音声言語情報処理 19−11
(1997. 12. 12)

会話スタイル依存の言語モデルを用いたキーフレーズの検出・検証

河原 達也　石塚 健太郎　堂下 修司
京都大学 工学研究科 情報工学教室
〒606-01 京都市 左京区 吉田本町

Chin-Hui Lee
Bell Laboratories
Murray Hill, NJ, USA

あらまし

頑健なキーフレーズの検出・検証のために、タスクに独立なフィラーモデルの構成法について提案する。本モデルは、タスク固有の語彙を前提とするかわりに、会話のスタイルに依存したフレーズを抽出するものであり、類似の異なった(大規模)コーパスから学習することが可能である。このポータブルで汎用的なモデルを2種類実装した。第1に、情報検索対話スタイル依存モデルを ATIS コーパスを用いて学習し、異なった情報検索タスクの音声理解におけるフィラーモデルとして利用することで、認識精度が向上した。第2に、講演スタイル依存モデルを口頭発表の書き起こしテキストから学習し、(別の)講義中に使用される音声操作プロジェクタのコマンドキーフレーズの検証に利用することで、音節連接モデルに比べて高い検証精度を得た。

キーワード　音声認識, 単語スポッティング, 発話検証, キーフレーズ検出, 音声理解

Speaking-Style Dependent Lexicalized Filler Model for Key-Phrase Detection and Verification

Tatsuya Kawahara　Kentaro Ishizuka　Shuji Doshita
Dept. of Information Science
Kyoto University, Kyoto 606-01, Japan

Chin-Hui Lee
Bell Laboratories
Murray Hill, NJ, USA

Abstract

A task-independent filler modeling for robust key-phrase detection and verification is proposed. Instead of assuming task-specific lexical knowledge, our model is designed to characterize phrases depending on the speaking-style, thus can be trained with large corpora of different but similar tasks. We present two implementations of the portable and general model. The dialogue-style dependent model trained with the ATIS corpus is used as a filler and shown to be effective in detection-based speech understanding on different dialogue applications. The lecture-style dependent filler model trained with transcriptions of various oral presentations also improves the verification of key-phrases uttered during lectures.

key words　speech recognition, word spotting, utterance verification
key-phrase detection, speech understanding

1 Introduction

In order to make an automatic speech recognition system deployable in real-world applications, it needs to have not only high accuracy but also flexibility to handle spontaneous utterances and reject irrelevant speech portions. We have introduced a combined detection and verification framework[1] that focuses on identifying the semantically significant portions and rejects the out-of-task parts of input utterances. Concept-based key-phrases are used as a detection unit, which enables more stable matching than simple word spotting.

Utterance verification technique is incorporated to obtain reliable detection and reduce false alarms. We adopt a vocabulary-independent approach for verification of detected phrases[2] so as to be applicable to various tasks. The verifier is subword-based. Specifically, we set up an *anti-subword model* for every subword to model the confusing patterns, and compute a likelihood ratio of the two models to represent confidence of the subword-level recognition. A confidence measure for phrase verification combines the subword-level verification scores. Thus, it is purely based on acoustic information.

On the other hand, it is well-known that lexical and language models are also effective for improving keyword detection and suppressing false alarms[3][4][5]. Most of the conventional works use task-dependent lexical entries and language models that are trained with a large corpus of the same task. However, it is not a realistic assumption that sufficient data is available for every single task in all applications.

In this paper, we present a task-independent approach for lexical filler modeling to enhance the key-phrase detection and verification. Instead of task-specific models, we propose a model depending on the speaking-style such as dialogue-style or lecture-style, which can be trained with different corpora of the same style.

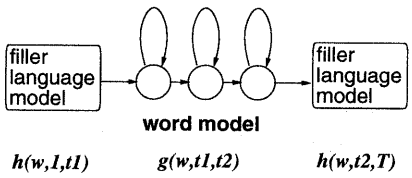

Figure 1: Language model for keyword spotting

2 Lexical and Language Model for Detection and Verification

2.1 Language Model for Keyword Spotting

The filler model is intended to cover typical patterns that accompany keywords or key-phrases. The whole model for keyword spotting is illustrated in Figure 1.

It is effective in suppressing false alarms and controlling the detection threshold based on acoustic evaluation of the whole utterance. Thus, it must achieve a wider coverage of possible filler patterns with a smaller complexity. We compared several filler language models for keyword spotting[6] and found that (1) a parallel network of phone models (phone network) is robust but insufficient, (2) lexical knowledge is very effective, (3) when a sufficient size of lexicon is incorporated, the phone network model is no longer needed to explicitly cover unknown words[4].

In the actual applications, however, sufficient data is not always available to obtain reliable language models as the data collection and labeling cost too much. Even a vocabulary set is not reliably given in many cases, as the filler has much more variety of patterns than the keyword vocabulary.

2.2 Lexical Model for Key-Phrase Verification

The same sort of lexical or language models can be used for utterance verification, namely to reject unknown words or out-of-task utterances. The output of the recognizer is tested, and accepted if its score is better than any entries of the verification model. Namely, the model is competitive to rec-

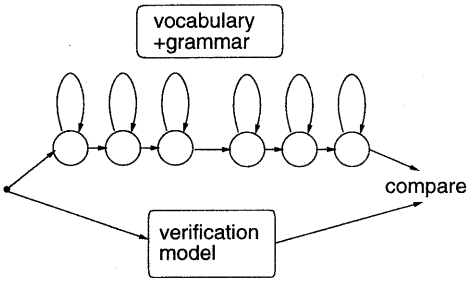

Figure 2: Language model for key-phrase verification

ognized candidates. The whole model for keyword spotting is illustrated in Figure 2.

Thus, the whole verification model has to cover possible confusing patterns that appear as non-key-phrases.

In many previous works, a general acoustic sink model[7] or a phone network model[8] is used to serve the purpose. However, such simple models are usually not sufficient to characterize non-key-phrase events especially when the task vocabulary size gets large.

3 Speaking-Style Dependent Model

The model we propose plays two roles described above: filler model for detection and competitive model for verification. They are unified in this study.

Our model is represented as a variable-length phrase model[9]. It is not a precise language model since stochastic information is not attached to the word connections. However, it still models word sequences, which can be implemented as a tree-structured lexicon or a simple automaton. It characterizes input utterances better than the simple phone network model. Moreover, we do not have to adjust a penalty that is usually imposed on them to fairly compare with the key-phrase models.

The key property of the model is that it is constructed in a task-independent manners. Instead of the task-dependent lexicon and corpus, we assume the model is dependent on the speaking-style. People use similar phrases in making an information

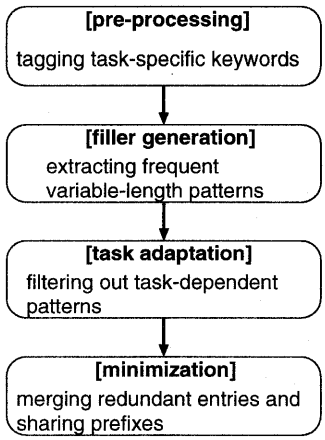

Figure 3: Training procedure for task-independent filler model

query dialogue whatever the content of the query is. And they use a different style in giving an oral presentation in public. Based on the assumption, we train the filler phrase model with large corpora that are not task-specific as long as their tasks are similar and so are the speaking-styles. It is trained by picking up characteristic word (or syllable) sequences with excluding task-specific words. The outline of the procedure is described as follows:

1. Task-specific keywords are tagged using domain knowledge and a lexical analyzer, so that they are not included in speaking-style dependent fillers.

2. Fillers are constructed by concatenating frequent word (or syllable) sequences

3. Task-specific patterns are filtered out by a simple task adaptation technique.

4. Minimization of the filler model is performed by merging redundant patterns and sharing prefixes.

It is also depicted in Figure 3.

The model is *lexicalized* in a sense that it uses the same subword model as the key-phrase recognizer. The property is also essential for task-independent portability[5].

We present two implementations and applications of this model in the following sections. The first one deals with dialogue on information query, and the other models lecture-style expressions. The dialogue-style filler and the lecture-style filler are quite different, thus specific modeling for each speaking-style is performed and evaluated in a task of the same style. Both of them are also completely different from the expressions used in written language such as a newspaper corpus.

4 Dialogue-Style Dependent Filler Model

4.1 Filler Phrase Model

At first, we construct a lexicalized filler model that is dependent on the dialogue-style. Specifically, we deal with information query which is a typical application of spoken dialogue systems. So the purpose of the filler model is to improve the detection rate of key-phrases which will lead to robust speech understanding. The lexicalized filler model here is defined as a set of word sequences or phrases. Instead of task-dependent key-phrases, we extract patterns related to the dialogue style in information query.

We made use of the ATIS-I corpus of 13099 utterances, which is one of the largest spoken dialogue corpus available.

The automatic phrase extraction algorithm[9] is applied. As a pre-processing, keywords are tagged based on the task specification, for example, CITY for Atlanta, Boston, Cleveland and so on. Initially, frequent non-keywords are picked up as the cores of fillers. Then, adjacent sticky words are concatenated to grow a filler phrase, until it encounters a keyword or it cannot satisfy a threshold on its occurrence count (coverage) normalized by the corpus text size. We picked up the 105 most frequent filler phrases, whose coverage exceeded the threshold (=0.001). A simple task adaptation technique is applied to remove the patterns that are dependent on the ATIS corpus and never appear on the other corpus (that is not of the task for evaluation) before porting to the specific task domain.

It is confirmed that we can obtain a reasonable set of phrases. Examples of the resultant phrases are as follows:

```
are there any, is there a, i want to,
show me all, i would like to
```

4.2 Evaluation on Speech Understanding

The generated filler model is applied to speech understanding based on our key-phrase detection and verification approach[1]. Experimental evaluation is performed on several sub-tasks of real-world spoken dialogue systems other than ATIS. All data are uttered by the general public and collected via dial-up lines at (former) AT&T Bell Labs. Task-independent context-dependent phone HMM is used as the acoustic model.

Here we show the results on spontaneous expressions specifying locations (LOCATION subtask) in the car reservation system. The vocabulary set covers 371 major locations in USA, plus hundreds of accompanying words extracted by our automatic procedure. The sample utterances are classified into three categories. *In-grammar* samples consist of valid key-phrases only. *Out-of-grammar* samples contain extraneous words, hesitations and repetitions in addition to expected key-phrases. *Out-of-task* samples contain no valid key-phrases and a null slot should be produced as rejection. The semantic accuracy is defined to evaluate how many semantic slots are correctly recognized.

The semantic accuracy with various recognition and verification methods is listed in Table 1. For comparison, decoding with a manual sentence grammar is also included. It achieves a good performance on grammatical samples, but fails to cope with ill-formed utterances. The key-phrase detection approach drastically improves the accuracy for out-of-grammar samples at the expense of small degradation for in-grammar samples. In the baseline detection method, an acoustic sink model as the simplest filler is already incorporated.

Then, verification process is incorporated. The subword-based acoustic verification method uses anti-subword model to eliminate false alarms caused by improper matching. The filler phrase model verification method uses the task-independent

Table 1: Semantic accuracy in speech understanding (LOCATION sub-task)

	in-grammar samples	out-of-grammar samples	out-of-task samples	total
number of samples	681	99	131	911
decoding with sentence grammar	94.2%	16.1%	26.0%	79.0%
key-phrase detection	92.6%	40.1%	20.6%	79.7%
+ subword-based acoustic verification	91.2%	59.1%	35.1%	82.1%
+ filler phrase model verification	92.7%	58.4%	21.4%	81.8%
+ combination of both verification	91.2%	67.8%	35.9%	83.1%

model trained with the ATIS-I corpus to suppress false alarms. In fact, it is used to generate competitive hypotheses in the detection process. Both verification methods give comparable understanding rates, which are better than conventional methods that do not apply verification. Moreover, the combination of both strategies further improves the accuracy for out-of-grammar samples and achieves the best performance. While the acoustic verification models confusing subwords and rejects improper matching, the task-independent phrase model filters out out-of-task portions.

It should be noticed that a reliable statistical language model can hardly be trained with this typical size of field trial data. The results show that our model trained with the other large corpus enhances the detection performance.

5 Lecture-Style Dependent Filler Model

5.1 Lexicalized Filler Model

Next, we apply our modeling on lecture-style speech. The task here is to detect several key-phrases during a lecture presentation.

The system in this section is developed for Japanese language. As Japanese are written without spacing between words, the definition and boundary of words are ambiguous and dependent on lexical analyzers. Thus, a lexicalized filler model is defined as a set of sequences of characters corresponding to syllables. They make pseudo phrases as a result.

We made use of a corpus that transcribes oral presentations at the meeting of SIG-SLP (Spoken Language Processing) held in Tokyo, May 1995. It has 18109 syllable characters.

Domain-specific keywords are removed by filtering out normal nouns labeled by a morphological analyzer (JUMAN) as a pre-processing. Then, frequent character sequences are picked up by a similar procedure as in the previous section. At a coverage threshold of 0.0005, we obtained 230 sequences (pseudo phrases) with a length of 3 to 6. They are totally different from the dialogue-style dependent phrases.

5.2 Evaluation on Utterance Verification

The extracted filler model is applied to key-phrase verification for the slide projector operated with speech input. Key-phrases are commands for the projector operation, such as "next slide please" or "two slides back". They are represented as a finite state grammar. The vocabulary size for the commands is 56.

A lecturer uses the same microphone to give a presentation and to utter commands to the projector. Thus, most of input speech segments are not command key-phrases and contain vocabulary of over thousands. Since the projector should not be triggered by false alarms, we impose following constraints: (1) a *magic word* ('operator' in this experiment) has to be uttered right before command key-phrases. (2) a pause must be put before and after the magic word and commands.

Table 2: Key-phrase verification performance

	FR	FA
no verification	8.0%	33.1%
syllable network model verification	8.0%	13.5%
filler phrase model verification	0.0%	0.0%

FR: false rejection of key-phrases
FA: false acceptance of lecture segments

A speech segment aligned with pauses is input to the recognizer that is made of task-independent subword HMM and the finite state grammar. Then, the recognizer's output is compared with the lexicalized filler model for verification. If the score of the key-phrase is better than that of the optimal sequence of filler phrases, then it is accepted as a command.

The test samples are 50 command key-phrase utterances and 133 speech segments of a lecture whose duration lengths are comparable to those of key-phrases (less than 5 sec.). The topic and speaker of the lecture is different from those of the training corpus. For comparison of the verification model, we tested a simple syllable network model that represents a parallel network of syllables. We also performed an experiment to make decision based on the absolute value of the recognition score without using any verification models.

The verification results are listed in Table 2. The figures given are at the best operating point for each method. There was no need for threshold adjustment on our proposed method because direct comparison of the recognizer's output and the verification model's realized a perfect verification. The syllable network model eliminates false alarms, but not to zero.

Thus, the verifier made of the lexicalized filler model trained with a different corpus makes the speech-input projector practical.

6 Conclusions

We have proposed the lexicalized filler model depending on the speaking style. It models task-independent phrases uttered in the same speaking style, thus can be trained with large corpora of different tasks. The key property of the model is portability and generality. It is a lexicalized model and can be ported to tasks of the same style without re-training. The model is realized in two different styles: dialogue-style and lecture-style. They are successfully applied to speech understanding and utterance verification, respectively. It is also shown that the proposed detection and verification framework with the filler model effectively works for various tasks and even different languages.

References

[1] T.Kawahara, C.H.Lee, and B.H.Juang. Combining key-phrase detection and subword-based verification for flexible speech understanding. In *Proc. IEEE-ICASSP*, pages 1159–1162, 1997.

[2] R.A.Sukkar and C.-H.Lee. Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition. *IEEE Trans. Speech & Audio Process.*, 4(6):420–429, 1996.

[3] J.R.Rohlicek, P.Jeanrenaud, K.Ng, H.Gish, B.Musicus, and M.Siu. Phonetic training and language modeling for word spotting. In *Proc. IEEE-ICASSP*, volume 2, pages 459–462, 1993.

[4] M.Weintraub. Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system. In *Proc. IEEE-ICASSP*, volume 2, pages 463–466, 1993.

[5] R.E.Meliani and D.O'Shaughnessy. Accurate keyword spotting using strictly lexical fillers. In *Proc. IEEE-ICASSP*, pages 907–910, 1997.

[6] T.Kawahara, T.Munetsugu, N.Kitaoka, and S.Doshita. Keyword and phrase spotting with heuristic language model. In *Proc. ICSLP*, volume 2, pages 815–818, 1994.

[7] J.G.Wilpon, L.R.Rabiner, C.H.Lee, and E.R.Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. Acoust., Speech & Signal Process.*, 38(11):1870–1878, 1990.

[8] A.Asadi, R.Schwartz, and J.Makhoul. Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. In *Proc. IEEE-ICASSP*, pages 305–308, 1991.

[9] T.Kawahara, S.Doshita, and C.H.Lee. Phrase language models for detection and verification-based speech understanding. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.