Proceedings of HLT-NAACL 2003
Student Research Workshop , pp. 37-42
Edmonton, May-June 2003

Word Fragment Identification Using Acoustic-Prosodic Features in Conversational Speech

Yang Liu[1,2]
[1] ICSI, Berkeley, CA 94704 U.S.A
[2] Purdue University, West Lafayette, IN 47907 U.S.A
yangl@icsi.berkeley.edu

Abstract

Word fragments pose serious problems for speech recognizers. Accurate identification of word fragments will not only improve recognition accuracy, but also be very helpful for disfluency detection algorithm because the occurrence of word fragments is a good indicator of speech disfluencies. Different from the previous effort of including word fragments in the acoustic model, in this paper, we investigate the problem of word fragment identification from another approach, i.e. building classifiers using acoustic-prosodic features. Our experiments show that, by combining a few voice quality measures and prosodic features extracted from the forced alignments with the human transcriptions, we obtain a precision rate of 74.3% and a recall rate of 70.1% on the downsampled data of spontaneous speech. The overall accuracy is 72.9%, which is significantly better than chance performance of 50%.

1 Introduction

Word fragments[1] occur frequently in spontaneous speech, and are good indicators for speech disfluencies (Heeman and Allen, 1999; Nakatani and Hirschberg, 1994). When expressed as a percentage of the disfluencies that contain a word fragment, Levelt found 22% for a pattern description task in Dutch (Levelt, 1983); Lickley reported 36% for casual conversations in British English (Lickley, 1994); Bear et al. found 60% for the ATIS corpus (Bear et al., 1992). We examined 83 conversations of Switchboard corpus (Godfrey et al., 1992) and found that about 17% of the disfluencies contain word fragments. However, accurate identification of word fragments is still an unsolved problem in speech community. In most cases, they are simply treated as Out-of-Vocabulary words or are often incorrectly recognized as words in the vocabulary. This not only affects the neighboring words, causing an increase in word error rate, but also fails to provide the important information that a word fragment is detected thus increasing the probability of a disfluency.

The following is an example of the human transcription and the speech recognition output[2] from the Switchboard corpus (Godfrey et al., 1992):

Human transcription:
and it's all just you know i've just eating more sort of eat to my apper- appetite

Recognizer output:
and it's all just see now i'm just eating more sort of need to my out bird's appetite

We can see that in the recognition results, the word fragment 'apper-' is incorrectly recognized as two words in the vocabulary. Additionally, due to the failure to identify the word fragment 'apper-', it will be extremely difficult to identify the disfluency in the recognition results.

The study of word fragments has been conducted from different standpoints. Psychologists and linguists (Levelt, 1989) suggest that speakers rarely interrupt a word when it is correct on its own, but they often do so when it is not. Levelt proposed that "by interrupting a word, a speaker signals to the addressee that the word is an error. If a word is completed, the speaker intends the listeners to interpret it as correctly delivered" (Levelt, 1989). So when a word is complete, the speakers are committing themselves to its correctness (at least at that moment).

While linguists and psycholinguists have considered this problem from the production point of view, we consider this problem from a recognition standpoint, with

[1] A word fragment, also called a partial word, happens when a speaker cuts off in the middle of a word.

[2] The presence of a word fragment in the example is represented by a '-' after the partial word. The recognition output is from SRI's recognizer system.

the goal of identifying disfluencies in spontaneous speech and improving speech recognition.

As noted in (Bear et al., 1992), knowledge about the location of word fragments would be an invaluable cue to both detection and correction of disfluencies. Heeman and Allen proposed an integrated model for the detection of speech repairs (Heeman and Allen, 1999). In that model, word fragments are used as an important feature. Nakatani and Hirschberg proposed a "speech-first" model for the detection of speech repairs using acoustic-prosodic cues, without relying on a word transcription (Nakatani and Hirschberg, 1994). They found that the presence of word fragments is an important indicator of speech repairs, along with the other prosodic-acoustic features such as silence duration, energy, and pitch. They analyzed the properties of word fragments, for example, the distribution of the fragments in syllable length, the distribution of initial phonemes in the fragments, and some acoustic cues (glottalization and coarticulation) in the fragments. Although the role of word fragments as an indicator of disfluencies is emphasized, they did not address the problem of how to detect the occurrence of word fragments, but only suggest that a word-based model for word fragment detection is unlikely. O'Shaughnessy (O'Shaughnessy, 1993) observed in the corpus of ATIS that when speaker stopped in the middle of a word and resumed speaking with no changed or inserted words (i.e. a repetition), the pause lasted 100-400 ms in 85% of the examples (with most of the remaining examples having pause of about 1 second duration). He also found that three-fourths of the interrupted words do not have a completion of the vowel in the intended word's first syllable (e.g., the speaker stopped after uttering the first consonant).

Although word fragments should play an important role for the disfluency processing in spontaneous speech, the identification of word fragments is still an unsolved problem in the speech community. It is impossible or possibly confusing to include all the partial words in the dictionary and therefore treat word fragments as regular words. If one acoustic model is built for all the word fragments, it may be quite difficult to train a good model to cover all the word fragments due to the variability of the possible partial words. Rose and Riccardi modeled word fragments (using a single word fragment symbol *frag*) in their system "How May I Help You" (Rose and Riccardi, 1999). Their system was improved by explicitly modeling all the filled pauses, word fragments and non-speech events; however, it did not report the effect that modeling word fragments made.

In this paper, we investigate the problem of word fragment detection using a new approach, i.e. from the properties of speech analysis. Our goal in this paper is to investigate whether there are reliable acoustic-prosodic properties for word fragments that can be used for automatically detecting their presence.

The paper is organized as follows. In Section 2 we introduce the acoustic and prosodic features that we investigate for word fragment detection. Section 3 describes the corpus and our experimental results. Conclusions and future work are found in section 4.

2 Acoustic and Prosodic Features

Our hypothesis is that when the speaker suddenly stops in the middle of a word, some prosodic cues and voice quality characteristics exist at the boundary of word fragments; hence, our approach is to extract a variety of acoustic and prosodic features, and build a classifier using these features for the automatic identification of word fragments.

2.1 Prosodic Features

Recently, prosodic information has gained more importance in speech processing (Shriberg and Stolcke, 2002). Prosody, the "rhythm" and "melody" of speech, is important for extracting structural information and automating rich transcriptions. Past research results suggest that speakers use prosody to impose structure on both spontaneous and read speech. Such prosodic indicators include pause duration, change in pitch range and amplitude, global pitch declination, and speaking rate variation. Since these features provide information complementary to the word sequence, they provide a potentially valuable source of additional information. Furthermore, prosodic cues by their nature are relatively unaffected by word identity, and thus may provide a robust knowledge source when the speech recognition error rate is high.

In the following we describe some of the prosodic features we have investigated for the word fragment detection task. These prosodic features have been employed previously for the task of detecting structural information in spontaneous speech such as sentence boundary, disfluencies, and dialog act. Experiments have shown that prosody model yields a performance improvement when combined with lexical information over using word level information alone (Shriberg and Stolcke, 2002).

We used three main types of prosodic features — duration, pitch and energy. Duration features were extracted from the alignments obtained from the speech recognizer. Examples of duration features are word duration, pause duration, and duration of the last rhyme in the word. Duration features are normalized in different ways such as by using the overall phone duration statistics, and speaker-specific duration statistics.

To obtain F0 features, pitch tracks were extracted from the speech signal and then post-processed by using a log-normal tied mixture model and a median filter (Sonmez et al., 1997), which computes a set of speaker-specific pitch

range parameters. Pitch contours were then stylized, fit by a piecewise linear model. Examples of pitch features computed from the stylized F0 contours are the distance from the average pitch in the word to the speaker's baseline F0 value, the pitch slope of the word before the boundary, and the difference of the stylized pitch across word boundary.

For energy features, we first computed the frame-level energy values of the speech signal, then similarly to the approach used for F0 features, we post-processed the raw energy values to get the stylized energy.

In addition to these prosodic features, we also included features to represent some ancillary information, such as the gender of the speaker, the position of the current word in the turn[3], and whether there is a turn change. We included these non-prosodic features to account for the possible interactions between them and the other prosodic features.

2.2 Voice Quality Measures

Human speech sounds are commonly considered to result from a combination of a sound energy source modulated by a transfer (filter) function determined by the shape of the vocal tract. As the vocal cords open and close, puffs of air flow through glottal opening. The frequency of these pulses determines the fundamental frequency of the laryngeal source and contributes to the perceived pitch of the produced sound.

The voice source is an important factor affecting the voice quality, and thus much investigation focuses on the voice source characteristics. The analysis of voice source has been done by inverse filtering the speech waveform, analyzing the spectrum, or by directly measuring the airflow at the mouth for non-pathological speech. A widely used model for voice source is the Liljencrants-Fant (LF) model (Fant et al., 1985; Fant, 1995). Research has shown that the intensity of the produced acoustic wave depends more on the derivative of the glottal flow signal than the amplitude of the flow itself.

An important representation of the glottal flow is given by the Open Quotient (OQ). OQ is defined as the ratio of the time in which the vocal folds are open to the total length of the glottal cycle. From the spectral domain, it can be formulated empirically as (Fant, 1997):

$$5.5 * OQ = log((H_1^* - H_2^* + 6)/0.27) \qquad (1)$$

where H_1^* and H_2^* are the amplitudes of the first and the second harmonics of the spectrum.

Different phonation types, namely, modal voicing, creaking voicing and breathy voicing, differ in the amount of time that the vocal folds are open during each glottal cycle. In modal voicing, the vocal folds are closed during half of each glottal cycle; In creaky voicing, the vocal folds are held together loosely resulting in a short open quotient; In breathy voicing, the vocal folds vibrate without much contact thus the glottis is open for a relatively long portion of each glottal cycle.

For our word fragment detection task, we investigate the following voice quality related features.

- **Jitter** is a measure of perturbation in the pitch period that has been used by speech pathologists to identify pathological speech (Rosenberg, 1970); a value of 0.01 represents a jitter of one percent, a lower bound for abnormal speech.

  The value of jitter is obtained from the speech analysis tool *praat* (Boersma and Wennik, 1996). The pitch analysis of a sound is converted to a point process, which represents a sequence of time points, in this case the times associated with the pitch pulses. The periodic jitter value is defined as the relative mean absolute third-order difference of the point process.

  $$jitter = \frac{\sum_{i=2}^{N-1} |2 * T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i} \qquad (2)$$

  where T_i is the ith interval and N is the number of the intervals of the point process. If no sequence of three intervals can be found whose durations are between the shortest period and the longest period, the result is undefined (Boersma and Wennik, 1996).

- **Spectral tilt** is the overall slope of the spectrum of a speech or instrument signal. For speech, it is, among others, responsible for the prosodic features of accent, in that a speaker modifies the tilt (raising the slope) of the spectrum of a vowel, to put stress on a syllable. In breathy voice, the amplitudes of the harmonics in the spectrum drop off more quickly as the frequency increases than do in the modal or creaky spectra, i.e. breathy voice has a greater slope than creaky voice. Spectral tilt is measured in decibels per octave. We use a linear approximation of the spectral envelope to measure spectral tilt. The average, minimum, and maximum value of the spectral tilt for the word, and a window before the word boundary are included in the feature set.

- **OQ** is defined in Equation (1), derived from the difference of the amplitude of the first and the second harmonics of the spectral envelope of the speech data. Studies have shown that the difference between these two harmonics (and thus the OQ) is a reliable way to measure the relative breathiness

[3]In discourse analysis, all the contiguous utterances made by a speaker before the next speaker begins is referred to as a conversational turn.

or creakiness of phonation (Blankenship, 1997). Breathy voice has a larger OQ than creaky voice. As an approximation, we used F0 and 2*F0 for the first and the second harmonics in the spectrum. Similar to the spectral tilt, we also computed the average, minimum, and maximum OQ value for a word duration or a window before the boundary.

3 Experiments

3.1 Experimental Setup

Our goal is to investigate whether there are some reliable acoustic-prosodic features for word fragments. The task of word fragment identification is viewed as a statistical classification problem, i.e. for each word boundary, a classifier determines whether the word before the boundary is a word fragment or not. For such a classification task, we develop an inventory of input features for the statistical classifier. A CART decision tree classifier is employed to enable easy interpretation of results. Missing features are allowed in the decision trees. To avoid globally suboptimal feature combinations in decision trees, we used a feature selection algorithm to search for an optimal subset of input features (Shriberg et al., 2000).

We used conversational telephone speech Switchboard corpus (Godfrey et al., 1992) for our experiments. In the human transcriptions, word fragments are identified (around 0.7% of the words are word fragments). We use 80% of the data as the training data, and the left 20% for testing. In order to avoid the bias toward the complete words (which are much more frequent than word fragments), we downsampled the training data so that we have an equal amount number of word fragments and complete words. Downsampling makes the decision tree model more sensitive to the inherent features of the minority class.

We generated forced alignments using the provided human transcriptions, and derived the prosodic and voice quality features from the resulting phone-level alignments and the speech signal. The reason that we used human transcriptions is because the current recognition accuracy on such telephone speech is around 70%, which will probably yield inaccurate time marks for the word hypotheses, and thus affect the feature extraction results and also make the evaluation difficult (e.g. determine which word hypothesis should be a word fragment). Even if the human transcription and the forced alignment are used to obtain the word and phone level alignments, the alignments could still be error-prone because the recognizer used for obtaining the alignments does not have a model for the word fragments. Note that we only used transcriptions to get the word and phone level alignments for computing prosodic and voice quality features. We did not use any word identity information in the features for the classification task.

At each boundary location, we extracted prosodic features and voice quality measures as described in Section 2. We trained a decision tree classifier from the downsampled training set that contains 1438 samples, and tested it on the downsampled test set with 288 samples (50% of the samples in the training and test set are word fragments).

3.2 Experimental Results

In Table 1 the results for word fragments vs. complete words classification are shown. The precision and recall for this fragment detection task are 74.3% and 70.1% respectively. The overall accuracy for all the test samples is 72.9%, which is significantly better than a chance performance of 50%. These results suggest some acoustic-prosodic features are indicative for word fragment detection.

Table 1: The word fragment detection results on the downsampled data of Switchboard corpus.

		hypothesis	
		complete	fragment
reference	complete	109	35
	fragment	43	101

Figure 1 shows the pruned decision tree for this task. An inspection of the decision tree's feature usage in the results can further reveal the potential properties that distinguish word fragments from complete words. In Table 2 we report the feature usage as the percentage of decisions that have queried the feature type. Features that are used higher up in the decision tree have higher usage values.

Table 2: The feature usage for the word fragment detection using the Switchboard data.

Feature	Percentage
jitter	0.272
energy slope difference between the current word and the following word	0.241
log ratio between the minimum median filtered F0 in a window before the boundary and the maximum value after boundary	0.238
average OQ	0.147
position of the current turn	0.084
pause duration after the word	0.018

Among the voice quality features, jitter is queried the most by the decision tree. We think that when the speaker suddenly cuts off in the middle of the word, there is abnormality of the vocal fold, in particular the pitch periods,

```
JITTER < 0.049782:  0
|  ENERGY_PATTERN_BOUNDARY in rf,fr,fX,rr,ff,rX :  0
|  |  F0K_WIN_DIFF_LOHI_N < -0.093224:  0
|  |  |  AVG_OQ < 0.60821:  0
|  |  |  |  TURN_CNT < 13.5:  FRAGMENT
|  |  |  |  |  ENERGY_PATTERN_BOUNDARY in rf,fX,Xr,ff,Xf :  0
|  |  |  |  |  |  JITTER < 0.018053:  0
|  |  |  |  |  |  JITTER >= 0.018053:  FRAGMENT
|  |  |  |  |  ENERGY_PATTERN_BOUNDARY in fr,rr,rX :  FRAGMENT
|  |  |  |  TURN_CNT >= 13.5:  0
|  |  |  |  |  AVG_OQ < 0.20956:  FRAGMENT
|  |  |  |  |  AVG_OQ >= 0.20956:  0
|  |  |  AVG_OQ >= 0.60821:  FRAGMENT
|  |  F0K_WIN_DIFF_LOHI_N >= -0.093224:  0
|  ENERGY_PATTERN_BOUNDARY in Xr,Xf :  FRAGMENT
JITTER >= 0.049782:  FRAGMENT
|  F0K_WIN_DIFF_LOHI_N < -0.14995:  FRAGMENT
|  F0K_WIN_DIFF_LOHI_N >= -0.14995:  FRAGMENT
|  |  ENERGY_PATTERN_BOUNDARY in rf,fX,rr,ff,rX :  0
|  |  |  PAU_DUR < 12.5:  0
|  |  |  PAU_DUR >= 12.5:  FRAGMENT
|  |  ENERGY_PATTERN_BOUNDARY in fr,Xr,Xf :  FRAGMENT
```

Figure 1: The pruned decision tree used to detect word fragments. The indent represents the tree structure. Each line corresponds to a node in the tree. A question regarding one feature is associated with each node. The decision is made in the leaf nodes; however, in the figure we also show the majority class passing along an internal node in the tree.

and this is captured by jitter. The average of OQ is also chosen as a useful feature, suggesting that a mid-word interruption generates some creaky or breathy voice. The questions produced by the decision tree show that word fragments are hypothesized if the answer is positive to the questions such as 'jitter $>$ 0.018053', 'average OQ $<$ 0.020956?' and 'average OQ $>$ 0.60821?'. All these questions imply abnormal voice quality. We have also conducted the same classification experiments by only using jitter and average OQ two features, and we obtained a classification accuracy of 68.06%.

We also observe from the table that one energy feature and one F0 feature are queried frequently. However, we may need to be careful of interpreting these prosodic features, because some word fragments are more likely to have a missing (or undefined) value for the stylized F0 or energy features (due to the short duration of the word fragments and the unvoiced frames). For example, in one leaf of the decision tree, word fragment is hypothesized if the energy slope before the boundary is an undefined value (as shown in Figure 1, the question is 'ENERGY_PATTERN_BOUNDARY in Xr, Xf?', where 'X' means undefined value).

Notice that the usage of the pause feature is very low, although a pause is expected after a sudden closure of the speaker. One reason for this is that the recognizer is more likely not to generate a pause in the phonetic alignment results when the pause after the mid-word interruption is very short. For example, around 2/3 of the word fragments in our training and test set are not followed by a pause based on the alignments. Additionally, there are many other places (e.g. sentence boundaries or filled pauses) that are possible to be followed by a pause, therefore being followed by a pause cannot accurately distinguish between a word fragment and other complete words.

4 Conclusion and Future Work

Word fragment detection is very important for identifying disfluencies and improving speech recognition. In this paper, we have investigated the problem of word fragment detection from a new approach. We extracted a variety of prosodic features and voice quality measurement to capture the possible acoustic cues at the location of word fragments. Experimental results show that acoustic-prosodic features provide useful information for word fragment detection. These results offer an alternative view of the approach from building acoustic models in a recognizer to handle word fragments and suggest that speech analysis can be quite relevant to building better speech recognition approaches.

These results are very preliminary. For example, experiments were only conducted using the downsampled data due to the extremely highly skewed data distribution. The current word fragment detection method would generate many false alarms in the real test situation, i.e. non-downsampled data. In addition, large corpora must certainly be examined and more sophisticated versions of the measures than we have used should be investigated, especially the voice quality measurements we used. However, as a first approximation of the characterization of word fragments via the acoustic-prosodic cues, we find these results encouraging. In particular, our ability to identify word fragments using only a few features seems promising. The potential features revealed by the experiments in this paper may be helpful to the method of building acoustic model for word fragment detection. Furthermore, we also need to investigate the performance when applying such an approach to the speech recognition results. Finally, a unified framework for word fragment and the disfluency detection is also a future direction of our work.

5 Acknowledgments

The author gratefully acknowledges Mary Harper for her comments on this work. Part of this work was conducted at Purdue University and continued at ICSI where the author is supported by DARPA under contract MDA972-02-C-0038. Thank Elizabeth Shriberg, Andreas Stolcke and Luciana Ferrer at SRI for their advice and help with the extraction of the prosodic features. They are

supported by NSF IRI-9619921 and NASA Award NCC 2_1256. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA, NSF, or NASA.

References

J. Bear, J. Dowding, E. Shriberg. 1992. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.

B. Blankenship. 1997. The Time Course of Breathiness and Laryngealization in Vowels. Doctoral Dissertations, UCLA.

P. Boersma, D. Wennik. 1996. http://www.praat.org/. Praat, a System for Doing Phonetics by Computer.

G. Fant, J. Liljencrants, Q. Lin. 1985. A Four-parameter Model of Glottal Flow. STL-QPSR, 4:1-13.

G. Fant. 1995. The LF-model Revisited. Transform and Frequency Domain Analysis. STL-QPSR, 2-3:119-156.

G. Fant. 1997. The Voice Source in Connected Speech. Speech Communication, 22:125-139.

J. Godfrey, E. Holliman, J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proceedings of IEEE Conference on Acoustics, Speech, and Signal processing, pp. 517-520.

P. Heeman, J. Allen. 1999. Speech Repairs, Intonational Phrases and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. Computational Linguistics.

W. J. M. Levelt. 1983 Monitoring and Self-repair in Speech. Cognition, 14:41-104.

W. J. M. Levelt. 1989. Speaking: From Intention to Articulation. MA: MIT Press.

R. J. Lickley. 1994. Detecting Disfluency in Spontaneous Speech. Doctoral Dissertation, University of Edinburgh.

C. Nakatani, J. Hirschberg, 1994. A Corpus-based Study of Repair Cues in Spontaneous Speech. Journal of the Acoustical Society of America, pp. 1603-1616.

R. C. Rose and G. Riccardi. 1999. Modeling Disfluency and Background Events in ASR For A Natural Language Understanding Task. In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing.

A. E. Rosenberg. 1970. The Effect of Glottal Pulse Shape on the Quality of Natural Vowels. Journal of The Acoustical Society of America vol. 49, pp. 583-590.

D. O'Shaughnessy. 1993. Analysis and Automatic Recognition of False Starts in Spontaneous Speech. In Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing, pp. 724-727.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, G. Tür. 2000. Prosody-based Automatic Segmentation of Speech into Sentences and Topics. Speech Communication vol. 32, pp. 127-154.

E. Shriberg, A. Stolcke, 2002. Prosody Modeling for Automatic Speech Recognition and Understanding. In Proceedings of the Workshop on Mathematical Foundations of Natural Language Modeling.

M. K. Sonmez, L. Heck, M. Weintraub, E. Shriberg. 1997. A Lognormal Tied Mixture Model of Pitch For Prosody-Based Speaker Recognition. In Proceedings of Eurospeech, pp. 1391-1394.