様式：課程博士用

平成 25 年 9 月 26 日

電子・情報工学専攻		学籍番号	099306
申請者 氏名	太田 健吾		

指導 教員	中川 聖一

論文要旨（博士）

論文題目	話し言葉特有の現象の統計的モデリングに関する研究

（要旨　1,200 字程度）

近年，音声言語処理の対象は書き言葉から話し言葉へと移行しつつある．例えば，我が国では，国会答弁の自動速記や，テレビ番組の自動生成のために音声認識が導入されている．また，多くの携帯端末に音声対話による質問応答システムが搭載され，より親しみ易い話し言葉によるインタラクションの実現が望まれている．

しかし，話し言葉を対象とした音声言語処理では，書き言葉を対象とした場合とは異なり，特有の様々な問題が発生する．そこで本研究では，話し言葉特有の現象をモデル化することにより，話し言葉を対象とした音声言語処理の高精度化を行った．

話し言葉特有の現象のうち，最も出現頻度が高いのはフィラー（“えっと”，“あの一”等の場繋ぎ語）である．フィラーに対応した音声認識用言語モデルを構築するためには，フィラーを含むコーパス（テキストデータベース）から言語モデルを学習する必要がある．しかし，そのようなコーパスは現実的には収集が難しい．そこで，本研究では，フィラーを含まないコーパスからフィラーに対応した音声認識用言語モデルを構築する手法を提案した．これにより，話し言葉を対象とした音声認識の認識精度を改善することが出来た．

また，音声対話システムにおいても，フィラーは重要な役割を果たす．特に，音声対話システムの応答文の文中におけるフィラーは，ユーザの感じる応答文の聞き易さや自然さに影響を与えると考えられる．そこで，文脈を考慮して，応答文の適切な位置にフィラーを挿入することで，ユーザの感じる聞き易さや自然さ，および応答文の文内容に関する理解度が改善することを示した．

さらに，話し言葉調の音声では，書き言葉中の読み上げ音声とは異なり，言語的な区切りとは異なる位置にポーズ（無音区間）が発生する．従って，従来の書き言葉を対象とした音声認識のように，コーパス中の句読点をポーズに対応させて言語モデルを学習する手法は不適切である．同様に，ポーズで区切られた認識処理単位をそれぞれ独立に処理する従来の認識手法も不適切である．これらの問題を考慮し，ポーズ情報を含まないコーパスから，話し言葉のポーズに対応した言語モデルを構築する手法を提案した．また，直前の認識処理単位の文脈情報を認識に利用する手法を提案した．これらの提案手法により，話し言葉を対象とした音声認識の認識精度を改善することが出来た．

最後に，話し言葉を対象とした音声認識において，正確な書き起こしを収集することは困難だが，速記者等によって整形が加えられた書き起こしを収集することは比較的容易であるという点に注目し，整形された書き起こしから整形箇所を自動検出する手法を

提案した．提案手法により，話し言葉と書き言葉のパラレルコーパスを効率的に収集できることを示した．また，提案手法によって抽出された非整形部分は，音響モデルの適応のための発音ラベルとして有効であることを示した．

year	month	day
2013	9	26

Department	Electronic and Information Engineering	ID	099306
Name	Kengo Ohta		

Supervisor	Seiichi Nakagawa

Abstract

Title	A study on statistical modeling of characteristic phenomena in spontaneous speech

(800 words)

Recently, the main target domain of spoken language processing applications is transitioning from read speech to spontaneous speech. For example, a speech recognition system has been introduced into an automatic creation of meeting records in the National Diet and an automatic closed captioning for TV programs. Additionally, a spoken dialogue-based question answering system has been incorporated into the most of cell phone. As a result, the demand for a familiar man-machine interaction based on spontaneous speech is increasing.

However, contrary to the spoken language processing against read speech, the spoken language processing against spontaneous speech faces the various issues. In consideration of such issues, in this study, we improved the performance of spoken language processing against spontaneous speech by modeling the characteristic phenomena in spontaneous speech. At first, we conducted a corpus study of read speech and spontaneous speech. Based on this study, we addressed each sub issues in the spoken language processing against spontaneous speech.

The most frequent issue among the characteristic phenomena in spontaneous speech is filled pause. The simplest approach to constructing a language model for speech recognition that covers filled pause is to train it from large scale corpus (text database) consisting of many faithful transcripts of filled pauses. However, the available corpora are usually limited because they are quite expensive to prepare. In this study, we proposed a new approach to constructing a language model that covers filled pause using a corpus that does not include filled pause. In our method, a filler prediction model for building a language model that includes fillers from a corpus without fillers. A filler prediction model is trained from a corpus that does not cover domain-relevant topics. It recovers fillers in inexact transcribed corpora in the target domain, and then a language model that includes fillers is built from the corpora. Our proposed approach improved the recognition accuracy of speech recognition against spontaneous speech.

Filled pause has an important role not only in speech recognition but also speech dialogue system. Particularly, filled pause in a response speech of spoken dialogue system affects the listenability and the naturalness of the response speech which users feel. Considering this issue, in subjective experiments of a tourist-guiding task, we compared user comprehension, naturalness, and listenability of the system's responses with and without filled pauses and silences. Experimental results showed that the listenability, the naturalness, and the comprehension of the response speech improve by inserting filled pauses into the adequate positions.

Furthermore, there are mismatches between speech processing units used by a speech recognizer and sentences of corpora in a spontaneous speech recognition task.

A standard speech recognizer divides an input speech into speech processing units based on its power information. On the other hand, training corpora of language models are divided into sentences based on punctuations. There is inevitable mismatch between speech processing units and sentences, and both of them are not optimal for a spontaneous speech recognition task. We proposed two sub methods to address this problem. At first, the words of the preceding units are utilized to predict the words of the succeeding units, in order to address the mismatch between speech processing units and optimal units. Secondly, we proposed a method to build a language model including short pause from a corpus with no short pause to address the mismatch between speech processing units and sentences. Their combination achieved better performance than the conventional method in the meeting speech recognition task.

Finally, we addressed the problem of availability of spontaneous speech corpora. Large-scale spontaneous speech corpora are crucial resource for various domains of spoken language processing. However, the available corpora are usually limited because their construction cost is quite expensive especially in transcribing speech precisely. On the other hand, loosely transcribed corpora like shorthand notes, meeting records and closed captions are more widely available than precisely transcribed ones, because their imperfectness reduces their construction cost. Because these corpora contain both precisely transcribed regions and edited regions, it is difficult to use them directly as speech corpora for learning acoustic models. Under this background, we have been considering to build an efficient semi-automatic framework to convert loose transcriptions to precise ones. In this study, we describe an automatic detection method of precise regions from loosely transcribed corpora for the above framework. Our detection method consists of two steps: the first step is a force alignment between loose transcriptions and their utterances to discover the corresponding utterance for the certain loose transcription, and the second step is a detector of precise regions with a support vector machine using several features obtained from the first step. Our experimental result shows that our method achieves a high accuracy of detecting precise regions, and shows that the precise regions extracted by our method are effective as training labels of lightly supervised speaker adaptation.