Improved Pronunciation Features for Construct-driven Assessment of Non-native Spontaneous Speech

Lei Chen, Klaus Zechner, Xiaoming Xi
Educational Testing Service
Princeton, NJ, USA
{LChen,KZechner,XXi}@ets.org

Abstract

This paper describes research on automatic assessment of the pronunciation quality of spontaneous non-native adult speech. Since the speaking content is not known prior to the assessment, a two-stage method is developed to first recognize the speaking content based on non-native speech acoustic properties and then forced-align the recognition results with a reference acoustic model reflecting native and near-native speech properties. Features related to Hidden Markov Model likelihoods and vowel durations are extracted. Words with low recognition confidence can be excluded in the extraction of likelihood-related features to minimize erroneous alignments due to speech recognition errors. Our experiments on the TOEFL®Practice Online test, an English language assessment, suggest that the recognition/forced-alignment method can provide useful pronunciation features. Our new pronunciation features are more meaningful than an utterance-based normalized acoustic model score used in previous research from a construct point of view.

1 Introduction

Automated systems for evaluating highly predictable speech (e.g. read speech or speech that is quite constrained in the use of vocabulary and syntactic structures) have emerged in the past decade (Bernstein, 1999; Witt, 1999; Franco et al., 2000; Hacker et al., 2005) due to the growing maturity of speech recognition and processing technologies. However, endeavors into automated scoring for spontaneous speech have been sparse given the challenge of both recognizing and assessing spontaneous speech. This paper addresses the development and evaluation of pronunciation features for an automated system for scoring spontaneous speech. This system was deployed for the TOEFL®Practice Online (TPO) assessment used by prospective test takers to prepare for the official TOEFL®test.

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 consists of three key categories: delivery, language use, and topic development. Delivery refers to the pace and the clarity of the speech, including performance on intonation, rhythm, rate of speech, and degree of hesitancy. 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. Most of the research on spontaneous speech assessment focuses on the delivery aspect given the low recognition accuracy on non-native spontaneous speech.

The delivery aspect can be measured on four dimensions: fluency, intonation, rhythm, and pronunciation. For the TPO assessment, we have defined pronunciation as the quality of vowels, consonants and word-level stress (segmentals). Intonation and sentence-level stress patterns (supra-segmentals) are not defined as part of pronunciation. Pronunciation is one of the key factors that impact the intelligibility and perceived comprehensibility of speech. Because pronunciation plays an important role in speech perception, features measuring pronunciation using speech technologies have been explored in many previous studies. However, the bulk of the research on automatic pronunciation evaluation concerns read speech or highly predictable speech (Witt, 1999; Franco et al., 2000; Hacker et al., 2005), where there is a high possibility of success in speech recognition. Automatic pronunciation evaluation is challenging for spontaneous speech and has been under-explored.

In this paper, we will describe a method for extracting pronunciation features based on spontaneous speech that is well motivated by theories and supported by empirical evaluations of feature performance. In conceptualizing and computing these features, we draw on the literature on automatic pronunciation evaluation for constrained speech. As described in the related work in Section 2, the widely used features for measuring pronunciation are (1) likelihood (posterior probability) of a phoneme being spoken given the observed audio sample that is computed in a Viterbi decoding process, and (2) phoneme length measurements that are compared to standard references based on native speech.

However, we have also come up with unique solutions to address the issue of relatively low accuracy in recognizing spontaneous speech. Our methods of feature extraction are designed with considerations of how to best capture the quality of pronunciation given technological constraints.

The remainder of the paper is organized as follows: Section 2 reviews the related research; Section 3 describes our method to extract a set of features for measuring pronunciation; Section 4 describes the design of the experiments, including the questions investigated, the data, the speech processing technologies, and the measurement metrics; Section 5 reports on the experimental results; Section 6 discusses the experimental results; and Section 7 summaries the findings and future research planned.

2 Related work

There is previous research on utilizing speech recognition technology to automatically assess non-native speakers' communicative competence (e.g., fluency, intonation, and pronunciation). Witt (Witt, 1999) developed the Goodness of Pronunciation (GOP) measurement for measuring pronunciation based on Hidden Markov Model (HMM) log likelihood. Using a similar method, Neumeyer et al. (Neumeyer et al., 2000) designed a series of likelihood related pronunciation features, e.g., the local average likelihood and global average likelihood. Hacker et al. (Hacker et al., 2005) utilized a relatively large feature vector for scoring pronunciation.

Pronunciation has been the focus of assessment in several automatic speech scoring systems. Franco et al. (Franco et al., 2000) presented a system for automatic evaluation of pronunciation quality on the phoneme level and the sentence level of speech by native and non-native speakers of English and other languages (e.g., French). A forced alignment between the speech read by subjects and the ideal path through the HMM was computed. Then, the log posterior probabilities for a certain position in the signal were computed to achieve a local pronunciation score. Cucchiarini et al. (Cucchiarini et al., 1997a; Cucchiarini et al., 1997b) designed a system for scoring Dutch pronunciation along a similar line. Their pronunciation feature set was more extensive, including various log likelihood HMM scores and phoneme duration scores. In these two systems, the speaking skill scores computed on features by machine are found to have good agreement with scores provided by humans.

A limited number of studies have been conducted on assessing speaking proficiency based on spontaneous speech. Moustroufas and Digalakis (Moustroufas and Digalakis, 2007) designed a system to automatically evaluate the pronunciation of foreign speakers using unknown text. The difference in the recognition results between a recognizer trained on speakers' native languages (L1) and another recognizer trained on their learned languages (L2) was used for pronunciation scoring. Zechner and Bejar (Zechner and Bejar, 2006) presented a system to score non-native spontaneous speech using features derived from the recognition results. Following their work, an operational assessment system, SpeechRater$^{\mathrm{TM}}$, was implemented with further improvements (Zechner et al., 2007).

There are some issues with the method to extract pronunciation features in the previous research on automated assessment of spontaneous speech (Zechner and Bejar, 2006; Zechner et al., 2007). For example, the acoustic model (AM) that was used to estimate a likelihood of a phoneme being spoken was well-fitted to non-native speech acoustic properties. Further, other important aspects of pronunciation, e.g., vowel duration, have not been utilized as a feature in the current SpeechRater$^{\mathrm{TM}}$ system. Likelihoods estimated on non-words (such as silences and fillers) that were not central to the measurement of pronunciation were used in the feature extraction. In addition, mis-recognized words lead to wrong likelihood estimation. Our paper attempts to address all of these limitations described above.

3 Extraction of Pronunciation Features

Figure 1 depicts our new method for extracting an expanded set of pronunciation features in a more meaning way.

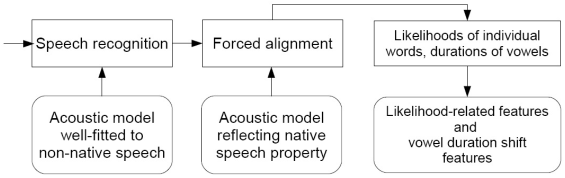
Speech recognition
Forced alignment
Likelihoods of individual words, durations of vowels
Acoustic model well-fitted to non-native speech
Acoustic model reflecting native speech property
Likelihood-related features and vowel duration shift features

Figure 1: Two-stage pronunciation feature extraction

We used two different AMs for pronunciation feature extraction. First, we used an AM optimized for speech recognition (typically an AM adapted on non-native speech to better fit non-native speakers' acoustics patterns) to generate word hypotheses; then we used the other AM optimized for pronunciation scoring (typically trained on native or near-native speech to be a good reference model reflecting expected speech characteristics) to force align the speech signals to the word hypotheses and to compute the likelihoods of individual words being spoken and durations of phonemes; finally new pronunciation features were extracted based on these measurements.

Some notations used for computing the pronunciation features are listed in Table 1. Based on these notations, the proposed new pronunciation features are described in Table 2. To address the limitations of previous research on automated assessment of pronunciation, which was described in Section 2, our proposed method has achieved improvements on (1) using the two-stage method to compute HMM likelihoods using a reference acoustic model trained on native and near-native speech, (2) expanding the coverage of pronunciation features by using vowel duration shifts that are compared to standard norms of native speech, (3) and using likelihoods on the audio portions that are recognized as words and applying various normalizations.

Table 1: Notations used for pronunciation feature extraction

Variable	Meaning
$L(x_i)$	the likelihood of word x_i being spoken given the observed audio signal
t_i	the duration of word i in a response
T_s	the duration of the entire response
T	$\sum_{i=1}^{n} t_i$, the summation of the duration of all words, where $T \leq T_s$
n	the number of words in a response
m	the number of letters in a response
R	$\frac{m}{T_s}$, the frequency of letters (as the rate of speech)
v_i	vowel i
N_v	the total number of vowels
P_{v_i}	the duration of vowel v_i
$\bar{P}$	the average vowel duration (across all vowels in the response being scored)
D_{v_i}	the standard average duration of vowel v_i (estimated on a native speech corpus)
$\bar{D}$	the averaged vowel duration (on all vowels in a native speech corpus)
S_{v_i}	$\lvert P_{v_i} - D_{v_i} \rvert$, duration shift of vowel v_i (measured as the absolute value of the difference between the duration of vowel v_i and its standard value)
Sn_{v_i}	$\lvert \frac{P_{v_i}}{\bar{P}} - \frac{D_{v_i}}{\bar{D}} \rvert$, normalized duration shift of vowel v_i (measured as the absolute value of the normalized difference between the duration of vowel v_i and its standard value)

4 Experiment design

We first raise three questions that we try to answer with our experiments. Then, we describe the data sets and the speech recognizers, especially the two

Table 2: A list of proposed pronunciation features

Feature	Formula	Meaning
L_1	$\sum_{i=1}^{n} L(x_i)$	summation of likelihoods of all the individual words
L_2	L_1/n	average likelihood across all words
L_3	L_1/m	average likelihood across all letters
L_4	L_1/T	average likelihood per second
L_5	$\frac{\sum_{i=1}^{n} \frac{L(x_i)}{t_i}}{n}$	average likelihood density across all words
L_6	L_4/R	L_4 normalized by the rate of speech
L_7	L_5/R	L_5 normalized by the rate of speech
$\bar{S}$	$\frac{\sum_{i=1}^{N_v} S_{v_i}}{N_v}$	average vowel duration shifts
$\bar{Sn}$	$\frac{\sum_{i=1}^{N_v} Sn_{v_i}}{N_v}$	average normalized vowel duration shifts

different acoustic models fitted to non-native and expected speech respectively. Finally, we describe the evaluation criterion used in the experiment.

4.1 Research questions

In order to justify that the two-stage method for extracting pronunciation features is a valid method that provides useful features for assessing pronunciation, the following questions need to be answered:

Q1: Can the words hypothesized be used to approximate the human transcripts in the forced alignment step?

Q2: Are the new pronunciation features effective for assessment?

Q3: Can the likelihood-related features be improved when using only words correctly recognized?

4.2 Data

Table 3 lists the data sets used in the experiment. Non-native speech collected in the TPO was used in training a non-native AM. For feature evaluations, we selected $1,257$ responses from the TPO data collected in 2006. Within this set, 645 responses were transcribed. Holistic scores were assigned by human raters based on a score scale of 1 (the lowest proficiency) to 4 (the highest proficiency).

In the TOEFL®Native Speaker Study, native speakers of primarily North American English (NaE) took the TOEFL®test and their speech files were collected. This TOEFL®native speech data and some high-scored TPO responses were used in the adaptation of an AM representing expected speech properties. In addition, $1,602$ responses of native speech, which had the highest speech proficiency scores in NaE, were used to estimate standard average vowel durations.

Type	Function	Source	Size
non-native speech	AM training	TPO	~ 30 hrs
	feature evaluation	TPO collected in 2006	$1,257$ responses (645 with transcripts)
native or near-native speech	AM adaptation	TPO and TOEFL Native	$\sim 2,000$ responses
	estimation of standard vowel durations	TOEFL Native	$1,602$ responses

Table 3: Data sets used in the experiment

4.3 Speech technologies

For speech recognition and forced alignment, we used a gender-independent fully continuous HMM speech recognizer. 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 TPO. For language model training, a large corpus of non-native speech (about 100 hours) was used and mixed with a large general-domain language model (trained from the Broadcast News (BN) corpus (Graff et al., 1997) of the Linguistic Data Consortium (LDC)). In the pronunciation feature extraction process depicted in Figure 1, this AM was used to recognize non-native speech to generate the word hypotheses.

The AM used in the forced alignment was trained on native speech and high-scored non-native speech. 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 batch-mode MAP adaptation. The adaptation corpus contained about $2,000$ responses with high scores in previous TPO tests and the TOEFL$^{\circledR}$Native Speaker Study. In addition, this AM was used to estimate standard norms of vowels as described in Table 1.

4.4 Measurement metric

To measure the quality of the developed features, a widely used metric is the Pearson correlation (r) computed between the features and human scores. In previous studies, human holistic scores of perceived proficiency have been widely used in estimating the correlations. In our experiment, we will use the absolute value of Pearson correlation with human holistic scores ($|r|$) to evaluate the features. Given the close relationship between pronunciation quality and overall speech proficiency, $|r|$ is expected to approximate the strength of its relationship with the human pronunciation scores.

5 Experimental Results

5.1 Results for Q1

When assessing read speech, the transcription of the spoken content is known prior to the assessment and used to forced-align the speech for feature extraction. However, when assessing spontaneous speech, we do not know the spoken content and cannot provide a correct word transcription for the forced alignment with imperfect speech recognition. A practical solution is to use the recognition hypothesis to approximate the human transcript in the forced alignment. Since the recognition word accuracy on non-native spontaneous speech is not very high (for example, a word accuracy of about 50% on the TPO data was reported in (Zechner et al., 2007)), it is critical to verify that the approximation can provide good enough pronunciation features compared to the ones computed in an ideal scenario (by using the human transcript in the forced alignment step).

We ran forced alignment on 645 TPO responses with human transcriptions, using both the manual transcription and the word hypotheses from the recognizer described in Section 4.3. Then, based on these two forced alignment outputs, we extracted the pronunciation features as described in Section 3.

Table 4 reports the $|r|$s between the proposed pronunciation features and human holistic scores when using the forced alignment results from either transcriptions or recognition hypotheses. The relative $|r|$ reduction (defined as $(|r|_{transcriptions} - |r|_{hypotheses})/|r|_{transcriptions} * 100$) is reported to measure the magnitude reduction.

Based on the results shown in Table 4, we find that the pronunciation features computed based on the forced alignment results using transcriptions have higher $|r|$s with the human holistic scores than the corresponding features computed based on the FA results using the recognition hypotheses. This is not surprising given that only $50\% \sim 60\%$ word accuracy can be achieved when recognizing non-native spontaneous speech. However, the pronunciation features computed using the recognition hypotheses that is feasible in practice show some promising correlations to human holistic scores. For example, L_3, L_6, and L_7 have $|r|$s larger than 0.45 and $\bar{Sn}$ has an $|r|$ larger than 0.35. Compared to the corresponding features computed using the FA results based on transcriptions, these promising pronunciation features that can be obtained practically, show some reduction in quality (from 13.4% to 21.1%) but are still usable. Therefore, our proposed two-stage method for pronunciation feature extraction is proven to be a practical way for the computation of features that have acceptable performance.

5.2 Result for Q2

Although our proposed modifications described in Section 3 have improved the meaningfulness of the features, an empirical study is needed to examine the actual utility of these features for the assessment of pronunciation.

In the experiment described in Section 5.1, four pronunciation features (including L_3, L_6, L_7, and

Feature	$\|r\|$ using transcription	$\|r\|$ using recognition hypothesis	relative $\|r\|$ reduction (%)
L_1	0.216	0.107	50.5
L_2	0.443	0.416	6.1
L_3	0.506	0.473	6.5
L_4	0.363	0.294	19
L_5	0.333	0.287	13.8
L_6	0.549	0.475	13.5
L_7	0.546	0.473	13.4
$\bar{S}$	0.396	0.296	25.3
$\bar{S}n$	0.451	0.356	21.1

Table 4: $|r|$ between the pronunciation features and human holistic scores under two forced alignment input conditions (using transcriptions vs. using recognition hypotheses) and relative $|r|$ reduction

$\bar{S}n$) show promising correlations to human holistic scores. To check the quality of the newly developed pronunciation features, we compared these four features with the *amscore* feature used in (Zechner et al., 2007) on the TPO data set collected in 2006 (with $1,257$ responses). We first ran speech recognition using the recognizer designed for non-native speech. The recognition results were used to compute the *amscore*, which is calculated by dividing the likelihood over an entire response by the number of letters. Then, we used the recognition hypotheses to do the forced alignment using the other AM trained on the native and near-native speech to extract those four pronunciation features. Finally, we calculated the correlation coefficients between features and the human holistic scores. The results are reported in Table 5.

feature	$\|r\|$ to human holistic scores
amscore	0.434
L_3	0.369
L_6	0.444
L_7	0.443
$\bar{S}n$	0.363

Table 5: A comparison of new pronunciation features to *amscore*, the one used in SpeechRater$^{\mathrm{TM}}$

Compared to the feature *amscore*, L_6 and L_7 have slightly higher $|r|$s with the human holistic scores. This suggests that our construct-driven approach yields pronunciation features that are empirically comparable or even better than the *amscore*. In addition, $\bar{S}n$, a new feature representing the vowel production aspect of pronunciation, shows a relatively high correlation with human holistic scores. This suggests that our new pronunciation feature set has an expanded coverage of pronunciation.

It is interesting to note that L_3 has a lower $|r|$ with human holistic scores than the *amscore* does. Although the computation of L_3 is quite similar to that of *amscore*, the major difference is that likelihoods of non-word portions (such as silences and fillers) are used to compute *amscore* but not L_3. This suggests that likelihood-related pronunciation features that involve information related to non-words may perform better in predicting human holistic scores. For example, for *amscore*, the likelihoods measured on those non-word units were involved in the feature calculation; for L_6 and L_7, the temporal information of those non-word units (e.g., duration of units) was involved in the feature calculation [1].

5.3 Result for Q3

In the feature extraction, we used the words hypothesized by the speech recognizer as the input for the forced alignment. Since a considerable number of words are recognized incorrectly (especially for non-native spontaneous speech), a natural way to further improve the likelihood related features is to only consider words which are correctly recognized. A useful metric associated with the recognition performance is the confidence score (CS) output by the recognizer, which reflects the recognizer's estimation about the probability that a hypothesized word is correctly recognized. The recognized words with high confidence scores tend to be correctly recognized. Therefore, focusing on words recognized with high confidence scores may reduce the negative impact caused by recognition errors on the quality of the likelihood related features.

On the TPO data with human transcripts, we used the NIST's *sclite* scoring tool (Fiscus, 2009) to measure the percentage of correct words (correct%), which is defined as the ratio of the number of words

[1] L_6 and L_7 use R, which is computed as $\frac{m}{T_s}$, where T_s contains durations of non-words.

correctly recognized given the number of words in the reference transcript. On all words (corresponding to confidence scores ranging from 0.0 to 1.0), the correct% is 53.3%. Figure 2 depicts the correct% corresponding to ten confidence score bins ranging from 0.0 to 1.0. Clearly, with the increase of the confidence score, more words tend to be accurately recognized. Therefore, it is reasonable to only use likelihoods estimated on the hypothesized words with high confidence scores for extracting likelihood related features.

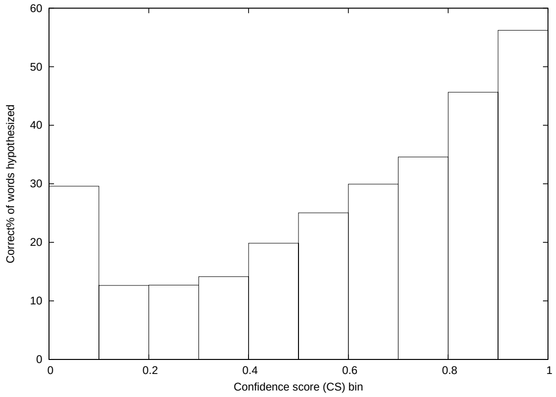

Figure 2: Correct% of words recognized across 10 confidence score bins

On the TPO data set collected in 2006, we computed three likelihood related features (including L_3, L_6, and L_7) only on words whose SC is equal to or higher than a threshold (i.e., 0.5, 0.6, 0.7, 0.8, and 0.9) and measured the $|r|$ of a feature with the human holistic scores. Table 6 lists the confidence score cutting thresholds, the percentage of words whose confidence scores are not lower than the cutting threshold selected, and $|r|$ between each likelihood feature to human holistic scores. In the Table 6, we observe that only using words recognized with high confidence improves the correlations between the features and the human holistic scores. One issue about only using words recognized with high confidence scores is that the number of words used in the feature extraction has been reduced and may reduce the robustness of the feature calculation.

T_c	percentage of words whose CS $\geq T_c$ (%)	L_3 $\lvert r\rvert$	L_6 $\lvert r\rvert$	L_7 $\lvert r\rvert$
0.0	100	0.369	0.444	0.443
0.5	84.21	0.38	0.462	0.461
0.6	77.07	0.377	0.465	0.464
0.7	69.31	0.363	0.461	0.461
0.8	60.86	0.371	0.466	0.466
0.9	50.76	0.426	0.477	0.475

Table 6: $|r|$ between L_3, L_6, and L_7 and human holistic scores using only words recognized whose CSs are not lower than a threshold (T_c)

6 Discussion

To assess the pronunciation of spontaneous speech, we proposed a method for extracting a set of pronunciation features. The method consists of two stages: (1) recognizing speech using an AM well fitted to non-native speech properties and (2) forced-aligning the hypothesized words using the other AM, which was trained on native and near-native speech, and extracting features related to spectral properties (HMM likelihood) and vowel production. This method of using one AM optimized for speech recognition and another AM optimized for pronunciation evaluation is well motivated theoretically. The derived pronunciation features have also been found to have reasonably high correlations with human holistic scores. The results support the linkage of the features to the construct of pronunciation and their utility of being used in a scoring model to predict human holistic judgments. Several contributions of this paper are described as below.

First, the two-stage method allows us to utilize an AM trained on native and near-native speech as a reference model when computing pronunciation features. The decision to include high-scored non-native speech was driven by the scoring rubrics derived from the construct, where the pronunciation quality of the highest level performance does not necessarily require native-like accent, but highly intelligible speech. The way the reference model was trained is consistent with the scoring rubrics, and makes it an appropriate standard based on which the pronunciation quality of non-native speech can be evaluated. By using the recognition hypotheses from the recognition step as input in the forced alignment step, our experiments show a relatively small reduction in correlations with human holistic scores in comparison to the features based on the human transcriptions. This suggests that our method has potential to be implemented in a real-time operational setting.

Second, a few decisions we have made in computing the pronunciation features are driven by considerations of how these features are meaningfully linked to the construct of pronunciation assessment. For example, we have excluded the HMM likelihoods on non-words (such as pauses and fillers) in the computations of likelihood-related features. In addition, only using words recognized with high confidence scores yields more informative likelihood-related features for assessing the quality of speech. The inclusion of vowel duration measures in the feature set expanded the coverage of the quality of pronunciation.

7 Summary and future work

This paper presents a method for computing features for assessing the pronunciation quality of non-native spontaneous speech, guided by construct considerations. We were able to show that using a two-stage method of first recognizing speech with a non-native AM and then forced aligning of the hypothesis using a native or near-native speech AM we can generate pronunciation features with promising correlations with holistic scores assigned by human raters.

We plan to continue our research in the following directions: (1) we will improve the native speech norms for vowel durations, such as using the distribution of vowel durations rather than just the mean of durations in our feature computations; (2) we will investigate other aspects of pronunciation, e.g., consonant quality and word stress; (3) we will add other standard varieties of English (such as British, Canadian, Australian, etc) to the training corpus for the reference pronunciation model as the current model is trained on primarily North American English (NaE).

References

J. Bernstein. 1999. PhonePass testing: Structure and construct. Technical report, Ordinate Corporation.

C. Cucchiarini, H. Strik, and L. Boves. 1997a. Automatic evaluation of Dutch Pronunciation by using Speech Recognition Technology. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Santa Barbara, CA.

C. Cucchiarini, H. Strik, and L. Boves. 1997b. Using Speech Recognition Technology to Assess Foreign Speakers' Pronunciation of Dutch. In *3rd international symosium on the acquision of second language speech*, Klagenfurt, Austria.

J. Fiscus. 2009. Speech Recognition Scoring Toolkit (SCTK) Version 2.3.10.

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

D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett. 1997. 1996 English Broadcast News Speech (HUB4).

C. Hacker, T. Cincarek, R. Grubn, S. Steidl, E. Noth, and H. Niemann. 2005. Pronunciation Feature Extraction. In *Proceedings of DAGM 2005*.

N. Moustroufas and V. Digalakis. 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech and Language*, 21(6):219–230.

L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. 2000. Automatic Scoring of Pronunciation Quality. *Speech Communication*, 6.

S. M. Witt. 1999. *Use of Speech Recognition in Computer-assisted Language Learning*. Ph.D. thesis, University of Cambridge.

K. Zechner and I. Bejar. 2006. Towards Automatic Scoring of Non-Native Spontaneous Speech. In *NAACL-HLT*, NewYork NY.

K. Zechner, D. Higgins, and Xiaoming Xi. 2007. SpeechRater: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech. In *Proc. SLaTE*.