Speech Understanding in a Multiple Recognizer with an Anaphora Resolution Process

Kazutaka Shimada **Akira Uzumaki** **Mai Kitajima** **Tsutomu Endo**
Department of Artificial Intelligence,
Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
{shimada, a_uzumaki, m_kitajima, endo}@pluto.ai.kyutech.ac.jp

Abstract

In this paper, we propose a simple and effective method for speech understanding. The method incorporates some speech recognizers. We use two types of recognizers; a large vocabulary continuous speech recognizer and a domain-specific speech recognizer. The multiple recognizer is a robust and flexible method for speech understanding. In this paper, we focus on two issues: (1) selection of outputs from the multiple recognizer and (2) anaphora resolution in the multiple recognizer. For the output selection, we use a simple edit distance measure computed from output sentences from each recognizer. Our method has high scalability and accuracy for the output selection. Next, we describe an anaphora resolution process using outputs from the multiple recognizer. The experimental results show the effectiveness of the proposed method.

1 Introduction

Speech understanding and dialogue systems have been developed for practical use recently. These systems often recognize user utterances incorrectly. It is important to deal with speech recognition errors for speech understanding systems. Extracting keywords and understanding an utterance using them reduce speech recognition errors (Bouwman et al., 1999; Komatani and Kawahara, 2000). Another approach is to use domain-specific grammars and linguistic models. However these methods can not handle out of domain and spontaneous utterances. One approach for the improvement is to repair recognition errors by users. There are many studies on detection of recognition errors in a speech output. Goto et al. (2005) have proposed some systems with nonverbal speech information, such as "SPEECH STARTER" and "SPEECH SPOTTER". Ogata and Goto (2005) have proposed a speech input interface with a speech-repair function. Although repairing recognition errors by humans is effective in terms of development of a speech understanding system with high recognition accuracy, it is costly for users. Combining some recognizers is one of the best approaches to improve the accuracy of speech understanding systems (Isobe et al., 2007; Utsuro et al., 2004). Utsuro et al. (2004) have obtained high accuracy by using some speech recognizers' outputs. However they dealt with word error reduction only. Although Isobe et al. (2007) have proposed a multi-domain speech recognition system based on some domain-specific recognizers, their system cannot treat out-of-domain utterances such as a chat between users. However the chat utterances often include significant information as the context of the dialogue.

In this paper we propose a simple and effective speech understanding method based on a large vocabulary continuous speech recognizer (LVCSR) and some domain-specific speech recognizers (DSSR). We call it "One Commoner and Some Specialists (OCSS) model". Figure 1 shows the outline of the model. In our system, the LVCSR is the commoner, namely domain-independent, and the DSSRs are specialists, namely domain-dependent. We focus on the difference between outputs generated from the commoner and specialists. By using this method, we can recognize speech inputs for domain-dependent with high accuracy and also handle context information in domain-independent speech inputs.

The task of this system is speech understanding for a livelihood support robot. The DSSRs recognize particular utterances about orders; e.g., order utterances from elders who need care and order utterances from nurses. We construct the grammar-based DSSR for order utterances with small vo-

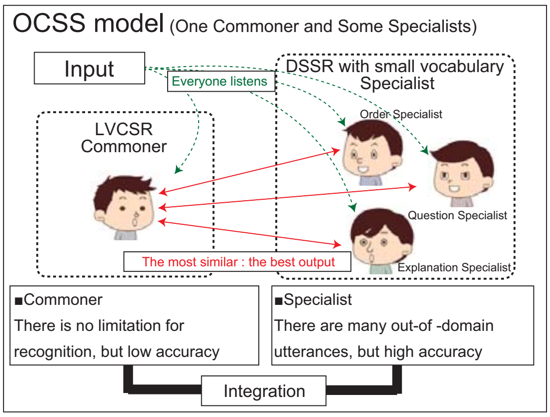

OCSS model (One Commoner and Some Specialists)
Input
Everyone listens
DSSR with small vocabulary
Specialist
LVCSR
Commoner
Order Specialist
Question Specialist
Explanation Specialist
The most similar : the best output
■Commoner
There is no limitation for recognition, but low accuracy
■Specialist
There are many out-of -domain utterances, but high accuracy
Integration

Figure 1: The OCSS model

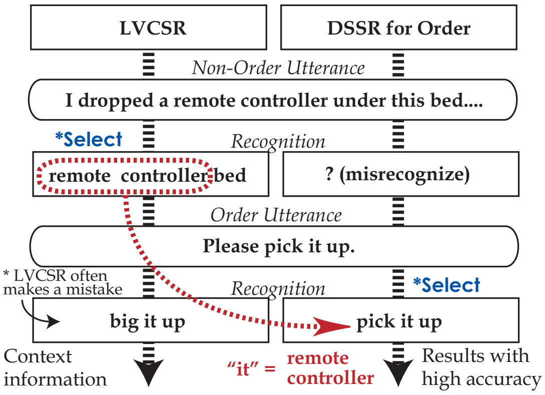


Figure 2: The effectiveness of our method.

cabulary and high accuracy for each order type. We use the LVCSR for recognition of utterances that the DSSR can not recognize, such as a chat between users. The information recognized by the LVCSR is of assistance for context construction of a dialogue. If we handle these different speech recognizers selectively and integratively, we realize a flexible and robust speech understanding method. Figure 2 shows the effectiveness of the proposed multiple recognizer. The DSSR achieves the order recognition with high accuracy and the LVCSR supplies lack of information in the order utterances.

In this paper we discuss two tasks for the OCSS model based method; (1) selection of the outputs and (2) anaphora resolution in the method. The 1st task is the selective usage of the multiple speech recognizer. In other words, it is to select outputs from each recognizer. For example, with respect to the utterance "Please pick it up" in Figure 2, it is important which result to select. The 2nd task is an integration process in our method. By using previous outputs from One Commoner (LVCSR) and Some Specialists (DSSRs), we resolve an anaphora in the current output. For example, with respect to the utterance "Please pick it up" in Figure 2, the system identifies that the word "it" in the utterance is the words "remote controller" which were recognized by the LVCSR in the previous utterance.

2 Output Selection in OCSS model

In this section, we explain the process of output selection in the OCSS model. In this process, we focus on a difference of outputs generated from each recognizer. Even human beings tend to misunderstand words which consist of similar pronunciations (Komatani et al., 2005). Here we focus on the output of the LVCSR. If an input is an order utterance, the DSSR and the LVCSR generate similar outputs on phoneme-level because the LVCSR is domain independence. On the other hand, if the input is not an order utterance, they often generate different outputs even on the phoneme-level because the DSSR never generates the correct result for non-order utterances.

Komatani et al. (2007) have reported an utterance verification method based on a difference of acoustic likelihood values computed from two recognizers. Kumar et al. (2005) have utilized Bhattacharyya distance to measure an acoustic similarity of different languages for multilingual speech recognition. In this paper, we use the edit distance as the similarity measure. The correspondence such as the edit distance is one of the most effective measures to identify high confidence words in outputs (Utsuro et al., 2004) and to extract similar word pairs (Komatani et al., 2005). In our method, if an input is an order utterance, the edit distance between the outputs from the DSSRs and the LVCSR becomes small. However if the input is not an order utterance, that between the outputs from the DSSRs and the LVCSR becomes large. In our method, we compute the edit distance of utterance-level and word-level by using a DP matching algorithm. In the process, we compute the edit distance between phonemes of words for both levels.

The rules to judge an utterance are applied in the following order:

1. Compute the edit distance of the utterance-level (ED_{utter}) between the LVCSR and each DSSR. For the outputs of which the edit distance is less than $thresh_{utter}$, we select the

output of the DSSR which contains the minimum ED_{utter}, as the final output.

2. Compute the edit distance of the word-level (ED_{word}) between the LVCSR and each DSSR. For the output of which the edit distance is less than thresh_{word}, we select the output of the DSSR which contains the minimum ED_{word} as the final output. Otherwise, the LVCSR as the final output.

The ED_{utter} is the edit distance value on the utterance-level. The ED_{word} is the average of the edit distance value computed on word-level. These values are normalized by the number of phonemes in the outputs. The thresh_{utter} and thresh_{word} are threshold values for the judgment. These values are decided experimentally.

In the computation of the word-level, we eliminate word pairs that are matched completely first. Next, we compute all the combinations of the other. Finally, we employ the minimum combinations as the word-level edit distance. Figure 3 shows an example of the calculation of the ED_{utter} and ED_{word}. In the figure, the dotted line denotes completely matched words. The numerals with arrows denote the original edit distance of the word pair. In the alignment process of word pairs, we select pairs which have the minimum value of the edit distance. In other words, we admit overlap of word pairs. For example," noue vs. no " and " no vs. no " in Figure 3.

3 Understanding and Anaphora Resolution

In this section we explain an anaphora resolution process in the OCSS model.

3.1 Understanding of Outputs from OCSS model

The output in the previous section, namely the output selection process, is an output of a speech recognizer. For the anaphora resolution process, we need to analyze the output.

For outputs from DSSRs, we convert them into a semantic frame. We utilize grammar information of DSSRs for the process. Each DSSR consists of 100-200 words and approximately 100 grammar patterns including approximately 50 categories. Figure 4 (a) shows an example of the grammar patterns and categories. The categories often contain semantic constraints such as "Drink_N" and "Location". For outputs from a LVCSR, we extract keywords by using some rules based on surface expression. Figure 4 (b) shows examples of the process. In the figure, "obj", "loc" and "agt" denote case markers.

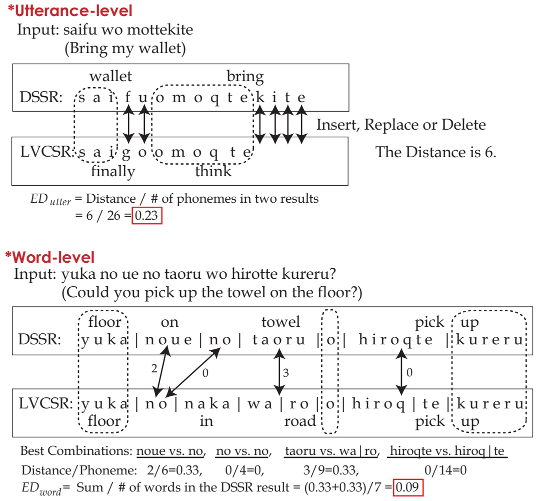
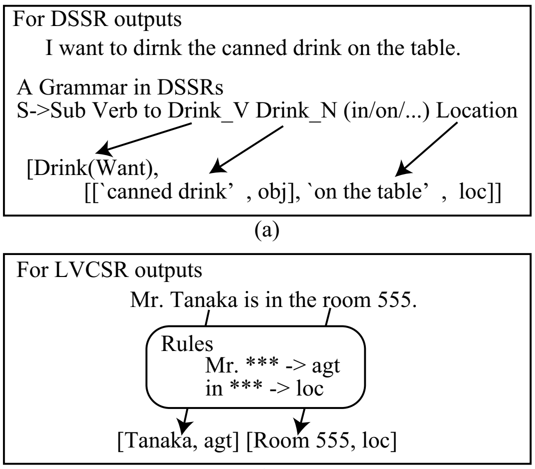

*Utterance-level
Input: saifu wo mottekite
(Bring my wallet)
wallet
bring
DSSR: s a i f u o m o q t e k i t e
Insert, Replace or Delete
LVCSR: s a i g o o m o q t e
The Distance is 6.
finally
think
ED utter = Distance / # of phonemes in two results
= 6 / 26 = 0.23
*Word-level
Input: yuka no ue no taoru wo hirotte kureru?
(Could you pick up the towel on the floor?)
floor
on
towel
pick
up
DSSR: y u k a | n o u e | n o | t a o r u | o | h i r o q t e | k u r e r u
2
0
3
0
LVCSR: y u k a | n o | n a k a | w a | r o | o | h i r o q | t e | k u r e r u
floor
in
road
pick
up
Best Combinations: noue vs. no, no vs. no, taoru vs. wa | ro, hiroqte vs. hiroq | te
Distance/Phoneme: 2/6=0.33, 0/4=0, 3/9=0.33, 0/14=0
ED word = Sum / # of words in the DSSR result = (0.33+0.33)/7 = 0.09

Figure 3: The edit distance calculation


For DSSR outputs
I want to dirnk the canned drink on the table.
A Grammar in DSSRs
S->Sub Verb to Drink_V Drink_N (in/on/...) Location
[Drink(Want),
[[`canned drink' , obj], `on the table' , loc]]
For LVCSR outputs
Mr. Tanaka is in the room 555.
Rules
Mr. *** -> agt
in *** -> loc
[Tanaka, agt] [Room 555, loc]

(a)

(b)

Figure 4: Examples of the output analysis

3.2 Anaphora Resolution

If an utterance contains an anaphoric expression, our system detects the antecedent from previous utterances. In this paper we handle demonstrative pronouns only. In this process, we focus on the following points: (1) the distance from the current utterance and (2) change of situation.

The anaphora resolution process is based on the following score calculation:

1. extract all words compatible with the case marker of anaphoric expressions[1]

[1] Correctly, words with the same marker and words with-

2. compute a base score:

$$base_i = \begin{cases} N - dist_i & \text{: Same situation} \\ \frac{N - dist_i}{2} & \text{: Different situation} \end{cases}$$

   where N is the number of previous utterances that our system treats in this anaphora resolution process. In this paper, $N = 10$. $dist_i$ is the distance between the current utterance that contains the anaphoric expression and the previous utterance that contains the antecedent. Here “change of situation” denotes “change of a speaker” or “change of the location of a robot”.

3. compute the following scores by using the base score:

   n-best: The word accuracy of the LVCSR is not often enough[2]. As a result, the antecedent often does not exist in the outputs from the previous utterances. To solve this problem, we use 10-best candidates of speech recognizers outputs.

$$nBest_i = \begin{cases} base_i \times CN_i & \text{: LVCSR outputs} \\ base_i & \text{: DSSRs outputs} \end{cases}$$

   where CN_i denotes the confidence measure computed from the LVCSR for each word.

   Marker: We tag a case marker to each word by using surface expression rules. However, all words are not always tagged by the rules because of lack of rules. Therefore we distinguish words with the same marker and words without a marker.

$$Marker_i = \begin{cases} \frac{base_i}{2} & \text{: Same marker} \\ 0 & \text{: No marker} \end{cases}$$

   Semantics: We tag some semantic labels to words in DSSRs. For example, the semantic label of “juice” and “tea” is “drink”. Here assume that the current utterance is “ I want to drink it”. In this situation, the antecedent of the word “it” has to contain the semantic label “drink”. Therefore we add a score to words that possess the same semantic label.

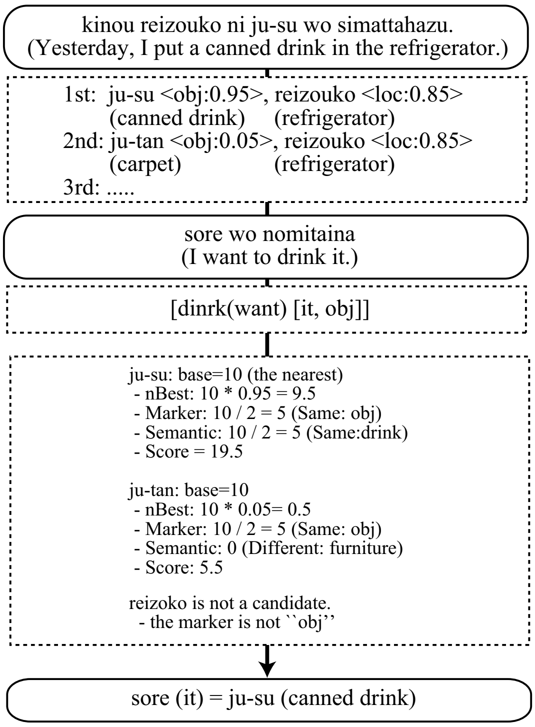

kinou reizouko ni ju-su wo simattahazu.
(Yesterday, I put a canned drink in the refrigerator.)
1st: ju-su <obj:0.95>, reizouko <loc:0.85>
(canned drink) (refrigerator)
2nd: ju-tan <obj:0.05>, reizouko <loc:0.85>
(carpet) (refrigerator)
3rd: .....
sore wo nomitaina
(I want to drink it.)
[dinrk(want) [it, obj]]
ju-su: base=10 (the nearest)
- nBest: 10 * 0.95 = 9.5
- Marker: 10 / 2 = 5 (Same: obj)
- Semantic: 10 / 2 = 5 (Same:drink)
- Score = 19.5
ju-tan: base=10
- nBest: 10 * 0.05= 0.5
- Marker: 10 / 2 = 5 (Same: obj)
- Semantic: 0 (Different: furniture)
- Score: 5.5
reizoko is not a candidate.
- the marker is not ``obj''
sore (it) = ju-su (canned drink)

Figure 5: An example of the anaphora resolution

$$Semantic_i = \begin{cases} \frac{base_i}{2} & \text{: Same Semantics} \\ 0 & \text{Otherwise} \end{cases}$$

4. compute the score of antecedent candidates in each utterance.

$$Score_i = nBest_i + Marker_i + Semantic_i$$

5. compute the final score of antecedent candidates

$$FScore_i = \sum_{i \in W} Score_i$$

   where W is all antecedent candidates.

6. extract the word that contains the maximum score as the antecedent.

Figure 5 shows an example of the process.

4 Experiments

In this section, we evaluate (1) the output selection based on the edit distance and (2) the anaphora resolution process described in the previous sections.

4.1 Speech recognizer in the experiment

We used Julius as the LVCSR and Julian as the DSSR (Lee et al., 2001). The Julian consists of a

out a marker are extracted. In other words, words with markers that are different from the anaphoric expression’s marker are not extracted for this anaphora resolution process.

[2] For the 1st output, it was less than 40% in a preliminary experiment.

vocabulary and a grammar file. For the grammar file we describe sentence structures in a BNF style, using word category names as terminate symbols. The vocabulary file defines words with its pronunciations (i.e., phoneme sequences) for each category. Here we design grammar and vocabulary files of the Julian which accepts only specific utterances from users. In this experiment, we used 4 DSSRs that we constructed by hand. The DSSRs are as follows:

- Order Utterances from patients (Order-P)
  e.g., Please bring the remote controller on the table.

- Order Utterances from nurses (Order-N)
  e.g., "Carry these meals to patient's rooms"

- System Commands
  e.g., "Move to the right by 50cm"

- Question Utterances
  e.g., "Where is my cellphone?"

4.2 Result of Output Selection

The 1st task in the experiment was to detect a speech recognizer which was suitable to generate the output in the multiple recognizer. The dataset consists of 20 utterances for each DSSR and 20 out-of-domain utterances such as greetings. The number of test subjects was 10. In other words, we evaluated our method with 1000 utterances: 5 categories (DSSRs and LVCSR) × 20 utterances × 10 test subjects. The thresh_{utter} and thresh_{word} were 0.26 and 0.08 respectively. These thresholds were determined on a preliminary experiment with another dataset.

Table 1 shows the experimental result. The F-value of the output selection was 0.916 on average[3]. Besides, we verified that the change of the F-value was small even if we changed the thresholds within the compass of 0.20-0.26[4]. Therefore, our method based on the edit distance is simple and robust.

4.3 Result of Anaphora Resolution

Next, we evaluated the anaphora resolution process in the OCSS model. The dataset of this experiment consisted of 206 utterances that included 50 anaphoric expressions. We compared the following combinations:

Table 1: Output Selection Accuracy.

Domain	Recall	Precision	F
Order-P	0.965	0.873	0.917
Order-N	0.965	1.000	0.982
Commands	0.930	1.000	0.964
Questions	0.975	0.878	0.924
LVCSR	0.765	0.827	0.795
Average	0.920	0.916	0.916

Table 2: Accuracy of the Anaphora Resolution.

Method	Accuracy
$\mathrm{Method}_{10}^{no\text{-}dist}$	32%
$\mathrm{Method}_{10}^{dist}$	41%
$\mathrm{Method}_{1st}^{dist+}$	48%
$\mathrm{Method}_{10}^{dist+}$	52%

- $\mathrm{Method}_{10}^{no\text{-}dist}$: In this method, $base_i$ was always 10.

- $\mathrm{Method}_{10}^{dist}$: In this method, $base_i$ was always computed from $N - dist_i$, i.e., it did not handle "Different situation" in Section 3.2 .

- $\mathrm{Method}_{1st}^{dist+}$: In this method, we used the 1st output only for the candidate extraction, i.e., it did not handle the *nBest* in Section 3.2 .

- $\mathrm{Method}_{10}^{dist+}$: The proposed method.

Table 2 shows the experimental result. The effectiveness of use of the distance from the current utterance was verified from the comparison between the $\mathrm{Method}_{10}^{no\text{-}dist}$ and the $\mathrm{Method}_{10}^{dist}$. Also the effectiveness of use of change of situation was shown from the comparison between the $\mathrm{Method}_{10}^{dist}$ and the $\mathrm{Method}_{10}^{dist+}$. The effectiveness of use of 10-best outputs was shown from the comparison between the $\mathrm{Method}_{1st}^{dist+}$ and the $\mathrm{Method}_{10}^{dist+}$ likewise. The $\mathrm{Method}_{10}^{dist+}$ obtained the best performance

However, the accuracy was insufficient (52%). Most of the mistakes in the anaphora resolution were due to misunderstanding of the speech recognizer (LVCSR). If the correct antecedent does not exist in the 10-best outputs from the LVCSR, our method can not detect it essentially. We compared results from actual outputs and transcripts, namely perfect outputs. Table 3 shows the difference between actual outputs and transcripts. This result

[3] In addition, the word recognition accuracy of each DSSR was 0.940 on average.

[4] The best F-value on this experiment was 0.924 in the case that thresh_{utter}=0.20.

Table 3: Accuracy of the Anaphora Resolution.

Input	Actual Output	Transcript
Accuracy	52%	88%

shows that it is important to improve the word accuracy of the LVCSR.

Our method was based on a simple scoring process. Iida et al. (2005) have proposed a machine learning-based approach to anaphora resolution. Applying other methods to our system is one of our future work.

5 Conclusions

In this paper, we described a speech understanding method based on a multiple speech recognizer. We called it “OCSS model”. The method was combination of one LVCSR and several DSSRs. By using this method, we realized a flexible and robust speech understanding method.

In this paper, we evaluated two processes of the method: (1) output selection and (2) anaphora resolution. For the output selection, the method was based on the edit distance between each output. In the experiment, we obtained high F-value (more than 0.9). This result shows that our method is simple and robust. For the anaphora resolution, the method was based on (1) the distance between an anaphora expression and an antecedent (2) change of situation such as speaker’s change. Although the proposed method was effective as compared with more simple methods, the accuracy was insufficient. The reason why the accuracy of the anaphora resolution was low was the accuracy of the LVCSR was low. To improve the accuracy, we need a LVCSR with more high accuracy.

Our future work includes (1) a large-scale experiment especially the anaphora resolution and (2) evaluation of the proposed method for other domains.

Acknowledgment

This research was supported by NEDO, Intelligent RT Software Project, 2009.

References

C. Bouwman, J. Sturm, and L. Boves. 1999. Incorporating confidence measures in the dutch train timetable information system developed in the arice project. In *Proceedings of ICASSP*.

M. Goto, K. Itou, and T. Kobayashi. 2005. Speech interface exploiting intentionally-controlled nonverbal speech information. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology (UIST 2005)*, pages 36–36.

R. Iida, K. Inui, and Y. Matsumoto. 2005. The issue of combining anaphoricity determination and antecedent identification in anaphora resolution. In *International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE)*, pages 244–249.

T. Isobe, K. Itou, and K. Takeda. 2007. A likelihood normalization method for the domain selection in the multi-decoder speech recognition system. *IEICE TRANSACTIONS on Information and Systems (Japanese Edition)*, 90(7):1773–1780.

K. Komatani, Y. Fukubayashi, T. Ogata, and H. G. Okuno. 2007. Introducing utterance verification in spoken dialogue system to improve dynamic help generation for novice users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 202–205.

K. Komatani, R. Hamabe, T. Ogata, and H. G. Okuno. 2005. Generating confirmation to distinguish phonologically confusing word pairs in spoken dialogue systems. In *Proceedings of 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 40–45.

K. Komatani and T. Kawahara. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proceedings of COLING 2000*, volume 1, pages 467–473.

S. C. Kumar, V. P. Mohandas, and H. Li. 2005. Multilingual speech recognition: A unified approach. In *Proceedings of InterSpeech 2005*, pages 3357–3360.

A. Lee, T. Kawahara, and K. Shikano. 2001. Julius - an open source real-time large vocabulary recognition engine. In *Proceedings of Eurospeech*, pages 1691–1694.

J. Ogata and M. Goto. 2005. Speech repair: Quick error correction just by using selection operation for speech input interfaces. In *Proceedings of Interspeech 2005*, pages 133–136.

T. Utsuro, H. Nishizaki, Y. Kodama, and S. Nakagawa. 2004. Estimating highly confident portions based on agreement among outputs of multiple lvcsr models. *Systems and Computers in Japan*, 35(7):33–40.