Difference of acoustic modeling for read speech and dialogue speech

Masato Mimura and Tatsuya Kawahara

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto, 606-8501 Japan

(Received 9 April 2001, Accepted for publication 7 June 2001)

Keywords: Speech recognition, Dialogue speech, Acoustic model, Triphone model

PACS number: 43.72.Ne

1. Introduction

As the dictation systems have been developed, the research target of speech recognition shifts from read speech to more spontaneous speech. Among them, lecture speech in public space and dialogue speech in some service domains are considered as proper tasks for the next step, rather than fully conversational speech. For the research of dialogue speech, domain-limited but large-scale spontaneous speech database was set up at ATR Interpreting Telecommunications Research Laboratories. In this report, we address the difference of read speech and dialogue speech by using this database, and propose an approach to dialogue speech recognition.

2. Databases and tasks

All experiments described in this report are male-dependent.

Table 1 lists the specification of databases for acoustic modeling. We use the database of phonetically-balanced sentences (ASJ/PB) and newspaper article sentences (ASJ/JNAS) edited by the Acoustic Society of Japan as read speech material. Acoustic models trained with these databases are provided in the IPA Japanese Dictation Toolkit [1]. We use the “speech dialogue database for spontaneous speech recognition —travel arrangement task—” of ATR [2] as dialogue speech material. The size of read speech data is about three times as large as dialogue data. We also attempt to train acoustic models using both databases.

Table 2 gives the specification of the tasks for evaluation. The samples of read speech are selected from ASJ/JNAS database, and the test set is used in the IPA project. We use the 20K-vocabulary model of the IPA Japanese Dictation Toolkit as the language model for recognizing this test set. The samples of dialogue speech are selected from normal dialogue database (ATR/SDB) and cross-lingual dialogue database (ATR/SLDB) of ATR. Cross-lingual dialogue is made via a human interpreter, thus it has relatively similar characteristics to read speech. For the dialogue speech recognition, the language model is trained with both ATR/SDB and ATR/SLDB database excluding the test set. The training data size of the language model for the dialogue task is smaller, but its perplexity is much smaller than the dictation task.

We use our Julius 2.2 as a decoder.

3. Difference in phonetic context

The distribution of phone contexts in respective corpus is shown in Fig. 1. In this figure the x-axis is the order of triphones by appearance counts in each corpus, and the y-axis means the accumulated ratio to the total counts of triphones. This figure shows that dialogue speech has larger bias in phonetic contexts than read speech. In fact, the contexts characteristic to colloquial expressions in spoken language, such as n-a+s a-s+u d-e+s e-s+u, amount to relatively large ratio in the dialogue corpus (By x-y+z we mean a triphone, where x and z are the left and the right phonetic context of the phone y, respectively).

4. Mis-match of model and task

All acoustic models are tied-state triphone HMM with 2000 states and 16 mixture components. Table 3 lists the word accuracy for combination of training databases and evaluation tasks. The accuracy of JNAS task by the IPA model trained on the ASJ database is same as the best result in Ref. [2]. The accuracy of ATR task by the acoustic model trained with the ATR database is almost equal to the result of Ref. [3], therefore these figures are reliable.

Followings are derived from the results.

1. It is important to match styles of training database and test set in read speech (JNAS) and normal dialogue (ATR/SDB).
2. Increasing training data by using speech database of different style is not effective in the case of these styles.
3. Since cross-lingual dialogue speech is similar to read speech, some effect is observed in increasing training data by incorporating read speech corpus.

5. Acoustic modeling of dialogue speech by incorporating read speech corpus

The reason the accuracy for dialogue speech is lower than read speech is considered that (1) triphone HMM of current method can not model characteristics of dialogue, or (2) training data is not sufficient. We deal with the latter problem in this report. However, it is very difficult to collect as much dialogue data as read speech. Simply adding read speech data
is not effective for dialogue speech. We propose methods that incorporate the read speech corpus by keeping the syllable structure of dialogue. Specifically, we present and compare two methods:

(a) Derive a decision tree clustering for triphone set from dialogue speech and apply it to combined data of dialogue and read speech corpora.

(b) Pick up most frequent triphone contexts in dialogue and separately train them with dialogue speech. The rest triphones are trained with combined data of two corpora.

We calculated correlation of triphone occurrences with each database (read and dialogue) and correlation in different tasks in Simulated Spoken Dialogue Corpus (Grant-in-Aid for Scientific Research on Priority Areas), and picked up the 100 best triphone contexts that are strongly correlated with dialogue and are independent of tasks. These triphone contexts are characteristic to colloquial expressions as described in section 3.

The results of dialogue speech recognition for two tasks are shown in Table 4.

For cross-lingual dialogue (ATR/SLDB), the accuracy is improved by both methods. For normal dialogue (ATR/SDB), the accuracy does not reach the result by the model trained with only the same corpus, but it is higher than the result by the model trained by the simple combination of both corpora.

6. Conclusion

We have investigated difference of dialogue speech from read speech in acoustic modeling, focusing on triphone characteristic, and also presented methods to supplement training of dialogue models with read speech corpus. The results described in this report may be affected by not only the style but also the task and input environment. We plan to evaluate on dialogue data of different tasks.

References

