 3-dimension smoothed pitch features (Lei et al., 2006). HTK lattices were generated with probabilities from the pruned trigram LM and rescored by the full trigram LM. The consensus decoding output from the rescored lattices was used for scoring FP detection. The AMs for this system were trained on 1642 hours of Mandarin broadcast news and conversation speech data and the LMs were trained on 1.4 billion words comprising a variety of resources. Details of training data and system development were illustrated in (Lei et al., 2009).

This procedure is similar to forced aligning the word lattices to the audio data (Finke and Waibel, 1997). Both Finke et al.'s approach (Finke and Waibel, 1997) and our approach built a lattice from each transcription sentence (in our approach, optional filled pauses are inserted between any two words and at the beginning and end of each utterance). Then Finke et al. force-aligned the lattice with utterance; whereas, we used multi-pass constrained decoding with within-word and cross-word models, MLLR adaptation of the acoustic models, and rescoring with a higher-order n-gram LM, so the performance will be better than just flexible alignment to the lattices. Note that when constructing the word lattices with optional FP words, for En-

glish, the optional FP words are a choice between *uh* and *um*. For Mandarin, the optional FP words are a choice between *uh*, *mm*, *zhege*, and *nage*. We assigned equal weights to FP words.

4 Experimental Results

Scoring of EARS MDE-style automatic disfluency detection output is done using the NIST tools [1], computing the error rate as the average number of misclassified words per reference event word. For English, the training and evaluation data were from the 40 hours CTS data in the NIST RT-04F MDE training data including speech, their transcriptions and disfluency annotations by LDC. We randomly held out two 3-hour subsets from this training data set for evaluation and parameter tuning respectively, and used the remaining data for training. Note that for Mandarin, there is no LDC released Mandarin MDE training data. We adapted the English MDE annotation guidelines for Mandarin and manually annotated the manual transcripts of 92 Mandarin broadcast conversation (BC) shows released by LDC under the DARPA GALE program, for edit disfluencies and filler words. We randomly held out two 3-hour subsets from the 92 shows for evaluation and parameter tuning respectively, and manually corrected disfluency annotation errors on the evaluation set.

Table 1 shows the results in NIST error rate (%) for edit word, IP, and filler word detection. We observe that adding POS features improves edit word, edit IP, and filler word detection for both languages, and adding a prosody model produced further improvement (note that filler word detection systems did not employ prosodic features). The gains from combining the word, POS, and prosody model over the word n-gram baseline are statistically significant for both languages (confidence level $p < 0.05$ using matched pair test). Also, adding the prosody model over word+POS yielded a larger relative gain in edit word+IP detection performance for Mandarin than for English data. A preliminary study of these results has shown that the prosody model contributes differently for different types of disfluencies for English and Mandarin conversational speech and we will continue this study in future work. We also plan to investigate the prosodic features considering the special characteristics of edited disfluencies in Mandarin studied in (Lin and Lee, 2009).

[1] www.itl.nist.gov/iad/mig/tests/rt/2004-fall/index.html

Table 1: NIST error rate (%) for edit word, IP, and filler word detection on the English and Mandarin test set, using word n-gram features, POS n-gram features, and prosody model.

Feature	**NIST Error Rate (%)**		
	Edit Word	Edit IP	Filler Word
	English		
Word	53.0	38.7	31.2
+POS	52.6	38.2	29.8
++Prosody	52.3	38.0	29.8
	Mandarin		
Word	58.5	42.8	33.4
+POS	57.7	42.1	32.9
++Prosody	56.9	41.5	32.9

For evaluating constrained speech recognition for FP detection, the English test set of conversational speech data and word transcripts is derived from the CTS subset of the NIST 2002 Rich Transcription evaluation. The waveforms were segmented according to utterance boundaries given by the human-generated transcripts, resulting in 6554 utterance segments with a total duration of 6.8 hours. We then excluded turns that have fewer than five tokens or have two or more FPs in a row (such as 'uh um' and 'uh, uh'), resulting in 3359 segments. This yields the test set from which we computed English FP detection scores. The transcripts of this test set contain 54511 non-FP words and 1394 FPs, transcribed as either *uh* or *um*. When evaluating FP detection performance, these two orthographical forms were mapped to a single token type, so recognizing one form as the other is not penalized. The Mandarin test set is the DARPA GALE 2008 Mandarin speech-to-text development test set of 1 hour duration. The transcripts of this test set contain 9820 non-FP words and 370 FP words, transcribed as *uh*, *mm*, *zhege*, and *nage*. We collapsed them to a single token type for FP scoring. We evaluated FP detection performance in terms of both false alarm (incorrect detection) and miss (failed detection) rates, shown in Table 2. We observed that adding pronunciation scores didn't change the P_{fa} and P_{miss}. On the English

test set, adding LM scores degraded P_{miss} but improved P_{fa}. However, on the Mandarin test set, increasing LM weight improved both P_{miss} and P_{fa}, suggesting that for the Mandarin LVCSR system in this study, the LM could provide complementary information to the AM to discriminate FP and non-FP words.

Table 2: Probabilities of false alarms (FAs) and misses in FP detection on the English and Mandarin test set w.r.t. acoustic model weight w_a, language model weight w_g, and pronunciation score weight w_p.

$\{w_a, w_g, w_p\}$	**FAs (%)**	**Misses (%)**
	English	
{1,0,8}	1.76	3.23
{1,8,8}	1.18	4.73
	Mandarin	
{1,0,8}	1.19	19.68
{1,8,8}	0.76	16.76

5 Conclusion

In conclusion, we have presented two automatic disfluency detection systems, one combining various lexical and prosodic features, and the other combining LVCSR acoustic and language model knowledge sources. We observed significant improvements in combining lexical and prosodic features over just employing word n-gram features, for both languages. When combining AM and LM knowledge sources for FP detection in constrained speech recognition, we found increasing LM weight improved both false alarm and miss rates for Mandarin but degraded the miss rate for English.

Acknowledgments

The authors thank all the anonymous reviewers of this paper for valuable suggestions. This work is supported in part by NSF grant IIS-0964556.

References

DARPA Information Processing Technology Office. 2003. Effective,affordable, reusable speech-to-text (EARS). http://www.darpa.mil/ipto/programs/ears.

Michael Finke and Alex Waibel. 1997. Flexible transcription alignment. In *IEEE Workshop on Speech Recognition and Understanding*, pages 34–40.

K. Georgila, N. Wang, and J. Gratch. 2010. Cross-domain speech disfluency detection. In *Proceedings of SIGDIAL*, pages 237–240, Tokyo.

X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf, and T. Lee. 2006. Improved tone modeling for Mandarin broadcast news speech recognition. In *Proceedings of Interspeech*.

X. Lei, W. Wu, W. Wang, A. Mandal, and A. Stolcke. 2009. Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversation. In *Proceedings of Interspeech*, Brighton, UK.

C. K. Lin and L. S. Lee. 2009. Improved features and models for detecting edit disfluencies in transcribing spontaneous mandarin speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1263–1278, September.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, and Mary Harper. 2005. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proc. Interspeech*, pages 3313–3316, Lisbon, September.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540, September. Special Issue on Progress in Rich Transcription.

A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields. In *Proceedings of the CoNLL*.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

W. Schuler, S. AbdelRahman, T. Miller, and L. Schwartz. 2010. Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, 36(1).

E. Shriberg and A. Stolcke. 1997. A prosody-only decision-tree model for disfluency detection. In *Proceedings of Eurospeech*, pages 2383–2386.

M. Snover, B. Dorr, and R. Schwartz. 2004. A lexically-driven algorithm for disfluency detection. In Susan Dumais, Daniel Marcu, and Salim Roukos, editors, *Proc. HLT-NAACL*, Boston, May. Association for Computational Linguistics.

Andreas Stolcke, Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lin, Tim Ng, Mari Ostendorf, Kemal Sönmez, Anand Venkataraman, Dimitra Vergyri, Wen Wang, Jing Zheng, and Qifeng Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Trans. Audio, Speech, and Lang. Pro-*

cess., 14(5):1729–1744, September. Special Issue on Progress in Rich Transcription.

W. Wang, Z. Huang, and M. P. Harper. 2007. Semi-supervised learning for part-of-speech tagging of Mandarin transcribed speech. In *Proceedings of ICASSP*, pages 137–140.

Y. Zhao and D. Jurafsky. 2005. A preliminary study of mandarin filled pause. In *Proceedings of DISS*, pages 179–182, Aix-en-Provence.

S. Zwarts and M. Johnson. 2011. The impact of language models and loss functions on repair disfluency detection. In *Proceedings of ACL/HLT*, pages 703–711, Portland.