The University of Birmingham 2017 SLaTE CALL Shared Task Systems

Mengjie Qian, Xizi Wei, Peter Jančovič and Martin Russell

Department of Electronic, Electrical and Systems Engineering, University of Birmingham

Introduction

The 2017 SLaTE Spoken CALL Shared Task [1] was led by the University of Geneva with support from the University of Birmingham and Radboud University.
Aim: label prompt and response pairs as "accept" or "reject".
Data: recordings of English responses from German-speaking Swiss teenagers interacting with the CALL-SLT system [2]. A development set, ST-DEV, of 5222 recordings and a test set, ST-TST, of 996 recordings were released.
System structure:

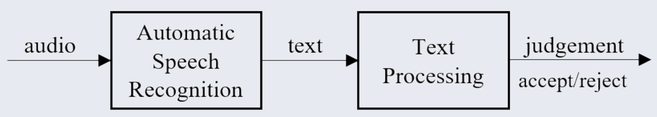

Figure 1: Structure of the system.

Scoring Metric

Comparing the system's judgement with the language and meaning gold standards, each response falls into one of the five categories described in Table 1.

English	Meaning	Judgment	Category
✓	✓	Accept	Correct Accept (CA)
✓	✓	Reject	False Reject (FR)
✗	✓	Reject	Correct Reject (CR)
✓	✗		
✗	✗		
✗	✓	Accept	Plain False Accept (PFA)
✓	✗	Accept	Gross False Accept (GFA)
✗	✗		

Table 1: Categories of Results

The evaluation of the overall quality of the systems is performed using a differential response score, D.

$$D = \frac{CR/(CR+FA)}{FR/(FR+CA)} = \frac{CR(FR+CA)}{FR(CR+FA)}, \quad (1)$$

where $FA = PFA + k \cdot GFA$, with k being a weighting factor that causes gross false accepts to have a more prominent effect ($k = 3$).

Automatic Speech Recognition

The provided baseline ASR is a hybrid deep neural network - hidden Morkov model (DNN-HMM) built using Kaldi. In cross-validation evaluations, this system achieved an average WER of 14.03%.

Training Data Selection

- ST-DEV: 5222 recordings, 4.8 hours, age ranging between 12 to 15 years.
- AMI: adults meeting recordings, 16.07 hours, mostly non-native speakers, 100% vs. 50% vs. **20%**.
- PF-STAR German: German children aged 10-13, 3.38 hours of read speech.

Acoustic Model

- Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transform (MLLT)
- feature-space MLLR (**speaker-id = utterance-id** vs. global speak-id)
- DNN adaptation

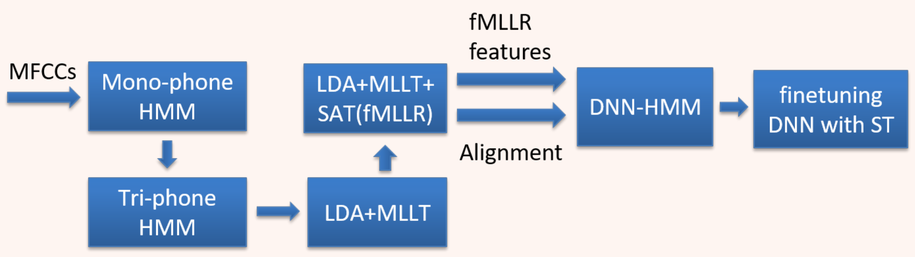

Figure 2: Structure of the ASR system.

Language Model

- back-off 3-gram language model trained on all the ST-DEV transcriptions

Results

- 9.27% WER average over 10-fold cross-validation experiments on ST-DEV
- 15.63% WER on ST-TST

Text Processing

The baseline text processing system uses a reference grammar and it gets D score of 2.358 and 1.694 on the Kaldi and Nuance baseline ASR output, respectively.

Pre-processing

1. Remove superfluous words
   - "ah, beh, mm, uh, um, . . . " or "hello, hi, ok, and, yes, oh, . . . "
2. Remove repetition
   - i would like tickets for tomorrow tomorrow → *repeated words*
   - can i have a ticket for friday night can i have a ticket for friday night → *repeated sentence*
   - i have i want a ticket for trafalgar square → *repeated meaning*

Expanded Reference Grammar

We expanded the reference grammar in the baseline text processing system using the similar method described in paper [3].

```
PromptTemplate i_want GERMAN ENGLISH
Text       Frag : Ich möchte GERMAN
Response   do you have ENGLISH
Response   i ( want | would like )
           ENGLISH ?please
EndPromptTemplate
```

Figure 3: Response template.

A few response templates were created according to ST-DEV transcriptions and these templates were applied to different situations to create full responses list for different prompts.

Fusion

Step1: *Format input data (output of text processing), convert 2-class (Accept, Reject) data into a matrix.*

$$T = [\,R\ R\ A\ R\ A \ldots] \Rightarrow score(x) = \begin{bmatrix} 0\,0\,1\,0\,1\ldots \\ 1\,1\,0\,1\,0\ldots \end{bmatrix} \Rightarrow \log(score(x) + \epsilon)$$

Step2: *Use linear logistic regression to train weights on K systems.*
Step3: *Apply weights on test data.*

$$score_c(x) = \textstyle\sum_{i=1}^{K} w_{c,i} \cdot score_{c,i}(x)$$

Step4: *Choose class which has higher score.*

$$class(x) = \arg\max_c score_c(x)$$

Submissions

For the final evaluation on ST-TST we submitted results from three systems:

- Submission 1 consists of our best ASR system, plus the expanded TP. The optimal parameters of ASR were estimated over 10-fold cross-validation experiments.
- Submission 2 is the result of fusing the outputs of six separate systems using linear logistic regression [4]. The systems all use our expanded TP with four variants of the ASR from Submission 1, the Kaldi baseline ASR and Nuance ASR.
- Submission 3 combines Nuance ASR with the expanded TP.

Submission 1, 2 and 3 achieved D scores of 4.71, 4.766 and 2.533, respectively [5].

References

[1] Claudia Baur, Johanna Gerlach, Emmanuel Rayner, Martin Russell, and Helmer Strik. A shared task for spoken CALL? In *Proc. Language Resources and Evaluation Conf. (LREC)*, 2016.

[2] Claudia Baur. *The Potential of Interactive Speech-Enabled CALL in the Swiss Education System: A Large-Scale Experiment on the Basis of English CALL-SLT.* PhD thesis, Université de Genève, 2015.

[3] Emmanuel Rayner, Claudia Baur, Cathy Chua, and Nikolaos Tsourakis. Supervised learning of response grammars in a spoken CALL system. In *Proc. Workshop on Speech and Language Technology in Education (SLaTE)*, 2015.

[4] Niko Brümmer. Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores-Tutorial and user manual. *Software available at http://sites. google. com/site/nikobrummer/focalmulticlass*, 2007.

[5] Spoken CALL shared task official website. `https://regulus.unige.ch/staging_spokencallsharedtask/`.