

Non-scorable Response Detection for Automated Speaking Proficiency Assessment

Su-Youn Yoon, Keelan Evanini, Klaus Zechner

Educational Testing Service

660 Rosedale Road, Princeton, NJ, USA

{syoon, kevanini, kzechner}@ets.org

Abstract

We present a method that filters out non-scorable (NS) responses, such as responses with a technical difficulty, in an automated speaking proficiency assessment system. The assessment system described in this study first filters out the non-scorable responses and then predicts a proficiency score using a scoring model for the remaining responses.

The data were collected from non-native speakers in two different countries, using two different item types in the proficiency assessment: items that elicit spontaneous speech and items that elicit recited speech. Since the proportion of NS responses and the features available to the model differ according to the item type, an item type specific model was trained for each item type. The accuracy of the models ranged between 75% and 79% in spontaneous speech items and between 95% and 97% in recited speech items.

Two different groups of features, signal processing based features and automatic speech recognition (ASR) based features, were implemented. The ASR based models achieved higher accuracy than the non-ASR based models.

1 Introduction

We developed a method that filters out non-scorable (NS) responses as a supplementary module to an automated speech proficiency assessment system. In this study, the method was developed for a telephony-based assessment of English proficiency for non-native speakers. The examinees' responses

were collected from several different environmental conditions, and many of the utterances contain background noise from diverse sources. In addition to the presence of noise, many responses have other sub-optimal characteristics. For example, some responses contain uncooperative behavior from the speakers, such as non-English speech, whispered speech, and non-responses. These types of responses make it difficult to provide a valid assessment of a speaker's English proficiency. Therefore, in order to address the diverse types of causes for these problematic responses, we used a two step approach: first, these problematic responses were filtered out by a "filtering model," and only the remaining responses were scored using the automated scoring model.

The overall architecture of our method, including the automated speech proficiency scoring system, is as follows: for a given spoken response, the system performs speech recognition, yielding a word hypothesis and time stamps. In addition to word recognition, the system computes pitch and power to generate prosodic features; the system calculates descriptive statistics such as the mean and standard deviation of pitch and power at both the word level and response level. Given the word hypotheses and pitch/power features, it derives features for automated proficiency scoring. Next, the non-ASR based features are calculated separately using signal processing techniques. Finally, given both sets of features, the filtering model identifies NS responses.

This paper will proceed as follows: we will review previous studies (Section 2), present the data

(Section 3), and then describe the structure of the filtering model (Section 4). Next, the results will be presented (Section 5), followed by a discussion (Section 6), and we will conclude with a summary of the importance of the findings (Section 7).

2 Previous Work

Higgins et al. (2011) developed a “filtering model” that is conceptually similar to the one in this paper. The model was trained and tested on a corpus containing responses from non-native speakers to an English proficiency assessment. This system used a regression model based on four features which were originally designed for automated speech proficiency scoring: the number of distinct words in the speech recognition output, the average speech recognizer confidence score, the average power of the speech signal, and the mean absolute deviation of the speech signal power. This model was able to identify responses which were also identified as NS responses by human raters with an approximately 98% accuracy when a false positive rate (the proportion of responses without technical difficulties that were incorrectly flagged as problematic) was lower than 1%.

Although there are few other studies which are directly related to the task of filtering out non-scorable responses in the domain of automated speech proficiency assessment, several signal processing studies are related to this work. Traditionally, the Signal to Noise Ratio (SNR) has been used to detect speech with a large amount of background noise. This method measures the ratio between the total energy of the speech signal and the total energy of the noise; if the SNR is low, then the speech contains loud background noise. A low SNR results in lower intelligibility and increases the difficulty for both human and automated scoring. Furthermore, spectral characteristics can be also applied to detect speech with loud background noise, since noise has different spectral characteristics than speech (noise tends to have no or few peaks in the spectral domain). If a response contains loud background noise, then the spectral characteristics of the speech may be obscured by noise and it may have similar characteristics with the noise. These differences in spectral characteristics have been used in audio information

retrieval Lu and Hankinson (1998).

Secondly, responses without valid speech can be identified using Voice Activity Detection (VAD). VAD is a technique which distinguishes human speech from non-speech. When speech is clean, VAD can be calculated by simply computing the zero-crossing rate which signals the existence of cyclic waves such as vowels. However, if the response also contains loud background noises, more sophisticated methods are required. In order to remove the influence of noise, Chang and Kim (2003), Chang et al. (2006), Shin et al. (2005) and Sohn et al. (1999) estimated the characteristics of the noise spectrum and the distribution of noise, and compensated for them when speech is identified. The performance of these systems is heavily-influenced by the accuracy of estimating characteristics of the background noise.

In this study, we used a set of ASR based features and non-ASR based features. ASR based features were similar to the ones used by Zechner et al. (2009). In addition to the features based on ASR hypotheses, the ASR based feature set contained basic pitch and power related features since the ASR system in this study also produced pitch and power measurements in order to generate prosodic features. The non-ASR based features were comprised of four groups of features based on signal processing techniques such as SNR, VAD, and pitch and power. Features related to pitch and power were included in both the ASR based features and the non-ASR based features. Since the non-ASR based features were originally implemented as an independent module from the ASR-based system (it was implemented for the case where the appropriate recognizer is unavailable), there is some degree of overlap between the two feature sets.

3 Data

The data for this experiment were drawn from a prototype of a telephony-based English language assessment. Non-native speakers of English each responded to 40 test items designed to evaluate their level of English proficiency. The test was composed of items that elicited both spontaneous speech (hereafter SS) and recited speech (hereafter RS). In this study, 8 items (four SS and four RS) were used for

each speaker.

Participants used either a cell phone or a land line to complete the assessment, and the participants were compensated for their time. The motivation level of the participants was thus lower than in the case of an actual high stakes assessment, where a participant’s performance could have a substantial impact on their future. In addition, the data collection procedure was less controlled than in an operational testing environment; for example, some recordings exhibited higher levels of ambient noise than others. These two facts led to the quality of some of the responses being lower than would be expected in an operational assessment.

The data for this study were collected from participants in two countries: India and China. For India, 4900 responses from 638 speakers were collected. For China, 5565 responses from 702 speakers were collected (some of the participants did not provide responses to all 8 test items). Each response is approximately 45 sec in duration.

After the data was collected, all of the responses were given scores on a three-point scale by trained raters. The raters also labeled responses as “non-scorable” (NS), when appropriate. NS responses are ones that could not be given a score according to the rubrics of the three-point scale. These were due to either a technical difficulty obscuring the content of the response or an inappropriate response from the participant.

The proportion of NS responses differs markedly between the two countries. 852 of the responses in the India data set (17% of the total) were labeled as NS, compared to 1548 responses (28%) in the China data set.

Table 1 provides the different types of NS responses that were annotated by the raters, along with the relative frequency of each NS category compared to the others.

Excluding the category “Other”, background noise, non-responses, and unrelated topic were the most frequent types of NS response for both data sets. However, the relative proportions of each type differed somewhat between the two countries. For example, the most frequent NS type in India was background noise; 33% of NS responses were of this type, 1.7 times higher than in China.

The proportion of unrelated topic responses was

| NS Type | India (%) | China (%) |
|--------------------|-----------|-----------|
| Background noise | 33.2 | 19.6 |
| Other | 25.0 | 15.4 |
| Unrelated response | 18.9 | 40.1 |
| Non-response | 10.6 | 8.8 |
| Non-English speech | 4.9 | 6.4 |
| Too soft | 2.8 | 1.0 |
| Background speech | 2.0 | 1.9 |
| Missing samples | 1.5 | 4.0 |
| Too loud | 0.8 | 0.1 |
| Cheating | 0.3 | 2.7 |

Table 1: Different types of NS responses and their relative frequency, in % of all NS for each country (ranked by frequency of occurrence in India)

| Data Partition | India | | China | |
|----------------|-----------------|--------|-----------------|--------|
| | # of re-sponses | NS (%) | # of re-sponses | NS (%) |
| SS-train | 1114 | 31.6 | 1382 | 32.2 |
| SS-eval | 1271 | 27.5 | 1391 | 33.8 |
| RS-train | 1253 | 8.0 | 1392 | 22.4 |
| RS-eval | 1275 | 4.8 | 1400 | 22.9 |

Table 2: Item-type specific training and evaluation data

also high in both countries, but it was much higher in the China data set: it was 19% in the responses from India and 40% for China (more than twice as high as in India). All responses which were not directly related to the prompt fell into this category. For SS items, the majority included responses about a different topic. For RS items, responses in which the speakers read different prompts were classified into this category.

The responses were divided into training and testing for NS response detection. Due to the significant difference in the proportion of NS responses and relative frequencies of NS types in the two data sets, filtering models were trained separately for each country. In addition, since the proportions of NS responses and the available features varied according to the item type, training and testing data were further classified by item types. The proportions of NS responses and the sizes of the partitions, along with the percent of NS responses in each item type, are shown in Table 2.

The partitions for testing the filtering model were selected to maximize the number of speakers with complete sets of responses; however, this constraint was not able to be met for the training partitions in the India data set (due to insufficient data). This explains the lower proportion of NS responses in the India test partitions, since speakers with complete sets of responses were less likely to provide bad responses. As Table 2 shows, NS responses were more frequent among SS items than RS items: the proportion of NS responses in SS items was four times higher than in RS items in India and 1.5 times in China.

4 Method

4.1 Overview

In this study, two different sets of features were used in the model training process; ASR-based features and non-ASR based features. For each item-type, an item-type-specific filtering model was developed using these two sets of features.

4.2 Feature generation

4.2.1 ASR based features

For this feature set, we used the features from an automated speech proficiency scoring system. This scoring system used an ASR engine containing word-internal triphone acoustic models and item-type-specific language models. Separate acoustic models were trained for the data sets from the two countries. The acoustic training data for the two models consisted of 45.5 hours of speech from India and 123.1 hours of speech from China. In addition, separate language models were trained for the SS and RS items for each country; for the RS items, the language models also incorporated the texts of the prompts.

A total of 61 features were available. Among these features, many features were conceptually similar but based on different normalization methods. These features showed a strong intercorrelation. For this study, 30 features were selected and classified into four groups according to their characteristics: basic features, fluency features, ASR-confidence features, and Word Error Rate (WER) features.

The basic features are related to power and pitch, and they capture the overall distribution of pitch and power values in a speaker’s response using mean and variance calculations. These features are relevant since NS responses may have an abnormal distribution in energy. For instance, non-responses contain very low energy. In order to detect these abnormalities in speech signal, pitch and power related features were calculated.

The fluency features measure the length of a response in terms of duration and number of words. In addition, this group contains features related to speaking rate and silences, such as mean duration and number of silence. In particular, these features are effective in identifying Non-responses which contain zero or only a few words.

The ASR-confidence group contains features predicting the performance of the speech recognizer. Low speech recognition accuracy may be indicated by low confidence scores.

Finally, the WER group provides features estimating the similarity between the prompts and the recognition output. In addition to the conventional word error rate (WER), term error rate (TER) was also implemented for the filtering model. TER is a metric commonly used in spoken information retrieval, and it only accounts for errors in content words. This measure may be more effective in identifying NS responses than conventional WER; for instance, the overlap in function words between off-topic responses and prompts can be correctly ignored. TER was calculated according to the following formula:

$$\begin{aligned} dif(W_c) &= \begin{cases} 0 & \text{if } C_{ref}(W_c) < C_{hyp}(W_c) \\ C_{ref}(W_c) - C_{hyp}(W_c) & \text{otherwise} \end{cases} \\ TER &= \frac{\sum_{c \in WC} dif(W_c)}{\sum_{c \in WC} C_{ref}(W_c)} \end{aligned} \quad (1)$$

where $C_{ref}(W_c)$ is the number of occurrences of the word W_c in reference, $C_{hyp}(W_c)$ is the number of occurrences of the word W_c in hypothesis, and WC is the set of content words in reference.

Formula 1 differs from the conventional method

| Group | List of features |
|-----------------|--|
| Basic | mean/standard deviation/minimum/maximum of power, difference between maximum and minimum in power, mean/standard deviation/minimum/maximum of pitch, difference between maximum and minimum in pitch |
| Fluency | duration of whole speech part, number of words, speaking rate (word per sec), mean/standard deviation of silence duration, number of silences, silences per sec and silences per word |
| ASR score | mean of confidence score, normalized Acoustic Model score by word length, normalized Language Model score by number of words |
| Word Error Rate | the word accuracy between prompt and ASR word hypothesis, correct words per minute, term error rate |

Table 3: List of ASR based features

of calculating TER in two ways. Firstly, content words which occurred only in the word hypothesis are ignored in the formula. Secondly, if a word occurred in the word hypothesis more frequently than in the reference, the difference is ignored. These modifications were made to address characteristics of the responses in the data. On the one hand, speakers occasionally inserted a few words such as “too difficult” at the end of a response. In addition, a few speakers repeated words contained in the prompt multiple times. The two modifications to TER address both of these issues.

All features from the four groups are summarized in Table 3.

4.2.2 Non-ASR based features

A total of 12 features from four different groups were implemented using non-ASR based methods such as VAD and SNR. These features are listed in Table 4.

| Feature Category | Feature |
|------------------|--|
| VAD | proportion of voiced frames in response, number and total duration of voiced regions |
| Syllable | number of syllables |
| Amplitude | maximum, minimum, mean, standard deviation |
| SNR | SNR, speech peak |

Table 4: List of non-ASR based features

VAD related features were implemented using the

ESPS speech analysis program. For every 10 millisecond interval, the voice frame detector determined whether the interval was voiced or not. Three features were implemented using this voiced interval information: the number of voiced intervals, ratio of voiced intervals in the entire response, and the total duration of voiced intervals.

In addition, the number of syllables was estimated based on the flow of energy. The energy of the syllable tends to reach its peak in the nucleus and the dip in the boundaries. By counting the number of such fluctuations in energy measurements, the number of syllables can be estimated. The Praat script from De Jong and Wempe (2009) was used for this purpose.

In order to detect the abnormalities in energy, amplitude based features were calculated. These features were similar to the basic features in ASR based features.

Finally, if a response contains loud background noise, the ratio of speech to noise is low. SNR, the mean noise level, and the peak speech level were computed using the NIST audio quality assurance package (NIST, 2009).

The VAD and syllable feature groups were designed to estimate the number of syllables, the proportion of speech to non-speech, and the total duration of speech intervals. These features were similar to the number of words and duration of speech features in the ASR-based feature set. Despite the conceptual similarity, these features were implemented since the two types of features were calculated using different characteristics of the spoken response.

The VAD and syllable features are based on the flow of energy and the zero crossing rate and the ASR-based features are based on the speech recognition. In particular, the speech recognizer tends to generate word hypotheses even for responses that contain no speech input, but VAD does not have such a tendency. Due to this difference, VAD based features may be more robust in the responses with no valid speech.

4.3 Model building

For each response, both ASR features and non-ASR features were calculated. In contrast to non-ASR features, which were available for all responses, ASR features (except the Basic group) were unavailable for some responses, namely, responses for which the ASR system did not generate any word hypotheses because no tokens received scores above the rejection threshold. This causes a missing value problem; about 7% of the responses did not have a complete set of attributes.

Missing values are a common problem in machine learning. One of the popular approaches is to replace a missing value with a unique value such as the attribute's mean. Ding and Simonoff (2008) proposed a method that replaces a missing value with an arbitrary unique value. This method is preferable when missing of a value depends on the target value and this relationship holds in both training and test data.

In this study, the missing values were replaced with unique values due to the relationship between the missing values and the target label; if the speech recognizer did not produce any word hypotheses, the response was highly likely to be a NS response. 63% of the responses where the speech recognizer failed to generate word hypotheses were NS responses. Since all ASR-based features were continuous values, we used two real values: 0.0 for fluency features and ASR features and 100.0 for word error rate features. The fluency features and ASR features tend to be 0.0 while the word error rate features tend to be 100.0 when the responses are NS responses.

A total of 42 features were used in the model building. The only exception was WER; since WER features were only available for the model based on recited speech, they were calculated only for RS items. Decision tree models were trained using the J48 algorithm (WEKA implementation of C4.5) of

WEKA machine learning toolkit (Hall et al., 2009).

5 Results

For each item-type, three models were built to investigate the impact of each feature group: a model using non-ASR features, a model using ASR features, and a model using both features (the "Combined" model). Tables 5 and 6 present the accuracy of the SS models and Tables 7 and 8 present the accuracy of the RS models. In all tables, the baseline was calculated using majority voting, and represented a system in which no responses were classified as NS; since the majority class was scorable, the baseline using the majority voting did not predict any response as non-scorable response. Therefore, precision, recall, F-score are all 0 in this case.

| Model | Acc. | Pre. | Rec. | F-score |
|----------|------|-------|-------|---------|
| Baseline | 72.5 | 0 | 0 | 0 |
| Non-ASR | 77.0 | 0.645 | 0.364 | 0.465 |
| ASR | 79.0 | 0.683 | 0.444 | 0.538 |
| Combined | 78.6 | 0.657 | 0.461 | 0.542 |

Table 5: Performance of the SS model in India

| Model | Acc. | Pre. | Rec. | F-score |
|----------|------|-------|-------|---------|
| Baseline | 66.2 | 0 | 0 | 0 |
| Non-ASR | 68.9 | 0.601 | 0.240 | 0.343 |
| ASR | 72.9 | 0.718 | 0.326 | 0.448 |
| Combined | 72.9 | 0.720 | 0.323 | 0.446 |

Table 6: Performance of the SS model in China

| Model | Acc. | Pre. | Rec. | F-score |
|----------|------|-------|-------|---------|
| Baseline | 94.8 | 0 | 0 | 0 |
| Non-ASR | 95.7 | 0.684 | 0.210 | 0.321 |
| ASR | 97.2 | 0.882 | 0.484 | 0.625 |
| Combined | 96.8 | 0.769 | 0.484 | 0.594 |

Table 7: Performance of the RS model in India

In both item-types, the models using ASR-based features achieved the best performance. The SS model achieved 79% accuracy in India and 73% accuracy in China, representing improvements of approximately 7% over the baseline. In both data sets, the RS model achieved high accuracies: 97% accuracy in India and 96% accuracy in China. In India,

| Model | Acc. | Pre. | Rec. | F-score |
|----------|------|-------|-------|---------|
| Baseline | 77.1 | 0 | 0 | 0 |
| Non-ASR | 78.3 | 0.555 | 0.268 | 0.361 |
| ASR | 95.6 | 0.942 | 0.860 | 0.899 |
| Combined | 95.1 | 0.912 | 0.872 | 0.892 |

Table 8: Performance of the RS model in China

this represents a 2.4% improvement over the baseline. Although the absolute value of this error reduction is not very large, the relative error reduction is 46%. In China, the improvement was more salient; there was 18% improvement over baseline, corresponding to a relative error reduction of 78%.

Additional experiments were conducted to determine the robustness of the filtering models to evaluation data from a country not included in the training data. The evaluation sets from both item types (SS and RS) in both countries (India and China) were processed using three different models: 1) a model trained using the ASR-based features for the responses from the same country (the ‘‘Same’’ condition, whose results are identical to the ‘‘ASR’’ results in Tables 5 - 8), 2) a model trained using the ASR-based features for the responses from the other country (the ‘‘Different’’ condition), and 3) a model trained using the ASR-based features for the responses from both countries (the ‘‘Both’’ condition). Table 9 presents the accuracy results for these four sets of experiments.

| Model | India | | China | |
|-----------|-------|------|-------|------|
| | SS | RS | SS | RS |
| Same | 79.0 | 97.2 | 72.9 | 95.6 |
| Different | 80.1 | 95.4 | 73.5 | 93.8 |
| Both | 80.0 | 96.5 | 74.0 | 95.9 |

Table 9: Accuracy results using training and evaluation data from different countries

These results show that the models are quite robust to evaluation data from a different country. In all cases, there is at most a small decline in performance when training data from the other country is used (in the case of the SS responses, there is even a slight increase in performance). Table 9 also shows that the RS models performed worse in the Different Country condition (compared to the Same Country

condition) than the SS models. This difference is likely due to the difference in the number of NS responses among the RS data in the two countries (as shown in Table 2). However, the decline is still relatively small, suggesting that it would be reasonable to extend the filtering models to responses from additional countries that were not seen in the training data.

6 Discussion

Approximately 40 features were available for the model building, but not all features had a significant impact on the detection of NS responses. For each item-type, the importance of features were further investigated using a logistic regression analysis. The training data of India and China were combined, and a stepwise logistic regression analysis was performed using the SPSS statistical analysis program.

For each item-type, the top 3 features are presented in Table 10; the features are presented in the same order selected in the models.

| Model | RS | SS |
|----------|--|--|
| ASR | TER, speaking rate, s.d. of pitch | mean of confidence scores, speaking rate, s.d. of power |
| Non-ASR | number of syllables, number and duration of voiced regions | number of syllables, s.d. and mean of amplitude |
| Combined | TER, speaking rate, s.s.dd. pitch | mean of confidence scores, speaking rate, number of voiced regions |

Table 10: Top 3 features in stepwise logistic regression model

For the RS items, TER was the best feature and it was the top feature for both the ASR feature based model and the combined model. The top 3 features in the combined model were the same as the ASR feature based model, and non-ASR features were not

selected. In non-ASR based features, the number of syllables was the best feature, followed by the VAD based features.

For the SS items, the top 2 features were the same in both the ASR feature based model and the combined model. The combined model selected one non-ASR based feature, namely, a VAD based feature. As with the RS items, the number of syllables was the best feature, followed by the energy related feature.

These results show the importance of WER features. Most of the current features are designed for signal level abnormalities such as responses with large background noise or non-responses. For instance, fluency features and VAD features are effective for non-response detection, since they can determine whether the responses contain valid speech or not. SNR and pitch/power related features are useful for identifying responses with large background noise. However, no features except the WER group can identify content-level abnormalities such as unrelated topic and non-English responses. The high proportion of these two types of responses (24% in India and 46% in China) may be the major explanation for the lower accuracy of the model for SS responses than for RS responses. In the future, content-related features should also be developed for spontaneous speech.

The features selected the first time in the logistic model differed according to item-types. The results support the item-type-specific model approach adopted in this paper; item-type-specific models can assign strong weights to the item-type-specific features that are most important.

As shown in Tables 5 - 8, the combination of non-ASR and ASR features could not achieve any further improvement over the model consisting only of ASR based features. However, in all cases, the non-ASR based model did lead to some improvement over the baseline. The magnitude of this improvement was greater in SS items than RS items; in particular, it was greatest among the SS items in the India data set. This difference may be due to the different distributions of the NS types among the data sets. The non-ASR based features can cover only limited types of NS responses such as non-responses and responses with background noise, and the proportion of these types is much higher among the SS

responses from India.

In addition, in RS items, the poor performance of the combined model may be related to the high performance of TER. The stepwise regression analysis showed that the combined model did not select any of non-ASR based features.

7 Conclusion

In this study, filtering models were implemented as a supplementary module to an automated proficiency scoring system. Due to the difference in the available features and proportion of NS responses, item-type specific models were trained.

The item-types heavily influenced the overall characteristics of the filtering models. First, the proportion of NS responses was significantly different according to item-type; it was much higher in spontaneous speech items than recited speech items. Secondly, the word error rate feature group was only available for recited speech. Although the word error rate feature group contained three features, they improved the performance of the filtering model significantly.

ASR feature based models outperformed non-ASR feature based models, but non-ASR based features may be useful for new tests. Finally, experiments demonstrated that the country-specific models using the ASR-based features are relatively robust to responses from a different country. This result suggests that this approach can generalize well to speakers from different countries.

In this study, large numbers of features (42 for RS items and 39 for SS items) were used in the model training, but some features were conceptually similar and not all of them were significantly important; the logistic regression analysis using training data showed that there was no significant improvement after selecting 5 features for RS items and 13 features for SS items. Use of non-significant features in the model training may result in the overfitting problems. In future research, the features will be classified into subgroups based on their conceptual similarities; groups of features with high intercorrelations will be reduced to include only the best performing feature in each group. Thus, based on careful pre-selection procedures, only high performing features will be selected, and the model will be re-

trained.

In addition, many different types of NS responses were lumped into one big category (NS); this may increase the confusion between scorable and non-scorable responses and decrease the model's performance. Some of NS types have very different characteristics compared to other NS types and this fact caused critical differences in the feature values. For instance, non-responses contained zero or close to zero words, whereas non-English responses and off-topic responses typically had a word count similar to scorable responses. This difference may reduce the effectiveness of this feature. In order to avoid this type of problem, we will classify NS types into small numbers of subgroups and build a separate model for each subgroup.

References

- Joon-Hyuk Chang and Nam Soo Kim. 2003. Voice activity detection based on complex Laplacian model. *Electronics Letters*, 39(7):632–634.
- Joon-Hyuk Chang, Nam Soo Kim, and Sanjit K. Mitra. 2006. Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing*, 54(6):1965–1976.
- Nivja H. De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.
- Yufeng Ding and Jeffrey S. Simonoff. 2008. An investigation of missing data methods for classification trees. *Statistics Working Papers Series*.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, volume 11.
- Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25:282–306, April.
- Guojun Lu and Templar Hankinson. 1998. A technique towards automatic audio classification and retrieval. In *Proceedings of the 4th International Conference on Signal Processing*, volume 2, pages 1142–1145.
- NIST. 2009. The NIST SPeech Quality Assurance (SPQA) Package Version 2.3. from <http://www.nist.gov/speech/tools/index.htm>.
- Jong Won Shin, Hyuk Jin Kwon, Suk Ho Jin, and Nam Soo Kim. 2005. Voice activity detection based on generalized gamma distribution. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 781–784.
- Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE Signal Processing Letter*, 6(1):1–3.
- Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.