Interspeech 2018
2-6 September 2018, Hyderabad

# Overview of the 2018 Spoken CALL Shared Task

*Claudia Baur*[1], *Andrew Caines*[2], *Cathy Chua*[3], *Johanna Gerlach*[1], *Mengjie Qian*[4]
*Manny Rayner*[1], *Martin Russell*[4], *Helmer Strik*[5], *Xizi Wei*[4]

[1]FTI/TIM, University of Geneva, Switzerland
[2]Automated Language Teaching & Assessment Institute, University of Cambridge
[3]Independent researcher
[4]Department of Electronic, Electrical and Systems Engineering, University of Birmingham
[5]Centre for Language Studies (CLS), Radboud University Nijmegen

claudia.baur@bluewin.ch, apc38@cam.ac.uk, cathyc@pioneerbooks.com.au,
Johanna.Gerlach@unige.ch, MXQ486@student.bham.ac.uk, Emmanuel.Rayner@unige.ch,
m.j.russell@bham.ac.uk, w.strik@let.ru.nl, XXW395@student.bham.ac.uk

## Abstract

We present an overview of the second edition of the Spoken CALL Shared Task. Groups competed on a prompt-response task using English-language data collected, through an online CALL game, from Swiss German teens in their second and third years of learning English. Each item consists of a written German prompt and an audio file containing a spoken response. The task is to accept linguistically correct responses and reject linguistically incorrect ones, with "linguistically correct" defined by a gold standard derived from human annotations. Scoring was performed using a metric defined as the ratio of the relative rejection rates on incorrect and correct responses. The second edition received eighteen entries and showed very substantial improvement on the first edition; all entries were better than the best entry from the first edition, and the best score was about four times higher. We present the task, the resources, the results, a discussion of the metrics used, and an analysis of what makes items challenging. In particular, we present quantitative evidence suggesting that incorrect responses are much more difficult to process than correct responses, and that the most significant factor in making a response challenging is its distance from the closest training example.

**Index Terms**: CALL, shared tasks, speech recognition, metrics

## 1. Introduction

The Spoken CALL Shared Task is a series of open challenges jointly organised by the University of Geneva, the University of Birmingham, Radboud University and the University of Cambridge[1]. The task is based on data collected from a speech-enabled online tool which has been used to help young Swiss German teens practise skills in English conversation [1, 2, 3]. Items are prompt-response pairs, where the prompt is a piece of German text and the response is a recorded learner English audio file. The task is to label pairs as "accept" or "reject", accepting responses which are grammatically and linguistically correct to match a set of hidden gold standard answers as closely as possible; many examples are shown in [4]. Results are scored using a metric, $D$, formally defined in §2.3, which rewards maximisation of the difference between the system's reaction to correct and incorrect student responses. The first edition of the task was announced at LREC 2016 [5], with training data released in July 2016 and test data in March 2017, and attracted 20 entries. Results, including seven papers, were presented at the SLaTE workshop in August 2017 (http://regulus.unige.ch/spokencallsharedtask/; [4, 6, 7, 8, 9, 10, 11]).

Here, we summarise results from the second edition of the task. As well as providing new test data, we approximately doubled the amount of training data, annotated it much more systematically, and released improved versions of the accompanying resources. In particular, we made available the open source Kaldi recogniser [12] developed by the University of Birmingham, which achieved the best performance on the original task, together with versions of the training and test data pre-processed through this recogniser. The second edition attracted 18 entries. The quality was considerably improved; by the $D$ metric, the worst entry from the second edition scored better than the best entry from the first edition, which served as the baseline here, and the best entry's score was nearly four times higher than the baseline ($D = 19.088$ versus $D = 5.343$).

The rest of the paper is structured as follows. Section 2 describes the data, resources and metric, and Section 3 presents the main results. Section 4 analyses what the results say about items that were easy or difficult for systems to process correctly. Section 5 discusses issues relating to metrics and concludes.

## 2. Data, resources and metric

### 2.1. Data

As with the first edition of the shared task, the data comes from an online English course developed for German-speaking Swiss teenagers in their second or third year of English lessons [1, 2, 3, 4]. For the second edition we annotated a new subset of the corpus consisting of 6698 student utterances to serve as additional training data. This new data was selected in a similar way to the first training set, to be balanced and representative of the collected data, with the additional constraint that there should be no overlap of individual students between the first task and second task. Speech data were processed through the two best speech recognisers from the first shared task [6, 8] after which the two sets of output transcriptions were merged and cleaned up by transcribers at the University of Geneva.

The cleaned, merged transcriptions were processed through four of the best assessment systems from the first shared task [6, 8, 7, 9] to give accept/reject decisions for the language criterion. The training data could then be divided into three groups according to the agreement among the four systems. There was

[1]Work at Cambridge was funded by Cambridge Assessment English. Work at Geneva was partially funded by the Swiss National Science Foundation under grant IZCOZ0_177065.



10.21437/Interspeech.2018-97

unanimous (4–0) agreement for 70% of the utterances, 3-to-1 agreement for 22%, and a 2–2 split for the remaining 8%.

We randomly selected 200 utterances from each group, which were then independently annotated by three English native speaker annotators familiar with the domain. For the 4–0 group, we found that the humans agreed about 98% with the machines, which was about as well as they agreed with each other. We consequently decided to consider this portion of the data reliably judged.

The three human annotators independently judged the remaining 3–1 and 2–2 portions of the data for both language and meaning. The utterances on which pairs of annotators disagreed were extracted and independently re-judged by the annotators, together with an additional 20% control set on which the annotators had agreed. Through this process the annotators discussed and resolved many differences of opinion, or simply agreed to disagree on certain linguistic specifics. At the end of the annotation process, the training items were divided into three bands by descending reliability, labelled as **A**, **B** and **C**:

**A** (5526 utterances). Either the machines are 4–0 and at least one human supports them, or the machines are 3–1 and all three humans support them.

**B** (873 utterances). All three humans agree, and either one or two machines support them.

**C** (299 utterances). Remaining cases.

The consolidated 'language' accept/reject judgement was defined to be the majority machine judgement in **A** and the majority human judgement in **B** and **C**.

The machines were not set to provide semantic/meaning judgements (cf. §2.3), though if they marked an item as correct for language, this implied that it was also correct for meaning. We verified that the **A** utterances were straightforward to annotate for meaning, and we decided that it was enough for one person to carry out meaning annotation on this set.

Otherwise the meaning annotations were combined as follows. If the machines unanimously rejected an item, it was marked as incorrect for meaning. In all remaining cases the accept/reject value for meaning was the majority decision of the three human annotators.

Due to the necessity of keeping the material secret from potential competitors in the task, we were forced to use a simpler methodology for annotating the test data. Two native speakers of English independently annotated 1800 items previously not used in the Shared Task, chosen so that the subjects were not ones who appeared in other training or test sets. Items were removed wherever the two annotators disagreed or at least one had flagged their judgement as 'uncertain', and 1000 items were randomly chosen from the remainder to be used as the test set.

After entries had been submitted, a final annotation pass was carried out, where one native English speaker with moderate German and one native German speaker with near-native English jointly listened to the items again, focusing in particular on examples where many entries disagreed with the initial judgement. This produced a small number of corrections. The final annotated test set, including links to audio files, is posted on the 'Test data' tab of the task site.

### 2.2. Other resources

The University of Birmingham group made available the Kaldi recogniser and the response grammar used for their winning entry in the 2017 edition of the Shared Task, to act as the baseline in the new task. Both resources are described in detail in [6]. For the benefit of groups who only wished to explore the language processing aspects of the task, we processed test and training data through the baseline recogniser, and supplied versions of the task metadata which included the recognition results produced. Finally, we supplied a Python script which instantiated a minimal example of a system capable of performing the shared task, using the "pre-recognised" set of metadata and the baseline response grammar.

### 2.3. Metrics

We define $D$ and other metrics as follows. As explained in the previous paper, there are two main intuitions. First, $D$ should measure the difference between the system's reaction to correct and incorrect responses; second, it should give a larger penalty to a false accept if the response is semantically as well as syntactically incorrect. We call false accepts of semantically incorrect responses "gross false accepts" and false accepts of semantically correct responses "plain false accepts".

We assume that we are given a set of annotated prompt/response interactions, where in each case the annotations show whether the response was correct or incorrect, both syntactically and semantically, and whether it was accepted or rejected. We write $CA$ for the number of correct accepts, $CR$ for the number of correct rejects, $PFA$ for the number of plain false accepts, $GFA$ for the number of gross false accepts and $FR$ for the number of false rejects. We set $FA = PFA + k.GFA$ for some constant $k$, weighting gross false accepts $k$ times more heavily than plain false accepts, and $Z = CA + CR + FA + FR$. Then we write $C_A = \frac{CA}{Z}$, $C_R = \frac{CR}{Z}$, $F_A = \frac{FA}{Z}$, $F_R = \frac{FR}{Z}$ and define metrics in terms of the four quantities $C_A$, $C_R$, $F_A$, $F_R$, which total to unity. Looking first at traditional metrics, we consider precision ($P = \frac{C_A}{C_A+F_A}$), recall ($R = \frac{C_A}{C_A+F_R}$), F-measure $F = \frac{2PR}{P+R}$) and scoring accuracy ($SA = C_A + C_R$).

Generally, all of the above metrics are based on the idea of minimising some kind of error. In contrast, $D$, the metric based on differential response which we used for the task, is defined as the ratio of the relative correct reject rate (the reject rate on incorrect responses) to the relative false reject rate (the reject rate on correct responses). We put $RC_R = \frac{C_R}{C_R+F_A}$ and $RF_R = \frac{F_R}{F_R+C_A}$, then define

$$D = \frac{RC_R}{RF_R} = \frac{C_R/(C_R + F_A)}{F_R/(F_R + C_A)} = \frac{C_R(F_R + C_A)}{F_R(C_R + F_A)}$$

When announcing the task, we said that entries would be scored using $D$, with a $k$ value of 3 and the added requirement that at least 25% of all incorrect utterances should be rejected.

## 3. Results

We received a total of 18 entries, 12 using data pre-processed through the baseline Kaldi recogniser and 6 using a custom recogniser. Table 1 presents the results. Two points immediately stand out. First, the marked improvement of all metrics compared with those in the first edition of the Shared Task: the top $D$ score is over 19, whereas it was under 5 in the first edition. ('Results' tab of http://regulus.unige.ch/spokencallsharedtask/; [4]). Second, the poor agreement of the rank orders based on the $D$ and $F$ metrics: LLL gets the best score on the $D$ metric, but DDD gets the best score on the $F$ metric. We return to both these issues in the final section. First, however, we consider the question of what makes an item easy or difficult to process.

| Id | Rec | Pr | R | F | SA | RCR | RFR | D | $D_A$ | $D_{full}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| LLL | Text | 0.742 | 0.984 | 0.846 | 0.760 | 0.305 | 0.016 | **19.088** | 1.417 | 5.200 |
| HHH | Speech | 0.758 | 0.975 | 0.853 | 0.772 | 0.342 | 0.025 | 13.492 | 1.481 | 4.470 |
| KKK | Text | 0.777 | 0.967 | 0.862 | 0.787 | 0.399 | 0.033 | 11.965 | 1.608 | 4.386 |
| GGG | Speech | 0.773 | 0.967 | 0.859 | 0.782 | 0.381 | 0.033 | 11.424 | 1.561 | 4.223 |
| III | Speech | 0.764 | 0.967 | 0.853 | 0.774 | 0.364 | 0.033 | 10.909 | 1.519 | 4.071 |
| FFF | Speech | 0.893 | 0.936 | 0.914 | 0.871 | 0.689 | 0.064 | 10.764 | 3.009 | 5.691 |
| DDD | Speech | 0.896 | 0.935 | **0.915** | 0.873 | 0.700 | 0.065 | 10.714 | 3.116 | 5.778 |
| BaselinePerfectRec | Text | 0.961 | 0.913 | 0.936 | 0.907 | 0.889 | 0.087 | 10.256 | 8.220 | 9.182 |
| EEE | Speech | 0.885 | 0.924 | 0.904 | 0.856 | 0.669 | 0.076 | 8.804 | 2.793 | 4.958 |
| OOO | Text | 0.759 | 0.955 | 0.846 | 0.764 | 0.362 | 0.045 | 7.993 | 1.497 | 3.459 |
| JJJ | Text | 0.797 | 0.941 | 0.863 | 0.793 | 0.458 | 0.059 | 7.804 | 1.736 | 3.681 |
| RRR | Text | 0.842 | 0.920 | 0.880 | 0.823 | 0.592 | 0.080 | 7.397 | 2.254 | 4.083 |
| QQQ | Text | 0.840 | 0.916 | 0.876 | 0.818 | 0.588 | 0.084 | 7.001 | 2.224 | 3.945 |
| MMM | Text | 0.794 | 0.933 | 0.858 | 0.785 | 0.445 | 0.067 | 6.677 | 1.682 | 3.351 |
| BBB | Text | 0.882 | 0.889 | 0.886 | 0.832 | 0.673 | 0.111 | 6.079 | 2.718 | 4.065 |
| AAA | Text | 0.881 | 0.889 | 0.885 | 0.831 | 0.672 | 0.111 | 6.068 | 2.708 | 4.053 |
| NNN | Text | 0.798 | 0.921 | 0.855 | 0.783 | 0.470 | 0.079 | 5.971 | 1.737 | 3.221 |
| CCC | Text | 0.873 | 0.891 | 0.882 | 0.825 | 0.643 | 0.109 | 5.885 | 2.498 | 3.834 |
| PPP | Text | 0.802 | 0.912 | 0.853 | 0.784 | 0.497 | 0.088 | 5.648 | 1.813 | 3.200 |
| Baseline | Text | 0.916 | 0.855 | 0.884 | 0.834 | 0.777 | 0.145 | 5.343 | 3.824 | 4.520 |

Table 1: *Results for 18 anonymised submissions and two baseline systems. "Rec" = recogniser used ("Text" = data pre-recognised using baseline Kaldi recogniser, "Speech" = other recogniser), "Pr" = precision, "R" = recall, "F" = F-measure, "SA" = scoring accuracy, "RCR" = relative correct rejections, "RFR" = relative false rejections, "D" = D-measure, "$D_A$" = D-measure on accepts, "$D_{full}$" = geometrical mean of D and DA, "Baseline" = system with baseline Kaldi recogniser and baseline XML grammar; "BaselinePerfectRec" = system with input from transcriptions and baseline XML grammar.*

## 4. What makes items difficult to label?

The test set contained 750 "correct" examples (the gold standard says the system should accept) and 250 "incorrect" examples (the gold standard says the system should reject). To study the degree of difficulty of the test items, we ordered both the correct and incorrect subsets of the test data by the number of entries supplying the wrong judgement, assuming that, at least to a first approximation, examples where many systems gave an incorrect judgement were hard to label correctly, and examples where few systems gave an incorrect judgement were easy. The approximate distribution is indicated in the first two columns of Table 2, where we aggregate the data by dividing the sets into three bands labelled "easy" (0–3 wrong judgements), "medium" (4–9 wrong judgements) and "hard" (10–18 wrong judgements).

To better understand the factors which might make items easier or harder, we performed an annotation of the test data. Three annotators listened to each audio file separately using an online tool and categorised them on the following six scales:

**Incomprehensible** Was any word spoken by the student incomprehensible to you? (yes/no)

**Pronunciation** Was any word spoken by the student clearly mispronounced, in the sense that at least one English sound was clearly substituted by a different and incorrect English sound? (yes/no)

**Stuttering/repetition** Did the student stutter, repeat themself, or in some other way clearly change their mind about what they were going to say?

**Crosstalk** Could you hear anyone other than the student talking? (yes/no)

**Strong non-speech noise** Could you hear any non-speech noises, for example background noise, comparable in loudness with the student's speech? (yes/no)

**Faint** Was the volume of the student's speech clearly much fainter than usual? (yes/no)

We also added the following automatically computed labels:

**OOV** The transcription contains an OOV word, i.e. a word not in the training data or grammar.

**d(gr)** Edit distance, in words, between the transcription and the closest in-coverage sentence in the grammar.

**d(tr)** Edit distance, in characters, between the transcription and the closest correct example in the training data.

**Word errors** The number of word errors in the 1-best recognition result produced by the baseline recogniser.

Table 2 shows the distribution of the resulting metrics over the three different bands. We used Light's $\kappa$ to estimate inter-annotator agreement; we obtained substantial agreement ($\kappa = 0.65$) for 'Stuttering/repetition' and fair agreement ($0.34 \leq \kappa \leq 0.4$) for 'Incomprehensible', 'Pronunciation', 'Crosstalk' and 'Faint'. The values are similar to those we obtained in the first edition of the Spoken CALL Shared Task (Table 7 of [4]).

We draw the following tentative conclusions from this data. First, incorrect utterances are much harder to process than correct utterances. In particular, they have over ten times as many incomprehensible words, over three times as many recognition errors, and over two and a half times as many clear pronunciation errors. On the linguistic metrics, they contain out-of-vocabulary words over fifteen times more frequently, their character edit distance to the closest training example is over five times as high, and their word edit distance to the closest in-coverage grammar example is over two and a half times as high.

| #Bad | #Utts | Incom | Pron | Stutt | xTalk | SNSN | Faint | OOV | d(gr) | d(tr) | #WER |
|---|---|---|---|---|---|---|---|---|---|---|---|
| "Correct" examples (should accept) | | | | | | | | | | | |
| 0–3 | 663 | 0.60 | 12.37 | 2.41 | 3.02 | 1.66 | 3.47 | 0.90 | 0.58 | 0.84 | 0.11 |
| 4–9 | 56 | 1.79 | 19.64 | 10.71 | 1.79 | 1.79 | 1.79 | 7.14 | 1.05 | 2.18 | 1.57 |
| 10–18 | 31 | 6.45 | 25.81 | 12.90 | 16.13 | 6.45 | 9.68 | 9.68 | 1.87 | 4.29 | 3.39 |
| all | 750 | 0.93 | 13.47 | 3.47 | 3.47 | 1.87 | 3.60 | 1.73 | 0.67 | 1.08 | 0.35 |
| "Incorrect" examples (should reject) | | | | | | | | | | | |
| 0–3 | 102 | 17.65 | 38.24 | 3.92 | 7.84 | 9.80 | 7.84 | 35.29 | 2.45 | 7.35 | 1.39 |
| 4–9 | 69 | 18.84 | 37.68 | 4.35 | 0.00 | 2.90 | 5.80 | 31.88 | 2.20 | 6.62 | 1.61 |
| 10–18 | 79 | 10.13 | 34.18 | 7.59 | 1.27 | 2.53 | 3.80 | 22.78 | 1.15 | 3.58 | 0.87 |
| all | 250 | 15.60 | 36.80 | 5.20 | 3.60 | 5.60 | 6.00 | 30.40 | 1.97 | 5.96 | 1.29 |

Table 2: *Possible indicators of difficulty, broken down by number of entries out of 18 assigning the wrong label. First two columns: "#Bad" = number of entries assigning wrong label. "#Utts" = number of examples in group. Middle seven columns: percentage of items displaying seven types of possible problems. "Incom" = at least one incomprehensible word, "Pron" = at least one mispronounced word, "Stutt" = stuttering etc, "xTalk" = crosstalk, "SNSN" = strong non-speech noise, "Faint" = low volume in speech, "OOV" = at least one out of vocabulary word. Final three columns: average per-utterance value for three metrics. "d(gr)" = word edit distance to closest in-grammar example; "d(tr)" = character edit distance to closest correct training example; "#WER" = number of word errors in recognition hypothesis from baseline recogniser.*

Comparing the easy, medium and hard bands in both groups, the clearest correlations seem to be in the linguistic metrics. For correct examples, hard utterances are consistently further from the grammar and training data; for incorrect examples, the reverse holds, and the hard utterances are consistently *closer* to the grammar and training data. This pattern fits the expectation that an example close to the training data will tend to be accepted, and one far from it will tend to be rejected. 'Word errors' appears to correlate strongly with difficulty for correct examples, but not for incorrect examples.

Carrying out a pair of ANOVA analyses supports these intuitive impressions. For the correct examples, all predictors except 'Pronunciation', 'Strong non-speech noise' and 'Faint' give correlations with '#Bad' significant at $p < 0.001$; 'Pronunciation' is significant at $p < 0.002$. By far the most important predictor is 'Word errors', which accounts for 42% of the variance in '#Bad'. The other predictors account for another 15%, giving a total of 57%. This contrasts sharply with the incorrect examples, where only 'd(gr)' is significant at $p < 0.001$, and accounts for 10% of the variance in '#Bad'. The other predictors together account for another 11% of the variance, giving a total of only 21%. An independent analysis using linear regression produced similar results.

## 5. Metrics and further directions

We are encouraged by the large improvement in the scores between the first and second editions of the Spoken CALL Shared Task. The fact that $D$ and $F$ give different results is, however, disquieting and evidently requires investigation. To begin, when we look at the definition of $D$ in § 2.3, we see that it only measures the informativeness of system rejects. ($D$ correlates very well with recall but poorly with $F$). The best entry, LLL, wins because it rejects 19.1 times as often on incorrect utterances as on correct utterances. The corresponding ratio for DDD, the entry which got the best $F$ score, is 10.7. An obvious question is whether the metric should not also take into account the informativeness of system accepts. It is straightforward to define a metric like $D$ which does this, and say that $D_A$ is the ratio of the relative correct acceptance rate ($RC_A$) to the relative false acceptance rate ($RF_A$). Using the notation of § 2.3, we put $RC_A = \frac{C_A}{F_R+C_A}$ and $RF_A = \frac{F_A}{C_R+F_A}$, then define

$$D_A = \frac{RC_A}{RF_A} = \frac{C_A/(F_R+C_A)}{F_A/(C_R+F_A)} = \frac{C_A(C_R+F_A)}{F_A(F_R+C_A)}$$

As with $D$, high values of $D_A$ are good. We see two plausible ways to continue this argument. First, one could say that, for the reasons discussed in the previous section, correct utterances are much easier to process reliably than incorrect ones, and it is consequently much easier to lower the false reject rate than the false accept rate. Experience also suggests that students tend to be much more unhappy about false rejects than about false accepts. As an initial goal, it consequently seemed sensible to choose a metric which focused on the reliability of rejects. From this point of view, $D$ is an appropriate metric.

A second point of view, which we tend to favour, would be that the results presented here suggest that the goal of producing systems which give reliable behaviour on rejects has now been more or less achieved. If we are to make further progress, it will be productive to replace $D$ with a metric which measures the reliability of both rejects and accepts; it is worth adding that *teachers* tend to be more worried about false accepts, since they can instil bad habits in their students. A natural way to perform the combination is to define a metric we can call $D_{full}$, which is the geometric mean of $D_A$ and $D$. It is defined by the formula

$$D_{full} = \sqrt{D_A D} = \sqrt{\frac{C_A(C_R+F_A)}{F_A(F_R+C_A)}\frac{C_R(F_R+C_A)}{F_R(C_R+F_A)}}$$

$$= \sqrt{\frac{C_A C_R}{F_A F_R}}$$

The values of $D_A$ and $D_{full}$ for each entry can be found in the last two columns of Table 1. According to the $D_{full}$ metric, DDD is slightly better than LLL, and BaselinePerfectRec is a great deal better than both of them. This seems to us an intuitively more reasonable answer than either the result from $D$ (LLL is much better than both DDD and BaselinePerfectRec) or the result from $F$ (BaselinePerfectRec is somewhat better than DDD, which is a great deal better than LLL).

If there is sufficient interest, we are considering organising a third edition of the Spoken CALL Shared Task, this time using $D_{full}$ as the ranking metric.

# 6. References


[1] C. Baur, M. Rayner, and N. Tsourakis, "A textbook-based serious game for practising spoken language," in *Proceedings of ICERI 2013*, Seville, Spain, 2013.

[2] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, no. 2, 2014.

[3] C. Baur, "The potential of interactive speech-enabled CALL in the Swiss education system: A large-scale experiment on the basis of English CALL-SLT," Ph.D. dissertation, University of Geneva, 2015.

[4] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 Spoken CALL Shared Task," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[5] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken CALL?" in *Proceedings of LREC 2016*, Portorož, Slovenia, 2016.

[6] M. Qian, X. Wei, P. Jančovič, and M. Russell, "The University of Birmingham 2017 SLaTE CALL Shared Task systems," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[7] A. Magooda and D. Litman, "Syntactic and semantic features for human like judgement in spoken call," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[8] Y. R. Oh, H.-B. Jeon, H. J. Song, B. O. Kang, Y.-K. Lee, J.-G. Park, and Y.-K. Lee, "Deep-Learning based automatic spontaneous speech assessment in a data-driven approach for the 2017 SLaTE CALL Shared Challenge," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[9] K. Evanini, M. Mulholland, E. Tsuprun, and Y. Qian, "Using an automated content scoring system for spoken CALL responses: The ETS submission for the Spoken CALL Challenge," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[10] N. Axtmann, C. Mehret, and K. Berkling, "The CSU-K rule-based pipeline system for Spoken CALL Shared Task," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[11] A. Caines, "Spoken CALL Shared Task system description," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[12] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.