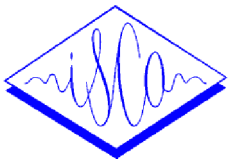
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Audio Quality Issue for Automatic Speech Assessment

Lei Chen

Educational Testing Service
Princeton, NJ, USA
LChen@ets.org

Abstract

Recently, in the language testing field, automatic speech recognition (ASR) technology has been used to automatically score speaking tests. This paper investigates the impact of audio quality on ASR-based automatic speaking assessment. Using the read speech data in the International English Speaking Test (IEST) practice test, we annotated audio quality and compared scores rated by humans, speech recognition accuracy, and the quality of features used for the automatic assessment under high and low audio quality conditions. Our investigation suggests that human raters can cope with low-quality audio files well, but speech recognition and the features extracted for the automatic assessment perform worse on the low audio quality condition.

1. Introduction

Speaking proficiency is an important component of the measurement of language skills. Speaking tests have been used in well-known standardized language tests, e.g., TOEFL and IELTS. In these tests, test takers' speech is recorded and scored by human raters. In the past decade, automatic speech recognition (ASR) has been used to automatically score speaking tests [1, 2, 3, 4]. For example, Ordinate provided a telephone-based speaking test, PhonePass [1]. Educational Testing Service (ETS) has been conducting research on scoring non-native spontaneous speech [5]. This research has resulted in a product, SpeechRater℠ [6], which was used to automatically score low-stakes test responses in the online practice test for TOEFL®. Recently, Pearson made public plans for a new test, Pearson Test of English (PTE), to be on the market in Fall of 2009. PTE test is a fully automated test used for high-stakes purposes, i.e., admitting international students into English-medium colleges and universities. PTE will use ASR to automatically score speaking responses [7].

Compared to the traditional way of manually scoring speech by human raters, the ASR-based automatic speech assessment is cheaper, faster, and less influenced by raters' uncontrolled variations, e.g., changes in emotion. However, ASR is still worse than human hearing in many ways. For example, ASR systems are less robust than human beings when facing audio files with a poor quality. Unfortunately, speech responses collected in speaking tests may have low quality for various reasons, e.g., using low quality microphones, setting up recording software incorrectly, and interferences from other test takers. Therefore, low quality audio files challenge the ASR-based speaking scoring system.

As described in Section 2, some previous research investigated the impact of audio quality on human raters and on accuracy of ASR. However, an investigation of the impact of audio quality on the ASR-based speaking assessment is still missing. In this paper, we report an investigation on this topic.

The remainder of paper is organized as follows: Section 2 describes the related research; Section 3 describes the data used in our experiment and the audio quality annotation; Section 4 reports on the speech recognizer used in our experiment and the extraction of speech features; Section 5 reports on our experimental results; Section 6 discusses our findings.

2. Related Work

McNamara and Lumley investigated the effect of degree of audibility of audio on human scoring [8]. On 142 audio tapes, they annotated the degree of audibility and found 61% of tapes were perceived as perfectly audible and the remaining 39% of tapes were perceived as imperfectly audible. They found that the tapes perceived as imperfectly audible were rated more harshly than if these tapes were judged to be perfectly audible. Accordingly, they gave several suggestions: (1) care needs to be taken to ensure that recordings are of the highest quality and (2) audio quality should be considered as a factor in the scoring process to neutralize its impact.

The performance of most current ASR systems degrades significantly when environmental noise occurs. Such performance degradation is mainly caused by mis-matches in training and operating environments [9]. Lippmann compared speech recognition done by human and machine [10]. He found that error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wide-band, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech.

For speaking tests, some recorded speaking responses may have low audio quality. For example, for internet-based practice tests, test takers use their own computers and microphones to record speech responses. Improper set-up of recording devices or using microphones with a poor quality cause low-quality audio samples. For tests held in testing centers, overlapped speech from other test-takers may also cause low-quality audio samples. Because poor audio quality challenges ASR systems, with an expanding use of ASR-based speaking assessment in language testing, we need investigate audio quality's impact on the ASR-based speaking assessment.

3. Data and Audio Quality Annotation

The International English Speaking Test (IEST) measures English oral communication skills in international business and professional settings. To help test-takers better prepare for this test, An on-line practice test that uses retired IEST Speaking test material is provided. Currently, this test is scored by trained human raters.

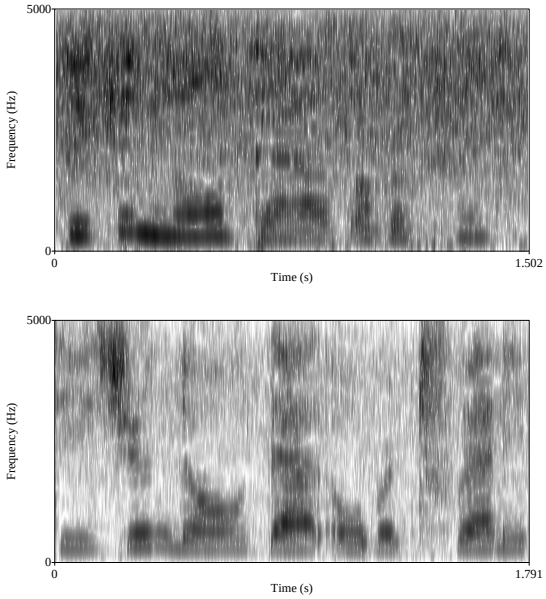


Figure 1: A comparison of spectrograms of two audio portions based on identical read content but with *poor* quality (on the top) and *satisfactory* quality (on the bottom)

The IEST speaking test includes reading-aloud tasks that require test-takers to read a short paragraph of 40-60 words aloud. The reading materials include announcements, advertisements, introductions, and so on. The read-aloud tasks are rated analytically on pronunciation and intonation on a 3-point scale.

The audio data used in our experiment were collected from previous IEST practice tests. From four test forms (corresponding to four different reading material), $1,065$ audio samples were obtained. All these audios were scored by human raters.

Some subjective and objective standards (e.g., PESQ [11]) have been designed to rate audio quality for measuring the fidelity of synthesized speech and the effect of speech enhancement. However, these standards do not fit our audio quality annotation task. One of the reasons is that these standards focus on differentiating subtle changes in audio quality, which is not the focus of our audio quality annotation. In addition, rating audio quality by following these standards requires using special equipment and collecting reference sounds. These are not available in our test assessment scenario. Therefore, we used a simple binary coding method designed based on our observation of the IEST data.

We rated audio quality as *satisfactory* and *poor*. The audio file rated as *satisfactory* has no or a few of problems, e.g., clipping, background noise, too soft or too loud sounds, etc, in recording. In contrast, the audio file rated as *poor* has problems in recording. Figure 1 displays spectrograms of two audio portions based on identical read content but with different audio quality ratings. Noise in the *poor* audio smears high frequency components of the audio signal and tends to negatively impact speech recognition that uses these high frequency components.

Two native American English speakers annotated these 1065 audio samples after their coding standards were calibrated. On the 200 audio files that were randomly selected from the data set, the two raters conducted a parallel annotation. Their annotations have an agreement of 90.00% and a Cohen's κ of 0.618. This indicates that an acceptable annotation consistency can be achieved between human raters. The author reviewed some of the two raters' annotations and found that one rater who had experience on audio quality rating before had a more accurate annotation. Therefore, on the kappa set, we used the annotation results provided by this rater. Among 1065 audio files, 938 files were rated as *satisfactory* and 127 files (11.92%) were rated as *poor*.

4. Speech Recognition and Feature Extraction

4.1. Speech recognition

As described in [12], a speech recognizer was used in recognizing speech and in forced aligning speech according to its hypotheses. Because the amount of read speech data from the IEST practice test is limited, we used external data resources to expand the data used for acoustic model (AM) and language model (LM) training.

Two different AMs were used in the recognition and forced alignment steps, respectively. The AM used in the recognition was trained on about 30 hours of non-native speech from the TOEFL®Practice Online (TPO) corpus [12]. Building on this recognizer, an MAP AM adaptation was conducted on the read data to reflect the phonetic patterns of IEST read speech. Because the content of the reading-aloud responses was constrained, a dictionary (with roughly 2,000 words) containing frequently used English words and words appearing in the reading-aloud passages was used. The LM was built as follows: a generic reading LM was trained on the Broadcast News (BN) corpus, which contains hundreds of hours of broadcast news read by anchors. Then, the LM trained on the BN corpus (LM_{BN}) was interpolated with the LM trained on the reading-aloud items (LM_{Read}) according to a weight setting of $0.9 * LM_{BN} + 0.1 * LM_{Read}$ to be the LM used in the reading specific recognizer.

The AM used in the forced alignment was trained on native speech and high-scored non-native speech data. It was trained as follows: starting from a generic recognizer, which was trained on a large and varied native speech corpus, we adapted the AM using MAP adaptation on a corpus containing about $2,000$ responses with high scores in previous TPO tests and the TOEFL®Native Speaker Study [12]. Ideally, the AM used in the forced alignment should be trained on read speech data rather than spontaneous speech data. This is in our future plan since the IEST read speech we have collected from native speakers is currently very limited.

4.2. Assessment features extraction

A construct is a set of knowledge, skills, and abilities measured by a test. The construct of the speaking test is embodied in the rubrics that human raters use to score the test. It generally consists of three key categories: delivery, language use, and topic development. Language use refers to the range, complexity, and precision of vocabulary and grammar use. Topic development refers to the coherence and fullness of the response. In practice, most of ASR-based speech assessment systems focus on the delivery given the challenge of recognizing non-native speech. The delivery in turn can be measured on four dimensions: fluency, intonation, rhythm, and pronunciation. For IEST read test,

since the speaking content is provided, the topic development category is not intended to be measured. For language use category, some low-scored test takers may not know some words shown in the reading passage. So, this category can be partially measured.

We extracted the following two types of features, including (1) speech features based on the speech recognition output as described in [13] and (2) pronunciation features that indicate the quality of phonemes and phoneme durations [12]. Among all extracted features, we selected 6 features that were found to be predictive of speaking proficiency. Note that our selected features were generally used for assessing spontaneous speech. Therefore, our investigation done on read speech data can be generalized to spontaneous speech. In addition, some features specific for read speech, e.g., features tightly related to word accuracy, were not used, since the audio quality's impacts on these features can be derived from the impacts on the recognition performance. Table 1 lists names, dimensions, categories in the assessment construct, as well as descriptions of these features. Details of computing these features can refer to [13, 12].

feature	dimension	category	description
wpsec	fluency	delivery	word per second (speaking rate)
tpsec	fluency & vocabulary diversity	delivery & language use	unique words normalized by total word duration
amscore	pronunciation	delivery	acoustic model score from speech recognition
lmscore	gramatical accuracy	language use	language model score
L_6	pronunciation	delivery	average likelihood per second normalized by the rate of speech
$\bar{Sn}$	pronunciation	delivery	average normalized vowel duration shifts

Table 1: A list of speech features used in our experiments

5. Experiments

5.1. Research questions

We intend to answer the following research questions:

- Does audio quality impact human raters' performance?
- Does audio quality impact speech recognition accuracy?
- Does audio quality impact the speech features' predictive ability for assessing speaking proficiency?

On the IEST read data, we used the speech recognizer described in Section 4.1 to recognize all audio files. Then, according to their reference texts (reading content), word accuracy was measured. The word accuracy ($wacc$) is computed as

$$wacc = \frac{1}{2} \times (\frac{c}{c+s+d} + \frac{c}{c+s+i}) \times 100,$$

where $c = \#correct$, $s = \#substitution$, $d = \#deletion$, and $i = \#insertion$. By giving equal weights to the reference and ASR hypothesis, the $wacc$ is unbiased to insertions or deletions. Next, we extracted features for speaking assessment according to the methods described in Section 4.2. Based on two analytic scores provided in the IEST data set, i.e., pronunciation score (pS) and intonation score (iS), an overall score was derived as the average of these two scores ($(pS + iS)/2$)). Finally we conducted statistical analyses to answer these three research questions, using human scores, the recognition results, and the extracted features.

5.2. Results

AQ=1 (N=938)	range	mean	std.
pS	[1-3]	2.25	0.649
iS	[1-3]	2.09	0.624
score	[1-3]	2.17	0.543
AQ=0 (N=127)	range	mean	std.
pS	[1-3]	2.15	0.668
iS	[1-3]	2.01	0.649
score	[1-3]	2.08	0.561

Table 2: Descriptive statistics of human scores under two audio quality conditions.

First, we compared human scores (pS, iS, and score) on audio files with different audio quality ratings. Descriptive statistics (including mean and standard deviation) of human scores under *satisfactory* and *poor* audio quality conditions were reported in Table 2. Variations of human scores have no notable difference under these two audio quality conditions. A following t-test confirms this finding. As reported in Table 3, for each kind of human score, there was no significant difference between two audio quality conditions. However, for the score, the p-value of the significance (0.079) is quite close to the threshold indicating a significant impact (0.05). This suggests that human scoring process is influenced by speech responses' audio quality. However, on our data, such influence is not statistically significant.

value	t	df	Sig. (2-tailed)
pS	-1.621	159.951	0.107
iS	-1.390	159.240	0.166
score	-1.767	159.554	0.079

Table 3: t-test of human scores under two audio quality conditions.

Next, we compared means of word accuracy under two audio quality conditions using t-test. The word accuracy on the *satisfactory* sound condition (M=84.69, SD=10.64) was significantly higher than the word accuracy on the *poor* sound condition (M=74.35, SD=15.06) according to a t-test ($t(1063) = -10.645, p = 0.00$).

At last, we investigated whether such recognition accuracy drop caused by low-quality sounds impacts the speech features' predictive ability for assessing speaking proficiency or not. To answer this question, we compared quality of the extracted speech features under these two audio quality conditions. A widely used metric for measuring feature quality is the Pearson correlation (r) computed between the features and human scores. In our experiment, we will use the absolute value of Pearson correlation with the overall score ($|r|$) to evaluate the features.

Table 4 compares $|r|s$ between each feature to human score under two audio quality conditions. We can find that except $\bar{S_n}$ and *tpsec* features, $|r|$s of *wpsec*, *amscore*, *lmscore*, and L_6

greatly decrease on audio files with *poor* audio quality compared to audio files with *satisfactory* quality.

feature	$\|r\\|_{AQ=0}$	$\|r\\|_{AQ=1}$
wpsec	0.19	0.28
tpsec	0.33	0.35
amscore	0.13	0.27
lmscore	0.23	0.31
L_6	0.008	0.1
$\bar{S_n}$	0.097	0.077

Table 4: Comparison of $|r|$ between features to score under two audio quality conditions

The $|r|$ reduction from *satisfactory* audio to *poor* audio suggests a potential performance drop for the automatic assessment. To further investigate each feature's discriminative ability on prediction of human rated scores, we conducted a one-way between-subjects ANOVA to compare the effect of speaking skill level (according to averaged human scores) on speech features under two audio quality conditions. As shown in Table 5, when $AQ = 1$, for each feature, there was a significant effect of scores on features. However, when $AQ = 0$, for only *tpsec* and *lmscore* features was there a significant effect of scores on features. It is interesting to note that only the features related to language-use still has significant effects between features and scores. This suggests that audio quality degradation has a more serious impact on features related to delivery aspect. An implication of this finding is that for a robust speech assessment system, features representing different aspects of language skill are required.

feature	$F(4, 122)\|_{AQ=0}$	$F(4, 933)\|_{AQ=1}$
wpsec	1.79, p=0.14	21.22, **p=0**
tpsec	4.02, **p=0.004**	34.22, **p=0**
amscore	0.80, p=0.53	20.93, **p=0**
lmscore	2.73, **p=0.032**	25.86, **p=0**
L_6	0.168, p=0.95	7.65,**p=0**
$\bar{S_n}$	1.04, p=0.38	5.40, **p=0**

Table 5: one-way ANOVA analysis on features in two audio quality conditions (bold p value indicates a significant effect between scores on features using $p < 0.05$)

6. Discussion

With the improvement of ASR, a trend toward utilizing ASR to automatically score speaking tests emerges in the language testing field. This paper argues that a major obstacle to using ASR to fully replace human raters, especially for high-stakes test, is that ASR performs poorly when facing low quality audio files. On 1065 audio files collected in the IEST practice test, we annotated audio quality into two levels, i.e., *satisfactory* (AQ=1) and *poor* (AQ=0). Under these two audio quality conditions, human raters showed consistent rating performance. However, the ASR performs significantly worse on *poor* audio quality condition than on *satisfactory* audio quality condition. We also investigated audio quality's impact of the quality of features extracted for the automatic scoring. Compared to features extracted on *satisfactory* audio files, features extracted on *poor* audio files have worse discriminative ability for scoring.

This study demonstrated the negative impact of audio quality on the ASR-based automatic speech assessment. To better utilize ASR for speech assessment, we suggest that (1) audio quality is monitored when recording test-takers' responses to make sure an acceptable quality is achieved, and (2) robust speech recognition [9] technology that is less influenced by audio quality may be utilized to improve the ASR's robustness to noises. In our future research, we plan to investigate using such robust speech recognition technology to cope with noisy responses in the automatic speech assessment.

7. Acknowledgments

The author thanks Ohls Sarah and Waverely Vanwinkle for their hard works on the annotation of audio quality. The author thanks Xiaoming Xi and Klaus Zechner for their help and suggestions on this study. The author also thanks David Williamson for his helpful reviews on this paper.

8. References

[1] J. Bernstein, "PhonePass testing: Structure and construct," Ordinate Corporation, Tech. Rep., 1999.

[2] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[3] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," in *InSTiLL (Intelligent Speech Technology in Language Learning)*, Dundee, Stotland, 2000.

[4] C. Hacker, T. Cincarek, R. Grubn, S. Steidl, E. Noth, and H. Niemann, "Pronunciation Feature Extraction," in *Proceedings of DAGM 2005*, 2005.

[5] K. Zechner and I. Bejar, "Towards Automatic Scoring of Non-Native Spontaneous Speech," in *NAACL-HLT*, NewYork NY, 2006.

[6] K. Zechner, D. Higgins, and X. Xi, "SpeechRater: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech," in *Proc. SLaTE*, 2007.

[7] "Versant english test: Test description and validation summary," Pearson, Tech. Rep., 2008.

[8] T. McNamara and T. Lumley, "The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational i settings," *Language Testing*, vol. 14, no. 2, pp. 140–156, Jul. 1997.

[9] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[10] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–16, 1997.

[11] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, and I., "Perceptual evaluation of speech quality (PESQ)-a new method forspeech quality assessment of telephone networks and codecs," in *Proceedings.(ICASSP'01)*, vol. 2, 2001.

[12] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[13] X. Xi, D. Higgins, K. Zechner, and D. Williamson, "Automated Scoring of Spontaneous Speech Using SpeechRater v1.0," Educational Testing Service, Tech. Rep., 2008.