Evaluation of Invalid Input Discrimination Using Bag-of-Words for Speech-Oriented Guidance System

Haruka Majima*, Rafael Torres*, Hiromichi Kawanami*, Sunao Hara**, Tomoko Matsui***, Hiroshi Saruwatari*, Kiyohiro Shikano*
*Graduate School of Information Science, Nara Institute of Science and Technology, Japan
**Graduate School of Natural Science and Technology, Okayama University, Japan
***Department of Statistical Modeling, The Institute of Statistical Mathematics, Japan

Abstract: We investigate a discrimination method for invalid and valid inputs based on machine learning using bag-of-words comprised from automatic speech recognition result as a classification feature. Changing the amount of training data, we elucidate that using 3000 of them (approx. 2 weeks of system inputs) shows enough classification performance.

1. Research background

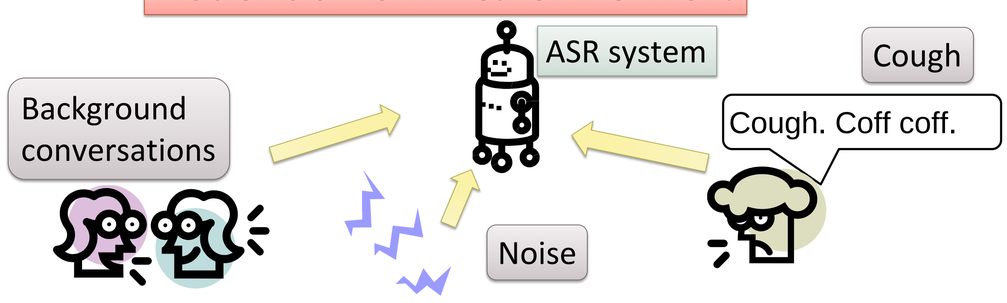
- Automatic speech recognition (ASR) has been widely applied to
  - ✓ dictation, Voice Search, car navigation, etc.

Problems

- Many invalid inputs that the system unnecessarily answers
- Invalid inputs decrease the response accuracy

→ Rejection of invalid inputs is desired.

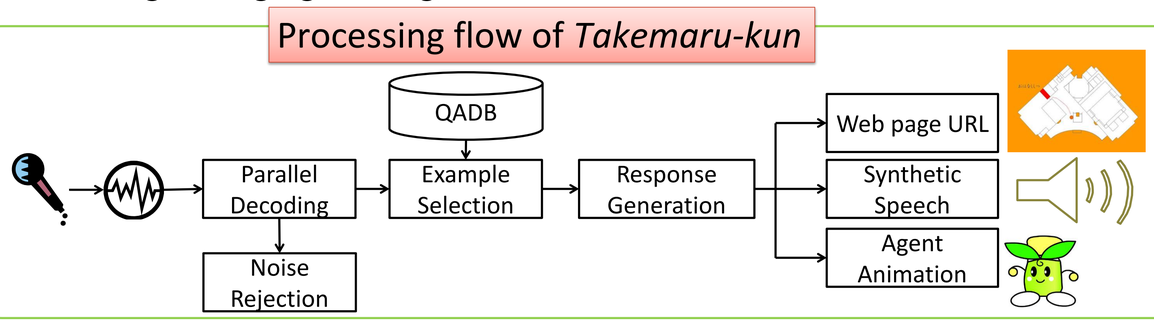
Problems of ASR in real environment



2. Speech-oriented guidance system *Takemaru-kun*

- A real-environment speech-oriented guidance system
  - ✓ placed inside the entrance hall of the *Ikoma City North Community Center,*
  - ✓ providing guidance to visitors, regarding the center facilities, services, neighboring sightseeing, weather forecast, news, etc.

Processing flow of *Takemaru-kun*



3. Detail of system inputs

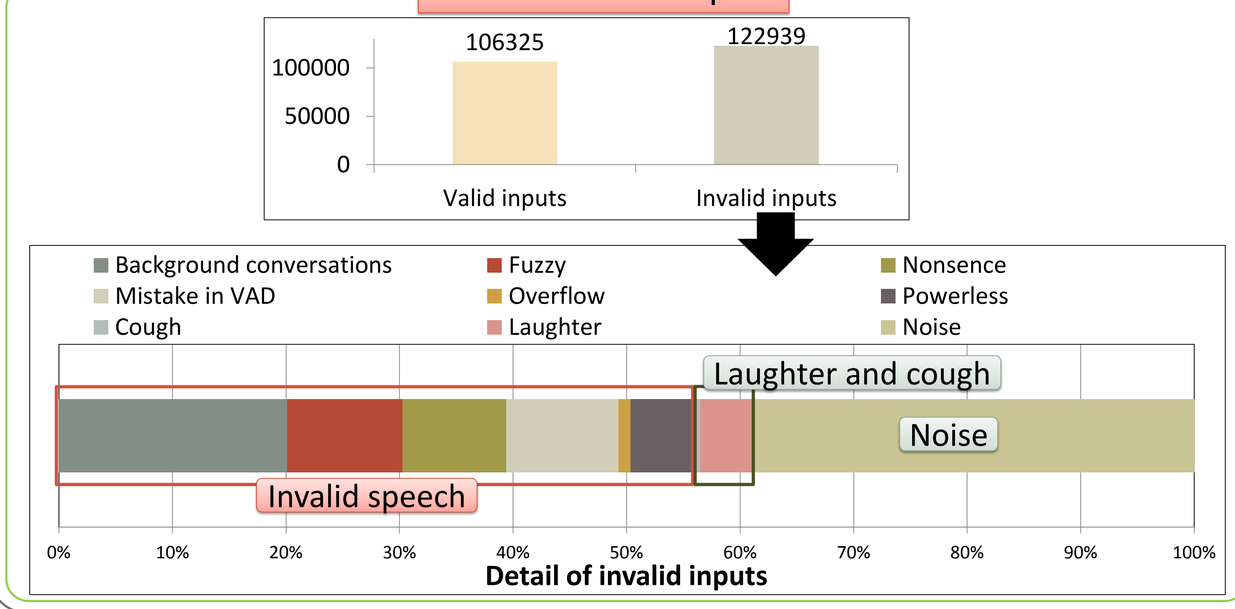
What is "invalid speech" ?

- Unintended inputs with tags of
  - ✓ background conversations, fuzzy speech, nonsense speech, mistake in VAD (voice activity detection), overflow speech, powerless speech
- Some tags of invalid inputs are overlapped

Detail of system inputs

- **Inputs** — All the inputs to the system
  - **Valid inputs** — Inputs that the system should respond
  - **Invalid inputs** — Inputs that the system should not respond
    - **Invalid speech** — Speech, but the contents are of invalid inputs
    - **Laughter and cough** — Non-speech sounds made by human
    - **Noise** — Non-speech sounds that are not made by human

Detail of invalid inputs



4. Proposed method

Employ Bag-of-Words as a classification feature

Knowledge: ASR results of valid inputs and invalid inputs have different tendency.

Examples of valid inputs

"Hello."
"Where is the restroom?"

Examples of invalid inputs

"Aaaaargh." (noise input)
"blah-blah-blah." (nonsence input)

→ Utilizing linguistic information may improve the discrimination of valid and invalid inputs.

5. Features employed for classification

1. **BOW (Bag-of-Words) vector**
   consists of frequencies of each word in a vocabulary word list, which is comprised of words from the 10-best ASR candidates of training data. The dimension of BOW vector is determined by the number of words in the word list.
2. **GMM likelihood**
   is given as the likelihood values of each utterance to six GMMs. The GMMs are trained using six kinds of data, for adults' valid speech, children's valid speech, laughter, cough, noise and other invalid speech respectively.
3. **Duration**
   is the duration of an utterance, determined by voice activity detection of Julius using amplitude and zero crossing.
4. **SNR**
   is the signal-to-noise ratio of an utterance.

6. Classification methods

SVM-based method

SVM (support vector machine)

- is a supervised learning binary classifier.
- estimates a separating hyper-plane with a maximal margin in a higher dimensional space.

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i$$

subject to $y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 - \xi_i,$
$\xi_i \geq 0 (i = 1, \ldots, l)$

$\boldsymbol{w}, b$: Parameters of discrimination function
C : Cost parameter of soft margin

ME-based method

ME (maximum entropy) models

- provide a general purpose machine learning technique for classification and prediction.
- attempts to maximize the log likelihood

$$\log P(E|D,\Lambda) = \sum_{(e,d)\in(E,D)} \log \frac{\exp\sum_i \lambda_i f_i(e,d)}{\sum_{e'}\exp\sum_i \lambda_i f_i(e',d)}$$

(E, D): training set
E: set of class labels
D: set of feature represented data points
$f_i(e, d)$: feature indicator functions

7. Experiments

To consider the amount of training data, we experimented the performance of invalid inputs discrimination using Bag-of-words, GMM likelihood, Duration and SNR by SVM or ME.

Speech database

Training data are the 15000 data of Nov. 2002 to Dec. 2002 and test data are 14881 data of Aug. 2003.

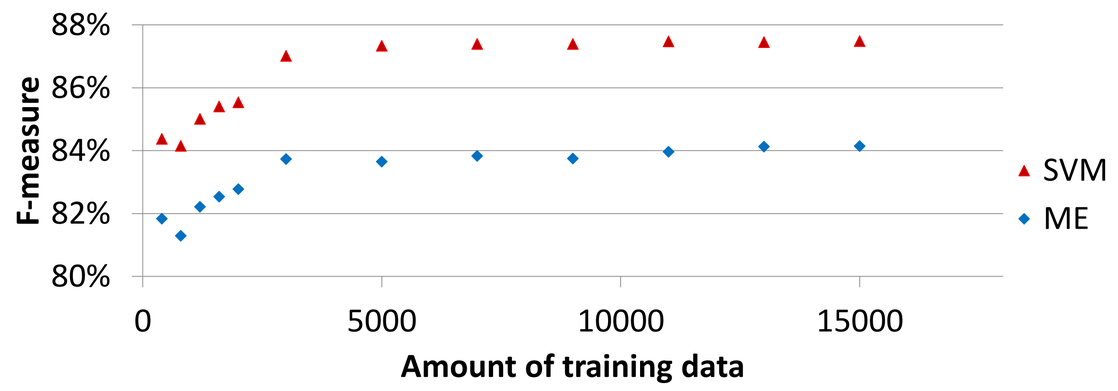
	Valid inputs	Invalid inputs	Total
Training data	9509	6491	15000
Test data	7607	7274	14881

Experimental conditions

ASR	Engine	Julius 4.2
	Language model and Acoustic model	*Takemaru-model*
	Output	10-best candidates
SVM	Tool	LIBSVM
	Kernel function	Radial Basis Function (RBF)
	Parameter C	10^{-2}, 10^{-1},…, 10^{4}
ME	Tool	Stanford Classifier 2.1.3

Results

- F-measures of SVM are always higher than that of ME.
- Both F-measures of SVM and ME are saturated using about 3000 training data.



8. Conclusions

- We investigated discrimination between invalid and valid spoken inquiries using multiple features, as Bag-of-Words, the likelihood values of GMMs, utterance durations and SNRs.
- The classification performance was better using larger amount of training data, however it saturated using only 3000 data.