Using a Uniform-Weight Grammar to Model Disfluencies in Stuttered Read Speech: A Pilot Study

Kristy Hollingshead and Peter A. Heeman

Technical Report CSLU-04-002
July 2004

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
{hollingk,heeman}@cslu.ogi.edu

Abstract

Stuttering is a speech disorder characterized by certain types of speech disfluencies, such as sound repetitions, which are frequent enough to be disruptive. Speech therapists frequently use manual counts of these speech disfluencies to diagnose whether a child stutters and to track improvement through a treatment program. However, these counts are subjective, inconsistent, and prone to error. We propose the use of speech recognition technology to automate these counts, thus providing an objective and consistent measurement. Since many of the disfluencies in stuttered speech obey certain regularities, we built several grammar-based language models that capture these regularities to detect disfluencies. These grammars have uniform transition weights. We tested the grammars on a short sample of stuttered speech and analyzed their performance based on word error rate and how well they detect and classify disfluencies. Results indicate that these grammars detect repetition disfluencies poorly, particularly phoneme repetitions. We are currently building a probabilistic language model and expect that the more accurate transition probabilities will lead to a decrease in false positives. We also expect that further improvement will be achieved by modeling the acoustic properties of stuttered speech.

1 Introduction

Both stutterers and non-stutterers have *speech disfluencies*, which are mistakes or interruptions in the stream of words a speaker intends to say, but disfluencies are noticeably more frequent in stutterers' speech. The disfluency classes most often associated with stuttering are termed *stuttering events*, and include abnormal repetition of sounds, syllables, and words ("He wa-wa-was a good king"), sound prolongation ("He wwwas a good king"), and hard blocks during speaking ("He w——as a good king").

A count of the stuttering events in a subject's speech is a common measure for stuttering diagnosis, but the counts are inconsistent across clinicians and clinics (Cordes et al., 1992; Kully and Boberg, 1988). Reasons for inconsistency include the influence of a clinician's subjective opinion, each clinic using slightly different definitions of stuttering events, and mistakes in the counts (Yaruss, 1997). Our long-term goal is to address this problem by developing an automated stuttering assessment tool that detects and classifies disfluencies in children's stuttered speech in order to produce an objective rating of the stuttering severity.

To detect the disfluencies of stutterers, we will harness previous work on disfluencies of non-stutterers, such as filled pauses, repetitions, and repairs. This work has used statistical language models to improve detection of disfluencies during *spontaneous speech* (Heeman and Allen, 1999; Stolcke et al., 1999). Since processing children's speech is often difficult for speech recognizers and children's stuttered speech will be more so, we built the current implementation of our system for a *read speech* task, where the subject reads a given text aloud. Recognizing read speech is easier for speech recognition systems because the system has prior knowledge about the subject's intended words; therefore, it has a smaller search space for predicting the next word that will be spoken. Although read speech is not typically as rich in disfluencies as spontaneous speech, children and stutterers are often disfluent when reading aloud (Banerjee et al., 2003a; Banerjee et al., 2003b; Mostow et al., 2002). Furthermore, the read speech task is used as a part of most stuttering assessments (Gregory et al., 2003; Riley, 1994).

Since the language model for a read speech task is so constrained, we theorize that language models based on simple uniform-weight grammars, rather than statistical language models, will be sufficient for disfluency recognition in this task. We test how well uniform-weight grammars detect 29 disfluencies in 3 minutes of read speech from a child stutterer. Our approach is similar to that used by Nöth et al. (2000), but includes more details about the location and class of disfluencies identified by the system. Therefore, we are able to provide a more thorough evaluation of the grammars and pinpoint the areas of greatest difficulty.

2 An Overview of Stuttering and its Diagnosis

2.1 Definition

Stuttering is a speech disorder with many definitions. According to Wingate (1964; as cited in Yaruss, 1997), part of the reason for so many definitions of stuttering is that its cause has yet to be identified. There have been many proposals as to the cause, per Johnson et al. (1959), including genetics, physical impairments, and psychosomatics. With no known cause, definitions of stuttering are based on the symptoms exhibited by stutterers. Stuttering symptoms can vary widely across the population of stutterers and include concomitant physical actions that occur at the same time as a disfluency, such as foot stomping and head nodding, and covert behaviors such as avoiding particular words or sounds. The most common symptom is the occurrence of speech disfluencies, which are widely used to define and identify stuttering.

2.2 Disfluencies

Identifying stuttering is complicated by the fact that nearly everyone, including non-stutterers, have some disfluencies in their speech. While speaking, nearly everyone makes mistakes that need to be corrected. Correcting the mistake might consist of changing what was said or simply repeating it; the result is that the speaker has repaired what he was saying, and so these corrections are termed *speech repairs*.

2.2.1 Disfluency Structure

Heeman and Allen (Heeman, 1999; Heeman and Allen, 1999) suggest that listeners use the regular structure of speech repairs to process the repairs seemingly effortlessly. This regular structure consists of a *reparandum*, an optional *editing term*, and an *alteration*, as shown in Figure 1. The reparandum is the stretch of speech the speaker intends to change. The interruption point (IP) follows the reparandum. Editing terms consist of filler words, such as “um” or “uh”, or cue words or phrases, such as “let’s see”. The alteration is the speech the speaker intends as a replacement for the reparandum. Some stuttering events, such as repetition disfluencies, belong to the class of speech repairs. In a repetition disfluency, each repeated unit (sound, word, or phrase) is part of the reparandum, the IP follows the last repeated unit, and the final, fluent unit is the alteration. Our current system implementation detects repetition disfluencies only; we can use their regularities to model them and thereby improve detection of them.

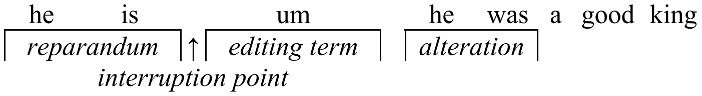

Figure 1: The structure of a speech repair

2.2.2 Disfluency Classes of Stutterers and Non-Stutterers

There are eight well-known disfluency classes, which Johnson et al. (1959) first defined in a seminal study on stuttering. The disfluency classes are: interjections, such as “um”; sound or syllable repetitions; word repetitions; phrase repetitions; revisions, to correct something said previously; incomplete phrases where a speaker abandons a topic without completing it; broken words; and sound prolongations. These disfluency classes are certainly not unique to stuttering. In fact, Johnson et al. found that there is considerable overlap in the types of disfluent behaviors produced by individuals who do and do not stutter.

There are not many, if any, clear and quantifiable characteristics to distinguish the disfluencies of non-stutterers and stutterers. Johnson et al. (1959) demonstrated that interjections, revisions, incomplete phrases, and broken words do not differentiate stutterers and non-stutterers. Some of their results showed that the frequency of word repetitions and sound or syllable repetitions provided the best distinction, but even these categories had some overlap between the groups. Wingate (1977) demonstrated that prolongations and sound or syllable repetitions, in a written transcription of speech, are necessary and sufficient to accurately differentiate stutterers and non-stutterers. Logan and LaSalle (1999) attempted to differentiate stutterers and non-stutterers by measuring the frequency of speech disfluencies occurring in close proximity to one another. Although the frequency was higher in stutterers' speech, the difference was not significant enough to distinguish the two groups. Throneburg and Yairi (1994) achieved greater success in differentiating the two groups by examining differences in the duration of word repetitions by preschoolers who do and do not stutterer, and determining that the total duration of the repetitions was shorter for the stutterers.

2.3 Diagnosis Methods

Clearly, distinguishing stutterers and non-stutterers is not a trivial problem. The challenge of distinguishing stutterers and non-stutterers is exacerbated in young children, because it is common for all children to have more disfluencies than adults do. In addition to using a stuttering diagnosis to distinguish stutterers from non-stutterers, clinicians also use the diagnosis to rate a patient's stuttering severity (Gregory et al., 2003). By tracking the severity rating over time, a clinician can measure a patient's progress through a treatment program, as well as the *efficacy* of a treatment program, or how well the program produces the intended effects.

There are a number of diagnosis methods for producing a stuttering severity rating for a patient. All methods require samples of the patient's speech, which might be gathered in real time through direct interaction with the subject, or on videotape. A variety of samples are usually collected because every individual's speech can vary greatly with the speaking context. For example, a child's speech is likely to be very different when he is speaking with a clinician than when he is talking to his sibling at home, regardless of whether or not he stutters. For reviewing the speech samples, some methods advocate the use of real-time analysis, where the clinician analyzes the speech while the subject is talking. Other methods use a transcript of the speech for analysis. Yaruss et al. (1998) provided evidence that the two different techniques produce similar results for rating the stuttering severity and concluded that the best strategy is to use both techniques, to take advantage of the different strengths of each.

Most methods for stuttering assessments require a count of the disfluencies in a patient's speech. Some features of stuttering that are more subjective in nature include overall tenseness, hyperarticulation, and inappropriate word stress, which all contribute to a vague sense of unnaturalness of the stuttered speech. Another measurement of stuttering is the presence of concomitant behavior or covert behavior such as avoiding saying certain words or sounds (Gregory et al., 2003). Another measurement is the speaker's reactions to his own stuttering (Yaruss, 2001). Most of these features are subjective and are not as well-defined as the more objective disfluency count.

An objective method for stuttering diagnosis is provided by the *Stuttering Severity Instrument 3* (Riley, 1994); clinicians use this tool to count the stuttering events per 100 syllables in samples of both spontaneous and read speech, compute the average duration of the three longest stuttering events, and note the presence of physical signs of stuttering. A more comprehensive method that combines objective and subjective features is provided by transcript-based methods such as the *Systematic Disfluency Analysis* (SDA) (Gregory et al., 2003). These methods require a fine-grained analysis of verbatim transcriptions of the subject's speech. For the SDA, clinicians mark and classify disfluencies on the transcript, based on repeated viewings of a videotape; additional markings are also made for audible and visible aspects of stuttering, as well as the duration of the stuttering event. In this way, both subjective and objective features can be documented.

2.4 Problems with Current Methods of Stuttering Diagnosis

Regardless of the method used to diagnose stuttering, the results are similarly problematic: each is still a manual diagnosis from a single clinician. To measure a patient's progress or to compare the efficacy of different treatment programs, there must be clear-cut diagnostic norms. High *intra-coder* and *inter-coder reliability* rates are necessary for establishing these diagnostic norms. Intra-coder reliability indicates that a measurement is repeatable by the same clinician, and inter-coder reliability indicates that a measurement is consistent between two different clinicians. Increasing the intra-coder and inter-coder reliability has been a concern of many researchers. Kully and Boberg (1988) examined the reasons for low reliability rates. In their experiment, clinicians from different speech treatment clinics rated the severity of the stuttering and counted the number of syllables and stuttered syllables in speech samples from stutterers and non-stutterers. One disagreement that arose was the number of syllables that should be counted in a sound repetition such as “k-k-king”. Clinicians also disagree on whether one should count just the stuttering events or all of the disfluency classes (Yaruss, 1997). For these and other reasons, the rate of inter-clinician agreement was low across all measures in (Kully and Boberg, 1988). In the same study, however, Kully and Boberg showed that measurements from clinicians working at the same clinic, and given the same training, had high inter-coder reliability. For intra-coder reliability, Yaruss (1997) encourages clinicians to realize that any stuttering diagnosis involves making some arbitrary decisions and to be aware of the choices they make during a stuttering diagnosis in order to be consistent.

Manual diagnoses are problematic for establishing diagnostic norms, hindering individual progress tracking as well as inter-clinic communication regarding the efficacy of treatment programs and suggestions for improving the programs. We propose that an automated approach would ensure objectivity and consistency across different clinicians and clinics.

3 Goals

3.1 Automated Stuttering Assessment Tool

To address the problems of the manual stuttering diagnosis, we propose the use of speech recognition to automate the stuttering diagnosis process. Our long-term goal is to create a tool that would take a sample of a patient’s speech as input, and output a severity rating.

3.2 Automated Disfluency Counts

As a first significant step towards our long-term goal, we propose the use of speech recognition to automate the disfluency count measure. The output of this tool will have the total number of disfluencies, the total number from each disfluency class, the number of words or syllables spoken, the calculated frequency of disfluencies, and the calculated frequency of each disfluency class. Obviously, to provide a count of total disfluencies and a count of disfluencies from each disfluency class, our system has to both detect and classify the disfluencies in the speech sample.

3.2.1 Model Regularities

Fortunately, stuttering events obey certain regularities that are simple to model automatically, which will make it easier to detect them. They have a regular structure, as discussed in Section 2.2.1. The disfluencies of stutterers also follow certain positional regularities. They are not randomly scattered throughout an individual's speech; they are more likely to occur on particular classes of words such as content words, or at particular positions in a sentence. For example, word and phrase repetitions are more likely to occur at the beginning of a sentence, clause, or prepositional phrase (Logan and LaSalle, 1999). Sound repetitions and prolongations tend to occur at the beginnings of words (Johnson et al., 1959; Johnson and Brown, 1935). These regularities seem to lend themselves to a regular grammar, which we create as described in Section 5.

3.2.2 Use Speech Recognition

Speech recognition systems have much to offer for identifying disfluencies. First, stuttering events are word-based, so with the correct sequence of words identified, it is a simple matter of scanning through the sequence to find stuttering events like word repetitions and phrase repetitions (Wingate, 1977). Furthermore, with the correct sequence of phonemes identified, and a knowledge of how phonemes group into words, it would be a simple matter to find sound repetitions. However, in order to identify the correct sequence of words or phonemes, the system must model disfluencies at the same time as it performs speech recognition. Without disfluency modeling, the system would be unlikely to hypothesize the occurrence of a disfluency, since such an event would be highly unlikely given the current fluency model. With disfluency modeling, however, the system can compare hypotheses of fluent speech with hypotheses of disfluent speech and produce the best fit for the data. Clearly, the task of detecting disfluencies is intertwined with the task of speech recognition, and the two should be performed together.

Second, durational measurements of each phoneme would allow a comparison for estimating which phonemes are most likely to be atypically prolonged sounds in the speech sample. Third, language-specific information is helpful in identifying stuttering events; knowing what the speech should sound like is necessary to identify abnormal speech that is likely to be disfluent. Speech recognizers automatically generate the recognized sequence of words and phonemes, as well as calculated durations of the phonemes. Additionally, speech recognizers are usually language-specific, because they use acoustic models trained on the typical sounds of a language. Therefore, speech recognition appears to be a promising solution for automatically detecting disfluencies.

3.2.3 Build Language Models

In the field of speech recognition, there has already been a fair amount of previous research on identifying the disfluencies of non-stutterers (Heeman, 1999; Stolcke et al., 1999). This work has focused on the use of statistical *language models* to improve speech recognition during disfluent speech. The goal of speech recognition is to find the most probable sequence of words given the acoustic signal. Language models augment speech recognizers by assigning a higher probability to the more logical sequence of words. For instance, based on acoustics alone, a speech recognizer might assign equal probabilities to the hypotheses "recognize speech" and "wreck a nice beach". The first phrase makes more sense and is more likely to occur. Therefore, a good language model would assign a higher probability to the first phrase so the speech recognizer would return that phrase as its final hypothesis.

There are two basic types of language models: grammar-based language models and statistical language models. Most of the current research focuses on statistical language models, which estimate the probability that any given word sequence will occur by computing the relative frequency of that sequence in a training corpus. A grammar-based language model is more restrictive in that it imposes a closed-set vocabulary and, typically, defines the order in which words can be recognized. Grammar-based language models do not require a corpus of training data and can be implemented by hand. Since we are using a read speech task, we know the words that the subject is supposed to say as well as the order of the words; therefore, our system uses grammar-based language models.

4 Related Work

4.1 Acoustic Characteristics Model

One approach to automating disfluency counts was proposed by Howell et al. (1997). The implemented system was an artificial neural network (ANN) for classifying a word either as a prolongation, a repetition (sound or part-word), or fluent. This system used the acoustic characteristics of stuttering to model disfluencies. Input to the network was pre-processed by marking the boundaries of each word unit by hand, and then removing the supra-lexical disfluencies, such as phrase repetitions and full-word repetitions.

The nine parameters of the ANN included the following: whole word and part word duration; whole word, first part, and second part fragmentation; whole word, first part, and second part spectral measures; and part word energies. The fragmentation measure indicated whether sections of energy and silence were present in the sample, for differentiating between prolongations and repetitions. The spectral measure indicated the extent of stability in the spectrum, for differentiating between fluent and disfluent (either prolongation or repetition) words. The properties of the first and second half of the words were input separately in order to take advantage of the fact that prolongations, sound repetitions, and part-word repetitions are more likely to occur at the beginning of a word (see Section 3.2.1).

The best-performing ANN used four parameters: whole word fragmentation and spectral measures, and part word duration and energy. The ANN correctly identified 95% of the fluent words and 78% of the prolongations and repetitions combined, but only correctly identified 58% of the prolongations and 43% of the repetitions. Howell et al. speculate that the difficulty in distinguishing between the disfluency classes might have been caused by amplitude fluctuations on prolonged, low-energy voiceless fricatives (such as the /f/ in “four”).

The network designed by Howell et al. does not use speech recognition. Therefore, the network has no prior knowledge regarding what the input word is, so it cannot use foreknowledge about the typical acoustic characteristics for that word to identify atypical characteristics and predict a disfluency.

4.2 Word Regularities Model

A speech recognition approach to automating disfluency counts was implemented by Nöth et al. (2000). The system was composed of a grammar that modeled the positional regularities of the disfluencies of stutterers (see Section 3.2.1) to recognize sound, word, and phrase repetitions, prolongations, hard blocks, and filled pauses during a read speech task. The grammar was constructed as a deterministic automaton, with the phonemes of each word of the text as the nodes, with connecting edges. Additional inserted edges allowed word and phrase repetitions at the beginning of sentences and clauses; sound repetitions were allowed at any point, and were represented by self-loops on every node. Extra nodes and edges were added to allow pauses and filler words at any point.

The system performed quite well, with a high correlation coefficient of 0.99 between the average actual disfluencies per word and the average detected disfluencies per word. Unfortunately, the evaluation of the system did not include many details. It did not include information regarding the location and classification of the disfluencies. Without location information to determine the distance between detected disfluencies and actual disfluencies, the system’s detection accuracy is unclear. Without classification information, the system cannot measure stuttering severity. Some disfluency classes are more noticeable and disruptive than others are. For example, sound repetitions are more disruptive to the listener and thus often associated with stuttering than phrase repetitions. The class label of the disfluencies is important for a stuttering diagnosis and should be included in judging the accuracy of any tool for automated stuttering diagnosis.

5 System to Automate Disfluency Counts

Our system to automate disfluency counts is similar in structure to the one described by Nöth et al. (2000). However, our system includes many more details about each detected disfluency, including its class category and its start and end times, to analyze the system's accuracy on each class of disfluency. We therefore provide a more thorough evaluation of our system than Nöth et al. provided of their system. To begin, we built several uniform-weight grammars. These grammars detect different classes of repetition disfluencies by modeling the positional regularities of the disfluencies of stutterers (see Section 3.2.1). We used each grammar to test the performance of the speech recognizer in the CSLU toolkit (Sutton et al., 1998) on the task of detecting and classifying the disfluencies in read speech from a child stutterer. This paper describes an initial investigation to determine whether a speech recognizer can succeed at the task with simple grammars but without any changes to its acoustic model: in other words, without capitalizing on all of the characteristics of stuttering.

5.1 Data

The data for this project consisted of a 3-minute excerpt of read speech from a child stutterer, provided by Prof. Scott Yaruss, Co-Director of the Stuttering Center of Western Pennsylvania. Using *Speech Viewer* (Sutton et al., 1998), we transcribed the words, partial words, and sounds in the excerpt, similar to the verbatim transcript approach of SDA (described in Section 2.3). We then used *DialogueView* (Heeman et al., 2002) to annotate the disfluencies: namely sound prolongations; hard blocks; hesitations; and sound, word, and phrase repetitions. We annotated the exact location and extent of the disfluencies, including compound disfluencies such as the prolongation of a sound in a word that is then repeated. Our annotation scheme is extended from the word-level scheme used for annotating speech repairs of non-stutterers (Heeman, 1999). We verified our annotations with those done by Prof. Yaruss on the same speech sample.

Disfluency Class	Number Annotated	Example
Prolongation	12	Once there was a lllllittle king.
Hard block/hesitation	1	Far away —in a small land.
Sound repetition	6	He was a g-g-good king.
Syllable repetition	1	He was always ready to help any-anyone in trouble.
Word repetition	7	His friends were fond fond of him.
Phrase repetition	2	He walked with he walked with her.
Total	29	

Table 1: Number of annotated disfluencies in the speech excerpt

The excerpt contained 238 words, including sound repetitions and partial words. We divided the speech into 36 utterances by inserting utterance boundaries at pre-determined locations, namely at prepositional and conjunction words and at the punctuation marks in the read text. We annotated 29 disfluencies in total, as shown in Table 1. While we realize that this is not enough data for statistical significance, we argue that it is enough to have a general sense of how well our approach is working; we are in the process of acquiring more data.

5.2 Scoring Criteria

A grammar is scored according to how well it *detects* disfluent speech. If the interruption point (IP; see Section 2.2.1) of a system-identified disfluency is within 0.5 seconds of the interruption point of an annotated disfluency, the disfluency was detected correctly. A disfluency is also *classified* as either a sound, word, or phrase repetition. If this classification matches the classification of an annotated disfluency with an IP within 0.5 seconds of the IP of the detected disfluency, then the disfluency was classified correctly. A disfluency is scored as being correctly classified only if it is both detected and classified correctly; not only is the disfluency at the right location, but it also has the right class label.

We use the IP to measure how well disfluencies are detected because the point between disfluent speech and fluent speech should be the moment of greatest change, and it is this change in the speech that we wish to capture. We included a margin of error of 0.5 seconds to account for misaligned words. Poor word alignment occasionally occurred, particularly at disfluencies, because acoustic models do not account for the disruptions in the signal caused by disfluencies.

```
                      ↓ interruption point: 0.343 ms
Transcription:  |he |was  |he  |was  |a  |good    |king  |
Hypothesis:         |he    |he      |was  |a |good  |king   |
                     ↑ interruption point: 0.234 ms
```

Figure 2: Transcription and hypothesis to illustrate scoring criteria

Our scoring methods are illustrated in Figure 2. The transcription of what the subject said is aligned with the words that the system recognized. The underlined words are the disfluencies. The system's hypothesis has a word repetition on the first word. The transcript has a phrase repetition on the first two words. The IPs of the two disfluencies are within 0.5 seconds of each other but the disfluencies have different class labels. Thus, the disfluency was detected correctly but incorrectly classified.

The grammars are also rated according to the word error rate (WER). We align the word strings and then count the number of mismatched words, as well as the number of words inserted and deleted, and divide this number by the total number of annotated words. We impose a more stringent error criterion than is used in the standard National Institute of Standards and Technology (NIST) scoring routines, by requiring that matched words overlap in time as well. An accurate estimation of word location is important since the locations of the words will affect how well the system can estimate the location of stuttering events.

5.3 Grammars

The speech recognizer was provided with a separate grammar for each utterance. The grammars, which are relatively simple and can be represented as regular expressions, limit the speech recognizer to a closed-set vocabulary and the order in which the words can be recognized. We use a *base text* to construct each grammar, by sequentially adding each word from the base text to the grammar. For a statistical language model, the base text would be the training corpus. For the grammar-based language model for a read speech task, there are several alternatives available.

5.3.1 Base Text

The first alternative for the base text is the *target text*, which is the text the subject is supposed to read. However, using the target text introduces the problem of *miscues*, or reading mistakes. When reading a text aloud, children often substitute one word for another, or insert extraneous words, or skip over a word entirely. Project LISTEN at Carnegie Mellon University, which is using speech recognition technologies to create a reading tutor that detects and corrects a child's reading mistakes, has implemented several approaches to improve the recognition of miscues (Banerjee et al., 2003a; Banerjee et al., 2003b; Mostow et al., 2002).

Since we do not wish to duplicate this work on detecting miscues, we choose the second alternative, using a transcript of what the subject said. In considering the transcript as the base text, there are two further options. The first is to use the *verbatim transcript*. The verbatim transcript contains the exact words the subject said, along with the start and end times of each word. With the verbatim transcript as the base text, the task becomes a simple word-alignment task. There are no branches in the grammar; the speech recognizer is tightly constrained to recognize exactly the words that were transcribed, in the order they were transcribed. The second option is to create and use a *fluent transcript* as the input. The fluent transcript is created by removing all disfluencies from the verbatim transcript; the remaining text is the subject's fluent speech with all of his or her miscues. With this information as the base text, the speech recognizer is forced to rely on accurate modeling of the disfluencies of stutterers to detect and classify them. Thus, the grammars for this experiment use the fluent transcript as their base text.

5.3.2 Disfluency Oracle

We use our disfluency annotations to create a *disfluency oracle* to provide the system with the location and class of each of the annotated disfluencies. With this information, certain disfluencies can be re-inserted into the fluent transcript; for example, our system only models repetition disfluencies, so the disfluencies of all other classes are included in the base text. The disfluency oracle allows us to vary the level of disfluency information provided to the system.

The disfluency oracle is also useful for estimating the probability of a disfluency occurring. We are in the process of implementing a probabilistic model to capture some of the regularities of the disfluencies of stutterers that are difficult to capture in a uniform-weight grammar. For now, we simulate a probabilistic model by using the disfluency oracle to add annotated disfluencies to the grammar as an optional path. Thus, at the location of each annotated disfluency, the probability of detecting a disfluency in the same class as the annotated disfluency is 1/2. The probability of detecting a disfluency in a different class at that location or at a location where no disfluencies were annotated (i.e., during fluent speech) is 0. Since the information provided by the oracle becomes an optional path in the grammar, we define such grammars to be the *optional oracle* grammars.

5.3.3 Components

Each grammar has three separate components, for modeling sound, word, and phrase repetitions. We say that a grammar *models* a disfluency class if the grammar contains an optional path for that disfluency class. The components can be activated separately or in combination. For an active component, optional paths for that disfluency class are inserted into the grammar, either according to the regularities of that class, defined below, or at the locations of the appropriate annotated disfluencies. For an inactive component, the disfluency oracle provides the location of disfluencies of that disfluency class and the system is constrained to recognize those disfluencies at the given locations. By activating and deactivating the components to model disfluency classes separately and in different combinations, we are able to determine which class of repetition disfluencies is most difficult to detect. We can also see the interaction between the tasks of modeling each disfluency class.

The regularities of repetitions are modeled in our system as follows. Sound repetitions consist of single-phoneme repetitions at the beginning of each whole word. Word repetitions are any whole word being repeated one or more times. Phrase repetitions are comprised of a sequence of two to five words repeated at pre-determined boundaries, namely at prepositional and conjunction words and at the punctuation marks from the target text.

5.3.4 Naming Conventions

We constructed 15 different grammars with uniform weights. Each of the grammars has a three-digit name to represent its three components and how the components are used in that grammar. The digit 0 indicates that the component is inactive, the digit 1 indicates that disfluencies in the component's disfluency class are modeled by regularities, and the digit 2 indicates that disfluencies in the component's disfluency class are modeled by the optional oracle. Each digit has a superscript to indicate which disfluency class it represents: P for phrase repetitions, W for word repetitions, and S for sound repetitions. For example, the hypothetical name $2^P1^W0^S$ indicates that this grammar's sound repetition component is inactive, and the word and phrase repetition components are active. The word repetitions are modeled by regularities and the phrase repetitions are modeled by the optional oracle.

We built a *gold-standard* grammar: $0^P0^W0^S$. All of the components of this grammar are inactive, so the grammar is a simple sequence of every word and sound transcribed in the verbatim transcript. The purpose of a gold standard is to provide a measurement of the best possible results. This grammar is used to test the accuracy of the acoustic model for the system, rather than to test the performance of the grammar. We also built grammars with just one active component, then with two active components, and finally with all three active components.

6 Results and Analysis

Each of the 15 grammars was used with the speech recognizer to detect the disfluencies in the three-minute excerpt of speech, which contained 29 disfluencies. Of the 29 annotated disfluencies, only 15 are sound, word, or phrase repetitions; the others are hard blocks, hesitations, and prolongations, which our system does not currently model. However, our system might inadvertently detect one of these other disfluencies. For example, the system might detect a prolongation and classify it as a sound repetition, as was the case for five of the 12 prolongation disfluencies with grammar $0^P0^W1^S$. We do not want to penalize our system in this case. Therefore, we score the results out of 29 actual disfluencies rather than out of 15.

For the detection and classification measures, we use the usual notions of *recall* and *precision* to evaluate how well the system detected disfluencies. Recall measures how many annotated disfluencies were correct out of the total number annotated, and precision measures how many of the system-identified disfluencies were correct out of the total number of system-identified disfluencies.

Note that allowing the system to receive extra credit for detecting any disfluency class affects the recall and precision rates such that the number of correct annotated disfluencies does not necessarily equal the number of correct system-identified disfluencies. Additionally, a system-identified disfluency may receive credit for detecting more than one annotated disfluency, in the case of compound disfluencies. For example, if the subject were to prolong a sound at the beginning of a word and then repeat the word, then it would be conceivable that the IP of a detected word repetition was within 0.5 seconds of the IP of an annotated word repetition as well as the IP of a prolongation.

6.1.1 Gold-Standard Grammar

Grammar $0^P0^W0^S$ does not explicitly model any of the repetition classes but relies solely on the disfluency oracle to determine the location of all three classes of disfluencies. This grammar had a word error rate of 4%, and an overall classification recall and precision rate of 10/29. It had an overall detection recall rate of 14/29 and an overall detection precision rate of 10/15. Since the scores for this grammar are not perfect, even though the system was forced to use the exact, correct sequence of words and sounds, it is clear that the system's ability to perform the word-alignment task is not perfect either. Improving the word-alignment performance will require an improved acoustic model so that the system is better able to match up the acoustics and the word sequence. Detailed results of this gold-standard grammar are reported, in gray text, alongside the results for the other models for ease of comparison.

6.1.2 Regularities Grammars

It is clear from the word error rate (WER) shown in Figure 3 for the Regularities grammars that grammars with an active sound repetition component have the worst performance. However, WER is too coarse of a measure to provide enough details to understand why some grammars performed better than others do. More details are required to understand why this is the case.

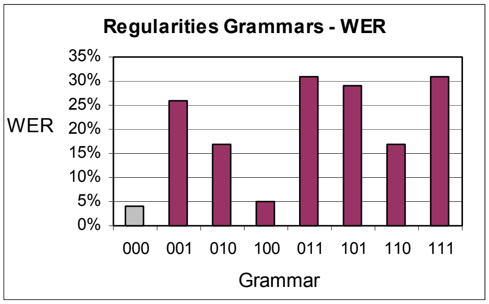

Figure 3: Word error rate (WER) for Regularities grammars

From Table 2 it is clear that the Regularities grammars have a high rate of *false positives*, where false positives are the total number of system-identified disfluencies minus the number of correct disfluencies. Since the grammars with active sound repetition components have the worst error rates, it is clear that modeling sound repetitions poses the greatest challenge to the system. The details captured by the detection measure allow for a more thorough analysis and understanding of how well the grammars are working.

Regularities Grammars - Detection									
	Modeled Repetitions						**Others**		
	Sound Repetition		**Word Repetition**		**Phrase Repetition**		**Prolongation**	**Hesitation**	**Syllable Repetition**
Grammar	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Recall**	**Recall**
$0^P0^W0^S$	4/6	4/6	6/7	6/7	0/2	0/2	4/12	0/1	0/1
$0^P0^W1^S$	2/6	9/35	3/7	--	0/2	--	5/12	1/1	0/1
$0^P1^W0^S$	0/6	--	3/7	6/19	1/2	--	2/12	1/1	0/1
$1^P0^W0^S$	0/6	--	0/7	--	0/2	0/3	0/12	0/1	0/1
$0^P1^W1^S$	2/6	8/31	4/7	4/16	1/2	--	7/12	1/1	0/1
$1^P0^W1^S$	2/6	9/37	3/7	--	0/2	0/2	5/12	1/1	0/1
$1^P1^W0^S$	0/6	--	2/7	4/18	0/2	0/2	2/12	1/1	0/1
$1^P1^W1^S$	2/6	9/31	4/7	2/14	0/2	0/3	7/12	1/1	0/1

Table 2: Detection rates for grammars with the disfluency class of one or more components modeled by regularities. (Precision rates are not available for inactive components, and also are not available for the disfluency classes that the system does not currently model, under the heading Others.)

Interestingly, most of the grammars detect prolongations, with recall rates up to 58%, even though the system does not model prolongations. The grammars which model sound repetitions have the highest prolongation recall rates, from 42% to 58%. It appears that the system can detect the disfluent speech of prolongations from the acoustic information, but since the system does not model prolongations, it must use a different disfluency class to indicate the location of this disfluent speech.

In addition to using sound and word repetitions to detect prolongations, the system also uses a word repetition to detect a phrase repetition. Unlike prolongations, however, the system does model phrase repetitions. The only grammars that use a word repetition to indicate a phrase repetition are those where the word repetition component is active and phrase repetition component is inactive (grammars $0^P1^W0^S$ and $0^P1^W1^S$). Since the phrase repetition component is inactive, the phrase repetition was included in the sequence of words that the system must recognize. The fact that the system inserts a word repetition as well is further indication of a problem with the word alignment.

The classification rates shown in Table 3 indicate how well the system was able to classify the disfluencies with the Regularities grammars. The classification precision rates in Table 3 are worse than the detection precision rates in Table 2. The lower precision rate for this task shows that disfluency classification is more difficult than detection.

Regularities Grammars - Classification						
	Sound Repetition		**Word Repetition**		**Phrase Repetition**	
Grammar	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
$0^P0^W0^S$	4/6	4/6	6/7	6/7	0/2	0/2
$0^P0^W1^S$	2/6	2/35	--	--	--	--
$0^P1^W0^S$	--	--	3/7	3/19	--	--
$1^P0^W0^S$	--	--	--	--	0/2	0/3
$0^P1^W1^S$	2/6	2/31	2/7	2/16	--	--
$1^P0^W1^S$	2/6	2/39	--	--	0/2	0/2
$1^P1^W0^S$	--	--	2/7	2/18	0/2	0/2
$1^P1^W1^S$	2/6	2/31	1/7	1/14	0/2	0/3

Table 3: Classification rates for grammars with the disfluency class of one or more components modeled by regularities.

Note that the classification rates improve as more of the components are activated. The system is no longer forced to use the wrong disfluency class to indicate the location of disfluent speech. By examining the detection rates in Table 2 for grammar $0^P0^W1^S$ and $0^P1^W1^S$, and the classification rates in Table 3 for the same grammars, we see that with grammar $0^P0^W1^S$, the system used sound repetitions to identify three of the word repetitions. However, once the word repetition component was activated in grammar $0^P1^W1^S$, the system used word repetitions, the correct disfluency class, to indicate the locations of the word repetitions. These results indicate an interaction between the tasks of identifying each class of repetition disfluency. Therefore, we argue that a single model that combines the detection and classification of all disfluency classes will perform better than single-disfluency models.

6.1.3 Optional Oracle Grammars

Comparing the WER of the Optional Oracle grammars in Figure 4 to the WER of the Regularities grammars in Figure 3, we see that the Optional Oracle grammars perform better than the Regularities grammars. This was expected, since by simulating a probabilistic model with improved transition probabilities, we limited the number of false positives that could be identified.

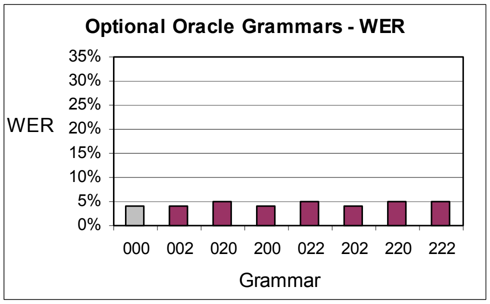

Figure 4: Word error rates (WER) for Optional Oracle grammars

From the detection rates shown in Table 4, we immediately see that the number of false positives has decreased; grammar $0^P0^W1^S$ identified 24 false positives, and grammar $0^P0^W2^S$ only identified two false positives.

Optional Oracle Grammars - Detection

	Modeled Repetitions						Others		
	Sound Repetition		Word Repetition		Phrase Repetition		Prolongation	Hesitation	Syllable Repetition
Grammar	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Recall	Recall
$0^P0^W0^S$	4/6	4/6	6/7	6/7	0/2	0/2	4/12	0/1	0/1
$0^P0^W2^S$	3/6	3/5	0/7	--	0/2	--	1/12	0/1	0/1
$0^P2^W0^S$	0/6	--	3/7	3/4	0/2	--	1/12	0/1	0/1
$2^P0^W0^S$	0/6	--	0/7	--	0/2	0/2	0/12	0/1	0/1
$0^P2^W2^S$	3/6	3/5	3/7	3/4	0/2	--	2/12	0/1	0/1
$2^P0^W2^S$	3/6	3/5	0/7	--	0/2	0/2	1/12	0/1	0/1
$2^P2^W0^S$	0/6	--	3/7	3/4	0/2	0/2	1/12	0/1	0/1
$2^P2^W2^S$	3/6	3/5	3/7	3/4	0/2	0/2	2/12	0/1	0/1

Table 4: Detection rates for grammars with the disfluency class of one or more components modeled by the optional oracle. (Precision rates are not available for inactive components, and also are not available for the disfluency classes that the system does not currently model, under the heading Others.)

The results in Table 4 suggest that the system is still using one class of disfluency to detect another. For example, grammar $2^P2^W2^S$ detected two of the 12 prolongations. These results seem surprising, since by definition of the optional oracle grammars, the probability of detecting a disfluency with a different class label than the annotated disfluency is 0. Closer inspection reveals that the detection of the additional disfluencies occurred during compound disfluencies. Two of the annotated disfluencies are compound: one is a prolongation of a sound that is then repeated, and the other is a prolongation in a word that is then repeated. Consequently, when grammar $0^P0^W2^S$ detected the sound repetition correctly, it received extra credit for the overlapping prolongation, and grammar $0^P2^W0^S$ detected the word repetition correctly, it received extra credit for the overlapping prolongation. With the exception of compound disfluencies, then, the restrictions of the optional oracle grammars prevented the system from using one disfluency class to indicate another.

Optional Oracle Grammars - Classification						
Grammar	Sound		Word		Phrase	
	Recall	Precision	Recall	Precision	Recall	Precision
$0^P0^W0^S$	4/6	4/6	6/7	6/7	0/2	0/2
$0^P0^W2^S$	3/6	3/5	--	--	--	--
$0^P2^W0^S$	--	--	3/7	3/4	--	--
$2^P0^W0^S$	--	--	--	--	0/2	0/2
$0^P2^W2^S$	3/6	3/5	3/7	3/4	--	--
$2^P0^W2^S$	3/6	3/5	--	--	0/2	0/2
$2^P2^W0^S$	--	--	3/7	3/4	0/2	0/2
$2^P2^W2^S$	3/6	3/5	3/7	3/4	0/2	0/2

Table 5: Classification rates for grammars with the disfluency class of one or more components modeled by the optional oracle.

The classification rates shown in Table 5 indicate how well the system was able to classify the disfluencies with the Optional Oracle grammars. In comparison to the Regularities grammars, the precision rates improved while the recall rates for each class either remained the same or improved slightly. Note that the values do not change as more components are activated, unlike the classification rates for the Regularities grammars. Since the system does not use the incorrect disfluency class to indicate the location of disfluent speech, the system produces the same results even as more components are activated in the grammars.

7 Conclusion

The main strength of our system was that it captured many details about each detected disfluency, including its class and exact location. The fine-grained measures of detection and classification rates, more than the course-grained WER, allowed for a thorough understanding and analysis of how the system performed with each grammar. Analyzing the details revealed that the system often inserted a disfluency from another class to account for disfluent speech if the correct class was unavailable for recognition, but rarely inserted a disfluency from another class if the correct class was available. The system's preference for a hypothesis that included a disfluency, even the incorrect class of disfluency, over a fluency hypothesis indicates that the two tasks of disfluency detection and speech recognition are intertwined, and performing both at the same time results in improved performance at each task.

Another central feature of our system was that it allowed us to activate and deactivate each component of a grammar. Modeling the disfluency classes separately and in combination allowed us to see the effects of each component individually. We were able to determine that the sound repetitions were the most difficult to detect accurately. Furthermore, by activating one, two, then three components, we were able to see the interaction between the tasks of detecting each disfluency class. This interaction makes the argument that the tasks should not be separated; systems that include each disfluency class in a single model should perform better than systems that model the disfluency classes separately.

Although the uniform-weight grammars succeeded at modeling some of the regularities of the disfluencies of stutterers, there were other factors affecting the performance of the system. The grammars were negatively affected by a high rate of false positives, particularly for sound repetitions. A probabilistic language model with more accurate transition probabilities would better capture and predict the regularities of repetitions and reduce the number of false positives identified by the system. However, even a language model with perfect prediction would not be sufficient, as shown by the 4% word error rate for grammar $0^P0^W0^S$ in Section 6.1.1. Improved acoustic and prosodic models, trained on stuttered speech in order to model some of its unique characteristics, should improve performance on the word-alignment task. Improved word alignment would in return improve performance on the detection and classification tasks since these tasks interact with the speech recognition task.

8 Future Work

One of our next steps is to gather more data, since the results on one 3-minute sample of speech are not conclusive. However, the system consistently performed poorly despite the use of generous models, so we are confident that our conclusions will remain the same after examining more data: uniform-weight grammars do not capture all of the regularities necessary to detect and classify stuttering events.

Once we have collected enough data, we will train acoustic and prosodic models of stuttered speech. With an acoustic model trained on stuttered speech, we will be able to expand the system to model prolongation disfluencies. We will also implement a probabilistic language model for stuttered speech. We plan to continue using the read speech task to minimize speech recognition complexities and to allow us to explore more complex models of the disfluencies of stutterers.

9 References

Banerjee, S., J.E. Beck and J. Mostow. (2003a). Evaluating the Effect of Predicting Oral Reading Miscues. In *Proceedings of Eurospeech*, September 2003a, Geneva, Switzerland.

Banerjee, S., J. Mostow, J. Beck and W. Tam. (2003b). Improving Language Models by Learning from Speech Recognition Errors in a Reading Tutor That Listens. In *Proceedings of the International Conference on Applied Artificial Intelligence (ICAAI)*, December 2003b, Pune, India.

Cordes, A.K., P.F. Ingham and J.C. Ingham. (1992). Time-Interval Analysis of Interjudge and Intrajudge Agreement for Stuttering Event Judgments. *Journal of Speech and Hearing Research*, *35*(3): 483-494.

Gregory, H.H., J.H. Campbell, C.B. Gregory and D.G. Hill. (2003). *Stuttering Therapy: Rationale and Procedures*. Boston, Pearson Allyn & Bacon.

Heeman, P.A. (1999). Modeling Speech Repairs and Intonational Phrasing to Improve Speech Recognition. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 1999, Keystone, Colorado.

Heeman, P.A. and J.F. Allen. (1999). Speech Repairs, Intonational Phrases and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. *Computational Linguistics*, *25*(4).

Heeman, P.A., F. Yang and S.E. Strayer. (2002). Dialogueview: An Annotation Tool for Dialogue. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, July 2002

Howell, P., S. Sackin and K. Glenn. (1997). Development of a Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children who Stutter: II. ANN Recognition of Repetitions and Prolongations with Supplied Word Segment Markers. *Journal of Speech, Language, and Hearing Research*, *40*(5): 1085-1096.

Johnson, W., R.M. Boehmler, W.G. Dahlstrom, F.L. Darley, L.D. Goodstein, J.A. Kools, J.N. Neeley, W.F. Prather, D. Sherman, C.G. Thurman, W.D. Trotter, D. Williams and M.A. Young. (1959). *The Onset of Stuttering; Research Findings and Implications*. Minneapolis, University of Minnesota Press.

Johnson, W. and S.F. Brown. (1935). Stuttering in Relation to Various Speech Sounds. *Quarterly Journal of Speech*, *21*: 481-496.

Kully, D. and E. Boberg. (1988). An Investigation of Interclinic Agreement in the Identification of Fluent and Stuttered Syllables. *Journal of Fluency Disorders*, *13*(5): 309-318.

Logan, K.J. and L.R. LaSalle. (1999). Grammatical Characteristics of Children's Conversational Utterances That Contain Disfluency Clusters. *Journal of Speech, Language, and Hearing Research*, *42*(1): 80-91.

Mostow, J., J. Beck, S.V. Winter, S. Wang and B. Tobin. (2002). Predicting Oral Reading Miscues. In *Proceedings of ICSLP*, Casual Productions, 2002, Denver, CO.

Nöth, E., H. Niemann, T. Haderlein, M. Decher, U. Eysholgt, F. Rosanowski and T. Wittenberg. (2000). Automatic Stuttering Recognition Using Hidden Markov Models. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2000, Beijing, China.

Riley, G. (1994). *Stuttering Severity Instrument for Children and Adults*. Third. Austin, TX, Pro-Ed.

Stolcke, A., E. Shriberg, D. Hakkani-Tür and G. Tür. (1999). Modeling the Prosody of Hidden Events for Improved Word Recognition. In *Proceedings of Eurospeech*, 1999, Budapest, Hungary.

Sutton, S., R. Cole, J. de Villiers, J. Schalkwyk, R. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, M. Massar and M. Cohen. (1998). Universal Speech Tools: The Cslu Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, November 1998, Sydney, Australia.

Throneburg, R.N. and E. Yairi. (1994). Temporal Dynamics of Repetitions During the Early Stage of Childhood Stuttering: An Acoustic Study. *Journal of Speech and Hearing Research*, *37*(5): 1067-1075.

Wingate, M.E. (1964). A Standard Definition of Stuttering. *Journal of Speech and Hearing Disorders*, *29*: 484-489.

Wingate, M.E. (1977). Criteria for Stuttering. *Journal of Speech and Hearing Research*, *20*(3): 596-607.

Yaruss, J.S. (1997). Clinical Measurement of Stuttering Behaviors. *Contemporary Issues in Communication Science and Disorders*, *24*: 33-44.

Yaruss, J.S. (2001). Evaluating Treatment Outcomes for Adults who Stutter. *Journal of Communication Disorders*, *34*: 163-182.

Yaruss, J.S., M.S. Max, R. Newman and J.H. Campbell. (1998). Comparing Real-Time and Transcript-Based Techniques for Measuring Stuttering. *Journal of Fluency Disorders*, *23*(2): 137-151.