論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉コーパスを用いた音声認識の検討
Title(English)	Automatic Speech Recognition Using a Spontaneous Spech Corpus
著者(和文)	篠崎隆宏, 細川貴生, 古井貞熙
Authors(English)	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会2001年春季講演論文集, Vol. , No. 1-3-14, pp. 31-32
Citation(English)	, Vol. , No. 1-3-14, pp. 31-32
発行日 / Pub. date	2001, 3

スペシャル・セッション〔自由発話の音声認識に向けて〕

1—3—14 話し言葉コーパスを用いた音声認識の検討*

◎篠崎 隆宏, 細川 貴生, 古井 貞熙 (東京工業大学)

1. はじめに

自由発話された音声を十分な精度で認識することは音声認識の用途を広げていく上で重要である。従来の書き言葉を対象としたコーパスに基づくモデルでは、話し言葉に対応できず低い認識率しか得られていない。このような背景から、話し言葉の構造を明らかにし、話し言葉の音声認識理解技術を高めることを目標として、新しいプロジェクトが開始された[1]。本論文では、プロジェクトに関連して東工大で進めている、話し言葉の音声認識に関する研究状況を報告する。

2. 話し言葉コーパス

プロジェクトでは種々の学会の講演音声や模擬的な講演を収録し、書き起こしを行っている。本論文の内容は2000年12月の時点で利用可能なデータに基づいている。書き起こしデータには未チェックのものとチェック済みのものがあるが、両方混ぜて使用した。言語モデル・音響モデル作成時には次章で示すテストセット講演と同一話者による講演は全て除外した。テストセットを除いた全講演数は610であり、内訳は多い順に模擬講演が336、音響学会が139、言語処理学会が63講演、その他72講演となっている。

3. 認識タスクと実験条件

3.1 テストセット

プロジェクトで録音した10講演を、認識対象(テストセット)として用いた。全て男性話者である。その概要を表1に示す。正式データ名は表第1列の通りであるが、以下では簡単のため第2列に示す表記を用いる。

表1 テストセットの概要

データ名称	略称	学会/研究会	講演時間
AS99SEP022	A22	日本音響学会	28分
AS99SEP023	A23	日本音響学会	30分
AS99SEP097	A97	日本音響学会	12分
PS99SEP025	P25	音声学会	27分
JL99OCT001	J01	国語学会	57分
KK99DEC005	K05	国語研究所	42分
NL00MAR007	N07	言語処理学会	15分
SG00MAR005	S05	社会言語科学会	23分
YG99JUN001	Y01	融合研究会	14分
YG99MAY005	Y05	融合研究会	15分

3.2 テストセットの特徴

テストセット中の各講演の発話速度を図1に、フィラー数、言い直し数を図2に示す。A22は発話速度が速くフィラーや言い直しが多い。

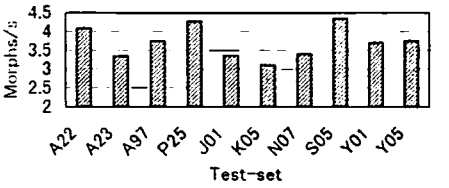

図1 発話速度

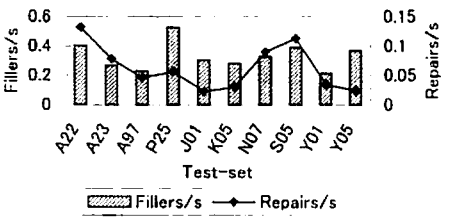

図2 発話中のフィラーと言い直し回数

3.3 実験条件

音声は16kHzで標本化、16bitで量子化した。音響パラメータはMFCC12次元、Δケプストラム12次元、対数エネルギーの1次差分の25次元で、切り出した発話区間ごとに平均ケプストラムによる正規化(CMS)を行った。形態素を単位とする統計的言語モデルを用い、正解文や言語モデルの作成にはNTTで開発された形態素解析ツールJTAGを使用した。デコーダにはJulius3.1を使用した。入力音声は書き起こしに含まれるラベルに基づき、500ミリ秒の無音を基準として切り出した。

4. 言語モデル

4.1 言語モデルの作成

統計的言語モデル(2-gram, 逆向き3-gram)を3組用意した。

Spon : プロジェクトで収録した講演の書き起こし610講演から作成した。読点に関してはテキストの段階で補った。

Web : World Wide Web上で公開されている講演書き起こしテキストを収集しコーパスを作成した

* Automatic speech recognition using a spontaneous speech corpus. By Takahiro Shinozaki, Takao Hosokawa and Sadaoki Furui (Tokyo Institute of Technology)

[1]。間投詞に関してはテキストの段階で補った。WebSp：Webのコーパスにテキストの段階で、教科書「音声情報処理」（総形態素数：63k）を加えてモデルを作成した。

各言語モデルの概要を表2に示す。

表2　言語モデルの概要

言語モデル	学習形態素総数	語彙数
Spon	1.5 M	20 k
Web	2 M	20 k
WebSp	2+0.06 M	20 k

4.2 テストセットパープレキシティ

表1中の始めの4講演について、図3にテストセットパープレキシティと未知語率を示す。これらの講演は音声に関連したものである。

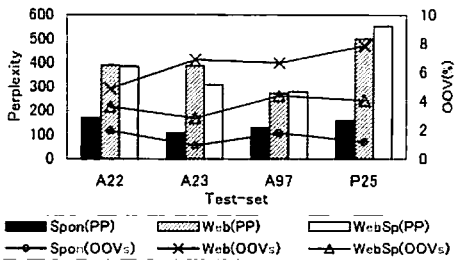

図3　パープレキティと未知語率

話し言葉から作成したSponでは他のモデルに比べ低いパープレキシティが得られた。Webは書き起こす際に文章として編集されていること、話の内容が音声に関連したものではないことなどからパープレキシティ、未知語率とも高くなっている。WebSpでは音声の教科書を加える話題適応により未知語率が下がっている。

5. 音響モデル

使用した音響モデルの概要を表3に示す。全て男性用モデルである。IPA2kはIPA99によるモデルであり、それ以外はプロジェクトの学習セットのうち男性部分から作成したモデルである。PTM2kは2k状態のトライフォンを元にしたPTM(phonetic tied-mixture)、その他は状態共有モデルである。TSf2kは発話速度が学習コーパス中上位6割の講演を選んで学習したモデルである。

表3　音響モデルの概要

モデル名	状態数	混合数	学習データ
PTM2k	129(2k)	64	講演59時間
TS1.5k	1.5k	16	講演59時間
TS2k	2k	16	講演59時間
TS3k	3k	16	講演59時間
TSf2k	2k	16	講演36時間
IPA2k	2k	16	読み上げ40時間

6 認識実験

6.1 言語モデルと認識率

言語モデルと認識率の関係を図4に示す。音響モデルはTS2kである。Sponを使用すると比較的高い認識率が得られる。WebSp は Web より高い認識率となり、話題適応が有効であった。

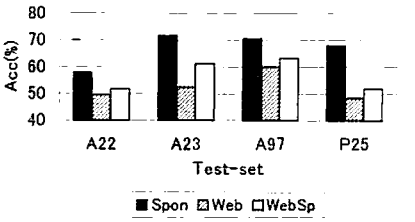

図4　言語モデルと認識率

6.2 音響モデルと認識率

音響モデルと認識率の関係を図5に示す。言語モデルはSponを使用した。図中の4講演については、TS3kで最も高い認識率となった。TSf2kに関しては期待した効果はなかった。

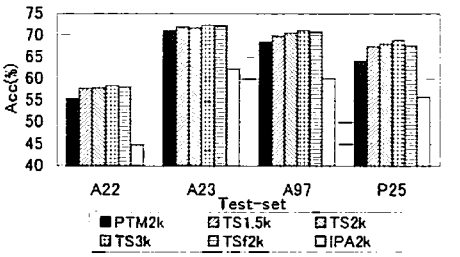

図5　音響モデルと認識率

テストセット中の残り6講演の認識率を表4に示す。言語モデルはSpon、音響モデルはTS2kである。

表4　講演音声の認識率

	J01	K05	N07	S05	Y01	Y05
TS2k	62.66	66.51	66.51	58.04	63.94	63.72

7. まとめ

話し言葉コーパスを利用した言語モデルや音響モデル、教科書を用いた話題適応などが講演音声認識に有効であることが分かった。

謝辞

プロジェクトの推進研究者各位に感謝する。

参考文献

[1] 篠崎隆宏、斎藤洋平、堀智織："話し言葉音声の認識を目指して"、信学技報、SP2000-96 (2000)