14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

# ANALYSIS OF DISFLUENT REPETITIONS IN SPONTANEOUS SPEECH RECOGNITION

*Vivek Rangarajan and Shrikanth Narayanan*

Speech Analysis and Interpretation Laboratory
University of Southern California, Viterbi School of Engineering
http://sail.usc.edu
vrangara,shri@sipi.usc.edu

## ABSTRACT

In this paper, we investigate the effect of disfluent repetitions in spontaneous speech recognition. We characterize the repetition errors in an automatic speech recognition framework using repetition word error rate (RWER). The problem is addressed by both building classifiers based on acoustic-prosodic features and a multiword model for modeling repetitions. We also analyze the repetition word error rate for different acoustic and language models in the Fisher conversational speech corpus. The classifier approach is not promising on recognizer output and generates a high degree of false alarms. The multiword approach to modeling the most frequent function word repetitions results in an absolute RWER reduction of 1.26% and a significant absolute WER reduction of 2.0% on already well trained acoustic and language models. This corresponds to a relative RWER improvement of 75.9%.

## 1. INTRODUCTION

Spontaneous speech is not well structured, acoustically and syntactically, as read speech. The presence of disfluencies makes the spontaneous speech disparate and provides a challenge for speech processing. State-of-the-art automatic speech recognition has achieved high recognition accuracy for read speech. However, the accuracy is still poor for spontaneous speech with disfluencies. The growing demand for reliable spontaneous speech recognizers has been manifested in applications such as dialogue systems, spoken document retrieval, call managers and automatic transcription of lectures and meetings. The disfluent speech portions in these tasks alter the smooth speaking style and therefore degrade the performance of the speech recognizer.

Disfluencies generally include filled pauses (uh, uhm, er etc.), repetitions, revisions, restarts and fragments. Typically, disfluencies can be broken down into a reparandum, interruption point (IP), editing phase and repair region [1]. Here we are interested in one specific type of disfluency, namely repetition. Repetitions in spontaneous speech in most cases involve a first instance of the repeated word ($\mathbf{R1}$), a possible silent pause (SIL), a second instance of the repeated word ($\mathbf{R2}$), and continuation of the utterance. An example is given below :

- I might ($\mathbf{R1}$) might ($\mathbf{R2}$) have to go to the cla- class.
- I might ($\mathbf{R1}$) SIL might ($\mathbf{R2}$) have to go to the cla- class.

Spoken language often also gives rise to disfluent repetitions that are a series of backchannel responses such as yeah yeah, right right. The studies and experiments conducted in this paper include these instances.

In the past, detection of repetitions in disfluent speech has been addressed from the perspective of acoustic [2, 3, 4], prosodic [5, 6] and language [5, 7, 8] models. Parsing techniques [9] have also been used to detect and correct these disfluencies at either the ASR n-best list or the lattice level by applying parsers, trained on disfluency-tagged treebank.

The acoustic modeling approaches have treated the disfluency as a general recognition model. With sufficient training instances, they attempt to capture the variations in duration and pause exhibited by the disfluencies. Acoustic analysis of disfluent repetitions was first presented in [3]. It suggests that repetitions are either prospective or retrospective depending on where the pause occurs in the repetition. [4] categorizes disfluent repetitions into canonical repetitions, covert self-repairs and stalling repetitions based on prosodic features. Acoustic-prosodic features used in building classifiers for detecting disfluencies are based on these trends observed in the signal.

Prosody based disfluency detection typically works with word boundaries, extracting the acoustic-prosodic features at each word boundary and building classifiers for each type. Decision tree classifiers are used with discrete features like duration, distance from pause and normalized f0 values [6]. Assuming that the ASR can give reliable end point estimates even if the hypothesized word sequence is erroneous, one can apply these classifiers at each word boundary to predict the disfluency type. However, most ASR systems do not provide accurate segmentations due to WER.

Language modeling approaches have been employed for detecting and correcting disfluencies in speech [7] primarily for natural language understanding. These approaches tag the training data with disfluency tags and during decoding predict the tags for the hypothesized word sequence. However, for ASR output these approaches are heavily dependent on the word error rate (WER). Also, the inherent nature of spontaneous speech is such that a speaker can introduce a repetition at any point in the conversation and a language model (LM) cannot completely characterize the repetition by itself. Hidden event modeling incorporating prosody in language models was presented in [5]. Recently, finite state grammar (FSG) based methods [8] have been proposed to detect and correct repetitions. These FSGs are typically formed from the training data and are constrained in that they are domain dependent and may not necessarily port well for a different application.

Typically, disfluencies have been detected and corrected as a post-processing step after first pass recognition or for isolated utterances with time alignment information. Our goal in this paper is to investigate the repetition phenomena in

spontaneous speech recognition and model them within the recognition framework. In this regard we characterize the accuracy of a spontaneous speech recognizer in terms of the repetition word error rate for different configurations of the ASR.

$$RWER = 100.0 * \frac{\mathbf{R1}^{(sub+ins+del)} + \mathbf{R2}^{(sub+ins+del)}}{\mathbf{total}^{(correct+sub+del)}} \quad (1)$$

where $\mathbf{R1}$ and $\mathbf{R2}$ denote the first and the second instance of the repetition in a repetition pair.

The motivation for doing so is two-fold. First, with reliable identification of repetitions we hope to prevent ASR performance degradation. This also reduces the need to perform additional post-processing of the ASR output to detect and correct errors caused due to repetition. Second, repetitions serve as an indicator of interruption point and help a speaker to think and maintain his turn in the conversation. They are also correlated with revisions and false starts and hence can be used in providing additional information about discourse structure.

The paper is organized as follows. In section 2, we present repetition statistics on popular conversational speech corpora. In section 3, we describe acoustic-prosodic features we have investigated for detection of repetition as a classification task. We describe the data used in this paper in section 4 followed by description of the ASR built for the task of analyzing repetitions in spontaneous speech. Section 6 explains our proposed multiword training for repetitions and reports the results. We end with conclusions and future work in section 7.

## 2. REPETITION STATISTICS ON SPONTANEOUS SPEECH CORPORA

Repetitions are common in spontaneous speech. Repetitions can be single word repetitions, where one word is repeated after another with a possible silent pause in between or multi word repetitions[1] where two or more words are repeated. Table 1 shows repetition statistics on commonly available spontaneous speech corpora. As the statistics suggest, repetitions are frequent in spontaneous speech and hence need attention.

| Corpus | Single word | Multi word | # words |
|---|---|---|---|
| Switchboard | 2.02 | 0.40 | 3.1 million |
| Call Home | 1.05 | 0.22 | 181 K |
| Fisher | 1.72 | 0.33 | 17.8 million |

**Table 1:** Repetition statistics in spontaneous speech corpora (%)

## 3. DATA

### 3.1 Data Description

The Fisher English Training data represents the conversational telephone speech (CTS) created at the Linguistic Data Consortium (LDC). The data used in this paper are from the second part of the collection designated as Fisher part 2. It contains speech data for 5849 complete conversations, each lasting upto 10 minutes. The transcription was performed by both LDC, BBN and WordWave. The cut times for the utterances are partially hand labeled and others are as a result of a first-pass speech recognizer for detecting sentence boundaries.

### 3.2 Training and Test Corpora

A total of 220 hours of data was selected as training material for the ASR system from the Fisher part 2 corpus. The 10 minute conversations were segmented into individual utterances based on the cut times provided by LDC. The results reported in this paper are based on acoustic models constructed using this training material. The test corpus consists of 2 hours of data from 20 speakers not seen in the training data. The percentage of repetitions in the test set is 1.91%. The language models were constructed from the training data transcripts as well as the transcripts from the Fisher part 1 collection and other conversational speech corpora .

## 4. ACOUSTIC-PROSODIC FEATURE ANALYSIS OF REPETITIONS

We first address the problem of detection of repetitions using acoustic-prosodic features in a classification framework. The resultant classifier is later applied to the output of the ASR to characterize the repetitions. Using the acoustical analysis of repetitions performed in [3, 4] as motivation, we extracted acoustic-prosodic features from the boundary of repetitions that quantify the inferred properties. We extracted duration, f0 and pause information across the repetition boundary. Specifically, the features considered were the pause duration after the repetition, the stylized f0 values around the boundary (f0 onset, offset, range, mean) and duration information (rhyme duration, duration of voiced regions before the boundary, duration of the vowels in $\mathbf{R1}$). The duration features were normalized by overall phone durations and the f0 values by speaker specific mean f0. We also experimented with raw values but the normalization provided better classification accuracy. The creakiness in the voice during a repetition can be characterized quantitatively by the Open Quotient (OQ) measure [8]. We use the definition of OQ as the difference in amplitude of the first and second harmonics of the spectral envelope.

We trained Gaussian Mixture Models (GMMs) for repetitions ($\lambda_{rep}$) and non-repetitions ($\lambda_{non-rep}$) based on the acoustic-prosodic features extracted at each boundary in the training data. The classifier thus trained is used in the two-way classification problem.

$$\text{boundary} = \begin{cases} \text{rep} & \text{if } p(X/\lambda_{rep}) > p(X/\lambda_{non-rep}) \\ \text{non-rep} & else \end{cases} \quad (2)$$

where $X$ denotes the acoustic-prosodic feature vector.

The classifier was trained with 30000 tokens from the training data after performing forced alignments. The forced alignments were performed using the transcriptions provided and the acoustic-prosodic features were extracted from the resultant phone-level alignments. The test data consisted of 1000 tokens and both the training and test set were downsampled to have equal priors. The classifier was also applied to the hypothesized word boundaries from the recognizer[2] out-

[1] Note that we refer to multi word repetitions as a group of two or three words repeated and not the same word repeated more than once

[2] Results reported here are for the SAT+m1 configuration of the recognizer specified in Table 4

put (also downsampled). The results of the classification is presented in Table 2.

| | Accuracy | Recall | False Alarms |
|---|---|---|---|
| Forced alignment | 75.9 | 74.0 | 25.97 |
| Recognizer ouput | 67.0 | 64.3 | 34.2 |

**Table 2:** Classification results of the acoustic-prosodic classifier (%)

The classification results for repetitions is similar to results presented in [6] though a direct comparison is not possible due to the different corpus used in our experiments. The results also indicate that the classifier based approach performs better when accurate time alignment of the words is known. Typically, this is not possible in a recognition scenario due to insertion, deletion and substitution errors. The false alarm rate is also reasonably high. In the next section we address an alternative view of disfluent repetitions, from an ASR perspective and also present a multiword modeling approach aimed at reducing both the WER and RWER.

## 5. ASR SYSTEM OVERVIEW

### 5.1 Dictionaries and Vocabularies

The Mississippi State switchboard dictionary with 38910 entries was augmented with the CMU dictionary that contains over 125000 words to form the base dictionary. Pronunciations for partial words were automatically derived from the baseform dictionary. Hypotheses for words not found in the dictionary during alignment were generated using CART based letter-to-sound rules [10] with the CART system trained on the base dictionary.

### 5.2 Language Model

Since language modeling data for conversational speech is sparse, we constructed the LM from all available data sources. The main sources used were the Fisher Training data transcripts, switchboard data transcripts and the HUB4 transcripts. The data from different sources were normalized using identical processes. Normalized spelling and uniform hyphenation was ensured across all corpora. Trigram LMs were trained using the SRI LM toolkit [11] with Kneser-Ney discounting and backoff. The interpolation weights were determined by minimizing the perplexity on held-out data from the fisher corpus. Table 3 shows the perplexity results on the test data chosen for the task. The WER reported here is for the baseline acoustic model.

| LM Corpora | Perplexity | WER | RWER |
|---|---|---|---|
| SWB + HUB4 (LM1) | 103.235 | 56.0 | 2.20 |
| Fisher Part 2 | 73.09 | 49.8 | 1.97 |
| Fisher Part 2 + LM1 | 72.60 | 49.5 | 1.92 |
| Fisher Part 1 & 2 (LM2) | 70.11 | 48.4 | 1.95 |
| LM2 + LM1 (LM3) | 69.85 | 48.0 | 1.90 |

**Table 3:** Repetition WER for different language models (%)

### 5.3 Acoustic Model and Adaptation

Standard acoustic models that are phonetic decision tree state clustered triphone models with left-to-right 3-state topology were trained using SONIC [12] speech recognition toolkit. The data at the leaves of the decision tree were modeled with Gaussian distributions via a BIC-based procedure and trained using multiple iterations of the EM algorithm. This is our baseline model. The acoustic models are further improved by performing cepstral variance normalization (CVN) and vocal tract length normalization (VTLN).

In addition to speaker independent acoustic models, we also built speaker adaptive models (SAT). The training was done via constrained maximum likelihood regression (CMLLR) transform on the feature space for each training speaker. The transform is applied to both means and variances of the system parameters. Once all speakers were transformed, we computed a new canonical acoustic model. In Table 4, (m1) refers to the acoustic model obtained after CVN and VTLN and the language model trained from Switchboard, HUB4 and Fisher transcripts. The SAT + (m1) model refers to speaker adaptation performed over cepstral variance and vocal tract length normalized acoustic model.

| Model | WER | RWER |
|---|---|---|
| Baseline + LM3 | 48.0 | 1.90 |
| CVN + VTLN + LM3(m1) | 46.0 | 1.80 |
| SAT + LM3 | 42.8 | 1.67 |
| SAT + (m1) | 42.1 | 1.66 |

**Table 4:** WER and RWER on test data for different acoustic models (%)

## 6. MULTIWORD TRAINING FOR MODELING REPETITIONS

We have shown the effect of repetitions on the accuracy of ASR systems in the previous sections. As results indicate, there is still room for improvement. In this section, we propose a multiword (also referred to as compound words) approach for modeling repetitions in spontaneous speech. A data driven approach to choosing multiwords for minimizing the WER has been presented in [13] though not from the perspective of modeling repetitions. Results in [13] also indicate that manual design of multiwords yield WER comparable to automatic selection. In this paper, we use a different approach by using a threshold for the raw count of the repetition pairs and by classifying them into function and content word categories.

Analysis of the repetitions in the training data indicate that most repetitions are function words[3] and content words form very small percentage of the total single word repetitions. Table 5 shows the distribution of repetitions among function and content words. Another motivation is that function words are on an average shorter in length (in terms of number of phones) and hence a function word repetition pair is much likely to be confused with another word during ASR decoding. We found that modeling all repetitions (function or content) as multiwords did not improve the performance, possibly due to the increase in acoustic confusability. This augurs well for most speech recognition tasks where the function word vocabulary is constant while the content words can vary with the task and may also be out of vocabulary (OOV).

In an attempt to reduce the RWER, we used a multiword training for the most frequent function words (we considered

[3]We categorize words as function or content based on part-of-speech tags obtained from a POS tagger

| Corpora | function word reps | content word reps |
|---|---|---|
| Switchboard | 91.0 | 9.0 |
| Fisher Part 2 | 91.9 | 8.1 |

**Table 5:** Function and content word repetitions in spontaneous speech corpora (%)

function word repetitions that occurred more than 250 times in the training data) in the corpus. The repetitions were included as entries in the dictionary as concatenated baseforms. We found that coarticulation is not common among repetitions and hence refrained from using coarticulated pronunciation variants. However, using the acoustic analysis presented in [3] as motivation, silent pauses (SIL) were included between the repetitions as alternate forms. This is illustrated in Table 6.

| Repetition pair | Dictionary entry |
|---|---|
| YEAH_YEAH (1) | Y AE Y AE |
| YEAH_YEAH (2) | Y AE SIL Y AE |

**Table 6:** Repetition entries in the pronunciation dictionary

Finally, we retrained acoustic models with the new multiword repetition dictionary. To incorporate the multiword repetitions in the LM, we replaced the appropriate bigrams and trigrams with corresponding multiwords and otherwise used the same LM training procedure as before. To verify the hypothesis that the multiword modeling approach does target the repetition errors, we examined the ratio of repetition errors to number of words in the repetitions. The results are reported in Table 7.

| Model | WER | RWER | $\frac{\text{R1-R2 errors}}{\text{total words in repetitions}}$ |
|---|---|---|---|
| SAT + (m1) ($*$) | 42.1 | 1.66 | 41.04 |
| ($*$) + multiword model | 40.1 | 0.40 | 20.02 |

**Table 7:** WER and RWER on test data for multiword approach (%)

The results in Table 6 indicate that the multiword approach to modeling repetitions in spontaneous speech is beneficial. The improvement in the WER and RWER is on the best trained acoustic and language models. Tabe 8 shows the performance on normal and back-channel repetitions. In the experiments, only 'yeah yeah' and 'right right' spoken in isolation were considered as back-channel repetition pairs and all the other repetitions constitute normal responses.

| Repetition Type | RWER | |
|---|---|---|
| | SAT + (m1) | SAT + (m1) + multiword |
| Normal Responses | 1.33 | 0.34 |
| Back-channel | 0.27 | 0.06 |

**Table 8:** RWER (%) of normal and back-channel responses in the test set

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an investigation into the nature of disfluent repetitions and their impact on spontaneous speech recognition. In particular we addressed the problem from the point of reducing RWER and hence the absolute WER, in ASR systems by employing different training procedures for language and acoustic models. We have also addressed the problem of detecting repetitions from acoustic-prosodic features by building GMM classifiers. An elaborate analysis of the RWER for various configurations of the acoustic and language model was presented. A multiword training procedure taking into account the most frequent function word repetitions in the training data provided an absolute RWER reduction of 1.26% and contributed to an absolute WER reduction of 2.0%. While the RWER is considerably reduced by using the multiword training procedure, the application of the acoustic-prosodic classifier to the hypothesized ASR output to detect and correct disfluencies is far from beneficial. The false alarm rate is also reasonably high for such a procedure. In future work, we plan to investigate our method on other conversational speech corpora. We also envision that incorporating scores from the disfluencies into the word lattice along with acoustic and language model scores will be beneficial to ASR performance on spontaneous speech.

## REFERENCES

[1] W. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[2] W. Ward, "Understanding spontaneous speech: The phoenix system," in *Proceedings ICASSP*, pp. 365–367, 1991.

[3] E. E. Shriberg, "Acoustic properties of disfluent repetitions," in *Proc. International Congress of Phonetic Sciences*, (Stockholm), pp. 384–387, Aug. 1995.

[4] M. C. Plauche and E. E. Shriberg, "Data-driven subclassification of disfluent repetitions based on prosodic features," in *Proc. International Congress of Phonetic Sciences*, vol. 2, (San Francisco), pp. 1513–1516, Oct. 1999.

[5] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. EUROSPEECH*, vol. 1, (Budapest), pp. 307–310, Sept. 1999.

[6] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. EUROSPEECH '97*, vol. 5, (Rhodes, Greece), pp. 2383–2386, Sept. 1997.

[7] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proceedings ICASSP*, vol. 1, (Atlanta, GA), pp. 405–409, 1996.

[8] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources," in *Proc. Eurospeech*, (Geneva), pp. 957–960, Sept. 2003.

[9] E. C. Matthew Lease and M. Johnson, "Parsing and its applications for conversational speech," in *Proceedings ICASSP*, 2005.

[10] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *3rd ESCA Workshop on Speech Synthesis*, (Jenolan Caves, Australia), pp. 77–80, 1998.

[11] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. International Conference on Spoken Language Processing*, vol. 2, (Denver, CO), pp. 901–904, Sept. 2002.

[12] B. Pellom, "Sonic: The university of colorado continuous speech recognizer," tech. rep., University of Colorado, Boulder, Colorado, 2001.

[13] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 327–332, May 2001.