Learner Outcomes

Explain the tradeoff between real-time and transcript-based disfluency counts.
Describe different types of disfluencies that they might want to count.
Describe how automatic speech recognition can speed up transcript-based disfluency counts.

Abstract

Clinicians often do real-time disfluency counts of read-speech. We are developing computer tools to improve the quality and usefulness of these counts. These tools merge the clinicians’ real-time disfluency counts with the output of an automatic speech recognizer, producing an annotated verbatim transcript, which can be easily corrected and reviewed.

Motivation

Stuttering is a communication disorder characterized by disfluencies that are frequent and disruptive to communication. Speech-Language Pathologists (SLPs or clinicians) use disfluency counts to decide whether a client should be treated, to assess treatment progress, and to document treatment outcomes (Yaruss, 1997; Yaruss, 1998; Conture, 2001). They often do disfluency counts in real-time as a client is talking. However, these are not very specific as we do not know what words are involved. Moreover, they cannot be re-examined in order to improve quality or assess validity. SLPs can also use a verbatim transcript approach, in which they first transcribe exactly what was said, and then mark up the transcript with disfluency codes (Bernstein Ratner, Rooney, and MacWhinney, 1996). For example, in Systematic Disfluency Analysis (SDA), there are separate codes for sound, word and phrase repetitions, interjections, revisions, prolongations, blocks and others (Gregory et al., 2003; Campbell, Hill, and Driscoll, 1991). This approach allows more detailed and accurate counts to be obtained.

Unfortunately, few tools exist to assist SLPs with disfluency counts. Automatic approaches to transcribe stuttered speech and count disfluencies have been attempted, but Automatic Speech Recognizers (ASRs) have problems even with fluent speech. The only computer tools that have proven useful are ones that allow an audio file to be played and annotated. However, these tools are not optimized for the task of annotating disfluencies, and cannot be used to annotate disfluencies in real-time. Transcribing the actual words is even more time consuming. Thus, SLPs almost exclusively use real-time counts that are not aligned to a word transcription.

Our goal is to build computer tools that will help SLPs collect disfluencies counts that have the quality and usefulness of transcript-based counts, but at a time-cost closer to real-time counts. As is common practice, SLPs will perform real-time disfluency counts during a therapy session, as the client is speaking. We will use these counts, along with automatic speech recognition, to create an annotated verbatim transcript, with the disfluency codes aligned to the appropriate words in the word transcription. The annotated verbatim transcription can then be reviewed and corrected.

For the current version of the tools, we are focusing on read-speech, where the client is reading a story out loud. Such speech samples are regularly used in stuttering diagnosis and treatment (Riley, 2009). Furthermore, read-speech allows us to leverage the highly constrained nature of the speech to have usable ASR output, even for stuttered speech. Ultimately, as we improve how ASRs model stuttered speech, we will move to less constrained speech, such as picture description and describing an activity.

This study was funded by the NIH under grant 2R42DC009944-02 to BioSpeech, where Drs. Heeman and Lunsford are employees. This potential conflict of interest has been reviewed and managed by OHSU.
Four Step Process

Step 1: Annotation Pass. The SLP, using our touch-screen annotation tool, counts the disfluencies in real-time as the client is speaking, using 8 disfluency codes. The tool creates an audio-video recording as well as time-aligned disfluency codes.

Step 2: Automatic Speech Recognition. The ASR is run on the audio file. The ASR employs an acoustic model, which maps the audio signal to phonemes and then to words, and a language model, which captures the story text, and common disfluency patterns. It uses these to find the most probable word transcription. The ASR also keeps track of alternate word transcriptions and how likely they are, in a word lattice. A small portion of a word lattice is shown in Figure 1. Following the arrows, one can see that the ASR is considering fluent (e.g., ‘lick the stitches’) and disfluent ones (e.g., ‘like like the stitches’, ‘like the st-stitches’).

Step 3: Merge Clinician’s Annotations with ASR output. A computer program searches through the word lattice to find one that is consistent with the clinician’s disfluency codes. This search takes into account that some of the clinician’s disfluency codes might be wrong and that they tend to lag the actual disfluency by several seconds. The output is an annotated verbatim transcript, which is an exact transcription of the words that were said, with the disfluency codes matched up to the words that are involved in the disfluency.

Figure 2 shows the tool that SLPs will use in the next step; but it also shows the actual output from Step 3. The 3 disfluencies are shown where the computer aligned them to the transcript; the horizontal line shows where the clinician annotated them. The first disfluency is a sound repetition (Rs that the clinician annotated at 38.3s. A plausible sound repetition was not found in the lattice, so the computer aligned it based on average lag times for SLPs in annotating sound repetitions. The second disfluency is a phrase repetition (Rp). The computer did correctly recognize the words involved, and placed the annotation code at the correct spot. The third disfluency is a block (B), which was correctly placed, due to the preceding silence.

Step 4: Review and Correction Pass. The SLP uses our touch-screen correction tool to review and correct the annotated verbatim transcript, as shown in Figure 2. The tool displays the waveform of the entire audio file, and shows the word transcription and disfluency codes time-aligned to the waveform. The tool also displays the video, which can be resized to fit the clinicians’ needs. By zooming in on the waveform, the SLP can read the transcript and see the disfluency codes as the waveform is playing. Any time the SLP thinks there might be an error in the word transcription or disfluency annotations, the SLP can stop playing the file, highlight and re-listen to the problematic
region, and correct any errors. For the disfluency codes, the SLP can add, change, or delete a code, or change its alignment to the word transcription.

If the word transcription is wrong, there is a good chance that the correct transcription is in the word lattice. The correction tool allows the SLP to easily *navigate* the lattice to see if the correct transcription is present, as illustrated in Figure 3. In the first panel, the user clicked on ‘lick’ and gets the four different words in the lattice that can follow that word. In the second panel, the user clicked on ‘the’, and is then shown the words that follow ‘lick the’. If the lattice does not include the correct word transcription, the SLP can add, change, or delete individual words.

Using an earlier version of the correction tool, we found that SLPs can review and correct the verbatim transcript in just 2 times real-time (one minute of audio takes 2 minutes to correct). In a questionnaire, several SLPs commented that this time was well spent, as they are reviewing critical areas of the client’s speech.

**Analysis**

We built a tool that summarizes the information in the annotated verbatim transcriptions. The first graph in Figure 4 compares the disfluencies counts (over at least 15 minutes of speech) differ between four people who stutter, by type of disfluency. In a clinical setting, this tool can show how a single client progresses over time, with a bar for each session. The second graph shows the power of transcript-based counts, as the duration of time for each disfluency type is shown.

The tool can also show how disfluencies occur over time, as shown in Figure 5. Color is used to show Disfluency type, while bar height and width show the duration of each disfluency. This picture shows subject13 disfluency patterns vary more over time than the other 3. Timelines can also be used to show how a single client varies over time.

**Conclusion**

In conclusion, these tools should help SLPs perform more detailed and more reliable disfluency counts, while not requiring an excessive expenditure of additional time. We hope that the resulting richness of clinical data can contribute meaningfully to the diagnosis and treatment of stuttering.

**References**


