TOWARDS THE AUTOMATIC EVALUATION OF FLUENCY

John Morgan
United States Military Academy
Department Of Foreign Languages
745 Brewerton Road
West Point, NY 10996
Email: john.morgan@usma.edu
Phone: 845-938-5329
Fax: 845-938-3585

Martha Gallagher, Ph.D.
United States Military Academy
Department Of Foreign Languages

Garrett Molholt Ph.D.
West Chester University Of Pennsylvania

ARL Sponsor: Mellissa Holand Ph.D.
U.S. Army Research Laboratory
Adelphi, Maryland

Stephen LaRocca Ph.D.
U.S. Army Research Laboratory
Adelphi, Maryland

October 18, 2004

1 ABSTRACT

Although the definition of fluency is not universally agreed upon, fluency is commonly accepted as one of the most important goals in foreign language instruction . The automatic evaluation of fluency would be useful to the Army, because it would facilitate the task of testing large numbers of linguists whose pay depends on the levels they achieve on oral proficiency tests. We also propose that an oral presentation tutor can be developed with current technology that would provide an environment in which students could practice reading outloud and receive feedback on their fluency.

In order to evaluate fluency, an Automatic Speech Recognition system (ASR) is configured to output duration data of speech segments. The duration data are used to calculate measures such as the pace, articulation rate, and phonation rate. Previous studies have shown that these measures of speaker fluency correlate well with expert ratings. This paper reports on baseline measurements for native Mandarin Chinese . Informal results comparing the application of the fluency measures to a small set of non-native speech with the native baseline are presented. A project for implementing fluency measures into software for foreign language learning is also outlined.

The results for native speakers' fluency measures fall in a relatively narrow band and are almost completely disjoint from the results obtained from the small set of non-native speakers. Anecdotal perception estimates correspond to the measurements given automatically.

2 INTRODUCTION

Computer Assisted Language Learning (CALL) applications commonly use ASR to provide pronunciation feedback at either the segment or word level. For example, in software packages like Auralogue's "TeLL Me More", the student is prompted to repeat a word or phrase and the ASR returns a score or some other type of feedback on the quality of the student's pronunciation. Other applications use ASR as an interface to a speech interactive environment. In the Tactical Language Training System [4]for instance, the video gaming environment "Unreal Tournament" has been adapted to accept speech input. In this system the goal is to simulate conversational speech in the context of a foreign culture and to rapidily aquire communicative skills. The Computer Science and Artificial Intelligence Laboratory (CSAIL)

at MIT has worked for a decade on robust dialogue systems for use in telephony applications such as travel reservation systems. Researchers at CSAIL [5]are taking advantage of this work to improve foreign language teaching. When applied to foreign language learning situations these dialogue systems give students practice at producing intelligible speech in the target language that is appropriate in certain scenarios.

These three types of systems make use of speech in limited quantities, the student is rarely required to speak more than a sentence at a time. In this paper we discuss a way to expand the use of ASR in CALL to speaking practice for larger chunks of text including sentences and paragraphs.

As part of most foreign language learning classroom activities, students are asked to make oral presentations to their classmates on cultural topics using new vocabulary and grammar . A critical part of achieving success in these presentations is speaking fluently. Hincks [3]describes an oral presentation practice checker that would listen to a student's spoken rendition of a written text and evaluate his or her "liveliness" along with the traditional checks for grammar and spelling. We Additionally envision the application providing feedback on rate of speech and other fluency measures.

Work on the automatic measurement of fluency has been done by researchers in Holland [1]for students learning dutch as a foreign language. Having reviewed the definition of fluency from several different perspectives, the dutch researchers proceed to show that regardless of the definition of fluency, the objective temporal measures obtained from an ASR system produce fluency assessments that are highly correlated with fluency ratings given by human experts. In the current paper we follow closely their methods. Specifically, we focus on rate of speech results since they found that these correlate best with the ratings from experts.

The fluency assessments we use are obtained from the measurement of the durations of speech segments including silent pauses. We obtained best results using context-independent phones as our speech segments. Time-aligning the endpoints of speech segments is an essential part of state of the art ASR, thus existing systems can be easily adapted to generate duration data for speech segments and hence fluency measurements .

We worked with a large data base of digital recordings of Mandarin Chinese speech from natives reading text from various sources. So far our work has concentrated on getting baseline measurements for the variables that are used to make fluency assessments. Part of this work involved detecting and removing "corrupt" data. That is, some of the recordings contained extra-

neous noise, whispering, complete silence etc. One method we used to find these corrupt data was to look for those speech segments that yielded extreme duration measurements, then we listen to them to decide whether or not they should be discarded.

Our non-native corpus consists of 6 speakers who read the same 50 phonetically rich prompts that were also read by all the native informants. We performed the same fluency measurements on these data and compared them with the native baseline measurements.

3 Methods

We used the Chinese Academy of Social Sciences (CASS) corpus to estimate statistics for variables that are associated with the assessment of a speaker's fluency. The CASS corpus is a large database of digital recordings of read speech from native speakers of Mandarin Chinese. Since Mandarin is a tonal language and since the tone is carried by the rhyme segment of the syllable, we initially made acoustic models of onsets and rhymes with the rhymes marked for five tones. We also used the standard speech frame encoding using a window size of 0.25 seconds and step size of 0.1 seconds. We noticed that in estimating the duration variable the standard deviation was often higher than the mean value for onset and rhyme segments. We also noticed that the alignments were not very accurate for silence and reduced segments. We were able to quickly get results for sevearal other languages including Arabic and Portuguese using monophone acoustic models. We obtained lower values for standard deviation with respect to the mean estimates for these other languages. So we made two modifications to our acoustic models.

1. We use toneless monophones instead of onsets and tone-marked rhymes as our speech segment models,

2. we made the speech frame encoding denser by using a 0.125 second window size and a step of 0.05 seconds.

By checking some data by hand we verified that the duration measurements were much improved by the new speech segment modeling and speech frame encoding.

Once we got the time alignments from the ASR system for each utterence we removed the initial and final silence from the transcriptions. These

silence segments are not consistently related to a speaker's performance of an utterence.

Then we calculated statistics for the following variables:

Articulation rate	$\frac{\text{Number of segments}}{\text{dur1}}$
Rate of speech	$\frac{\text{Number of segments}}{\text{dur2}}$
Phonation/time ratio	$100\% \frac{\text{dur1}}{\text{dur2}}$
Mean length of silent pauses	Mean length of all silent pauses
Duration of silent pauses	$\frac{\text{Total duration of all silent pauses}}{\text{dur2}}$
Number of silent pauses	$\frac{\text{Number of silent pauses}}{dur2}$

Table 1: Fluency variables.

1. silent pause: a stretch of silence,
2. dur1: duration of speech without utterance internal silences,
3. dur2: duration of speech including utterance internal silences.

3.1 Materials

The CASS corpus has a total of 315 speakers 244 of which were used in calculating the baseline. Some of the remaining 71 speakers' recordings were intentionally performed in noisy environments. Other speakers were left out to be used later as test data. The entire CASS corpus has approximately 100 hours of speech.

The non-native corpus we used comes from 4 cadets and 2 of their USMA Chinese instructors. All of them read the first 50 promptsthat were read by all the native informants.

3.2 Tools and Hardware

The Hidden Markov Model toolkit (HTK) alpha release of version 3.3. was used for data preparation, acoustic model training , and decoding.

A Beowulf class supercomputer consisting of a cluster of 7 workstations was used for the data preparation, training and decoding. The machines are Intel Pentium 1200 MHz dual processors with 500 MB of RAM running Redhat Linux 8.1. The machines are networked via a gigabyte hub.

speakers	ros	mean	sd
native	7.52	0.133	0.199
non-native	4.25	0.235	0.237

Table 2: Mandarin Chinese rate of speech using onset-rhyme acoustic models.

4 RESULTS

First, we modeled the onset and rhyme parts of the Mandarin Chinese syllable and we obtained the results in tables 2 and 3. The rate of speech (ros) was calculated by dividing the number of onsets and rhymes by the duration of all the segments, onsets, rhymes, and silence. The data under the columns labeled "mean" and "sd" gives the value of the mean and standard deviation for the duration of all the segments. The rate of articulation (roa) was calculated by dividing the number of onsets and rhymes by the duration of all the segments except silence.

Second, we changed our acoustic models from onsets and tone marked rhymes to toneless monophones and we made the speech frames shorter and more frequent. We got the results in tables 4 and 5.

Phonation is another indicator of fluency. It is calculated by dividing the duration of the segments without silence by the duration of the segments including silence. The baseline results we calculated are in table 6.

The ros and roa results for the 6 non-native speakers are in tables 7, 8, 9, and 10. The tables captioned "Army Linguists" shows data from 2 West Point Chinese instructors and the other tables contain data from their cadets.

5 DISCUSSION

This study describes baseline results that will be used later in experiments to test the effectiveness of tools designed to practice fluent speech. In further research we propose to establish baseline norms regarding fluency measures for other languages. We would also like to investigate how learners from one language move toward fluency in another language.

The rate of speech (ros) measures appear to show that fluency is an aspect of language learning that could benefit from a fluency checker. The highest ros value in our small non-native corpus was 8.084 phones per second registered by an experienced Army linguist. The very slowest speaker in the corpus of baseline native speakers had an ros of 9.20 phones per second.

Following are Some interesting research questions that arise from these baseline results:

- Does faster ros usually lead to faster phonation?
- Does faster ros usually lead to less dynamic pitch variation?
- Which parameters are most salient for language learners? and
- How can the most salient parameters be displayed so that learners grasp the concepts easily and put them to work?

5.1 Caviats

The text source of the data is a weakness in this study. The informants in the corpora we used mostly read short sentences. we would like to collect speech from speakers reading long chunks of text. Idealy, we would use speech collected from real oral presentations given in language learning contexts. Informants could for example read text they themselves had written.

speakers	roa	mean	sd
native	7.65	0.131	0.197
non-native	4.28	0.234	0.238

Table 3: Mandarin Chinese rate of articulation using onset-rhyme acoustic models.

speakers	ros	mean	sd
native	12.87	0.078	0.056
non-native	6.49	0.154	0.141

Table 4: Mandarin Chinese rate of speech using toneless monophone acoustic models and denser and shorter frames in the speech encoding.

speakers	roa	mean	sd
native	13.38	0.075	0.050
non-native	7.579	0.132	0.086

Table 5: Mandarin Chinese rate of articulation using monophone acoustic models and denser and shorter frames in speech encoding.

speakers	phonation
native	93.76
non-native	76.47

Table 6: Mandarin Chinese phonation/time ratio using monophone acoustic models and denser and shorter frames in speech encoding.

speakers	ros	mean	sd
2	6.442	0.155	0.117
5	8.084	0.124	0.061

Table 7: Army linguists Mandarin Chinese rate of speech using toneless monophone acoustic models.

speaker	roa	mean	σ
2	7.192	0.139	0.080
5	8.302	0.120	0.075

Table 8: Army linguists Mandarin Chinese rate of articulation using monophone acoustic models.

Speaker 7 is a good example of how the duration variables can indicate fluency problems. Notice in tables 9 and 10 the difference in the standard deviation between the rate of speech and the rate of articulation. This occurs because the silence segments are much longer than the duration of the monophone segments. So, even though the runs of speech sound well pronounced, the speaker doesn't sound fluent because of the long inturruptions by segments of silence.

In order to conduct a valid experiment on the performance of a fluency checker, the data used for testing must be different from the data used to obtain the baseline measurements. Thus the results obtained for the non-native data are not valid as an evaluation of a future fluency checker since the non-native informants read the same prompts as the native speakers.

The future fluency checker will likely use the sonic system to obtain segment alignments, so it should be incorporated in future experiments.

In our calculations we removed initial and final silence segments. We note that this might not be possible in a real time application and this could have an effect on the performance of a fluency checker.

When we changed our speech segment modeling to toneless monophones we paid a price. We'd like to investigate the differences in fluency between natives and students that occur as a result of tonal contexts or sandhi. This is not a show stopper however, since we can run separate recognizers in separate threads using tonal and context dependent acoustic models respectively.

The dutch research team found that their methods were not as successful with spontaneous speech.[2]

References

[1] C Cucchiarini, H Strik, and L Boves. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *J. of the Acoustical Society of America, Vol. 107 (2)*, pages 989–999, 2000.

[2] C Cucchiarini, H Strik, and L Boves. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *J. of the Acoustical Society of America, Vol. 107 (2)*, pages ?–?, 2002.

[3] R. Hincks. Tutors, tools, and assisstants for the l2 user. *Fonetik*, pages 173–176, 2003.

[4] W. L. Johnson, S. Marsella, N. Mote, H. Viljhalmsson, S. Narayanan, and S. Choi. Tactical language training system: Supporting the rapid acquisition of foreign language and cultural skills. In *Proceedings Of InSTILL/ICALL 2004- NLP And Speech Technologies In Advanced Language Learning Systems*, Venice, Italy, 2004.

[5] S Seneff, C Wang, and J Zhang. Spoken conversational interaction for language learning. In *Proceedings Of InSTILL/ICALL 2004- NLP And Speech Technologies In Advanced Language Learning Systems*, Venice, Italy, 2004.

speaker	ros	mean	sd
1	7.057	0.142	0.106
4	5.954	0.168	0.155
6	7.298	0.137	0.113
7	5.101	0.196	0.214

Table 9: Cadet Mandarin Chinese rate of speech using toneless monophone acoustic models.

speaker	roa	mean	sd
1	7.859	0.127	0.086
4	7.092	0.140	0.097
6	8.494	0.118	0.073
7	6.836	0.146	0.100

Table 10: Cadet Mandarin Chinese rate of articulation using monophone acoustic models.