E-022

生活環境下でのボタンレス音声認識のための評価データベース構築

Evaluation Database for Buttonless Speech Recognition in Home Environments

大淵康成*，戸上真人*，住吉貴志*
Yasunari OBUCHI, Masahito TOGAMI, and Takashi SUMIYOSHI

1 はじめに

生活環境下での家電品の音声操作への期待が高まっている．現在，多くの家電品がリモコンによる操作を基本としているが，リモコン上のボタン操作体系の複雑化や，リモコンそのものの紛失などの問題があり，ユーザーの不満は大きい．これに対し，家電品を音声コマンドにより操作できるようになれば，利便性が著しく向上すると期待されている．特に，現在では多くの音声認識製品に搭載されている起動ボタンを廃することができれば，利便性の向上は著しい．ただしそのためには，起動ボタンを無くした場合に生ずるであろう大量のフォルスアラームに対する対応が不可欠である．

我々は，このようなブレークスルーを実現するための第一歩として，システム評価のためのデータベースの構築を行っている．これまでにも，音声認識システム開発用の様々なデータベースが提供されてきたが，我々の主眼は，むしろユーザが音声認識起動を望まないときのデータにある．類似のアプローチとして，電子協や RWCP などの非音声データベース [1] があり，多種多様な環境音・雑音をカバーしているが，実生活の中でこれらがどのように現れるかについての情報は含まない．一方，ここに含まれない顕著な「雑音」として，ユーザ自身の世間話や笑い声といったものへの対処も重要である．

これまで，実環境でのボタンレス音声認識の開発のため，模擬生活環境における網羅的な音声データの収集に着手した [2]．その際，音声認識機能を持つ TV リモコンを併用して収録を行うことにより，音声認識使用時とそれ以外の比較を可能とした．本報告では，リモコンの代わりに天井に取り付けたマイクで音声認識を行いながらのデータ収録を行った結果を示すとともに，リモコン使用の場合からの変化についても論じる．

*日立製作所中央研究所，東京都国分寺市

表 1: データベースの概要

	HITHOME	
	07	08
サイズ (時間)	278	375
セッション数 (日) (被験者数=1/2/3)	36 (0/23/13)	49 (12/26/11)
ASR 使用マイク	リモコン	天井
ASR 待受語数 (平均音素数)	17 (11.1)	14 (24.2)
ASR 機能数	11	

2 データベースの概略

一般的な生活環境を整えた部屋で，被験者が朝から夕方まで約 7.5 時間ほど生活する間の音声をすべて録音した．得られたデータは、収録方法の違いにより “HITHOME07” および “HITHOME08” の二つに分けられている．各データベースの概要を表 1 に示す．部屋には多数のマイクを設置したが，本報告では TV リモコンマイク (HITHOME07) および天井マイク (HITHOME07/08) で収録された計 3 チャネルのデータのみを扱う．なお，データは 44.1kHz から 16kHz にダウンサンプリングして使用した．部屋の中の被験者の数は 1 人〜3 人である．収録中は,「テレビの操作は必ず音声認識で行う」以外の指示は与えず，被験者は自由に過ごしている．ただし，収録中に，掃除・洗濯・炊事などを行うよう依頼した．

音声による TV 操作は 11 機能から成るが，一部の機能に複数の待受語を対応させてある．音声認識 (ASR) に使用したのは，HITHOME07 では TV リモコンマイク，HITHOME08 では天井マイクである．なお，収録中の音声認識はボタンにより起動し，ボタン押下時刻のログを保存する．本研究の最終目標は，このボタン押下時刻のログ情報を用いることなく，連続的に録音された音声データから被験者のテレビ操作用発声だけを的確に抽出し，なおかつ認識することである．

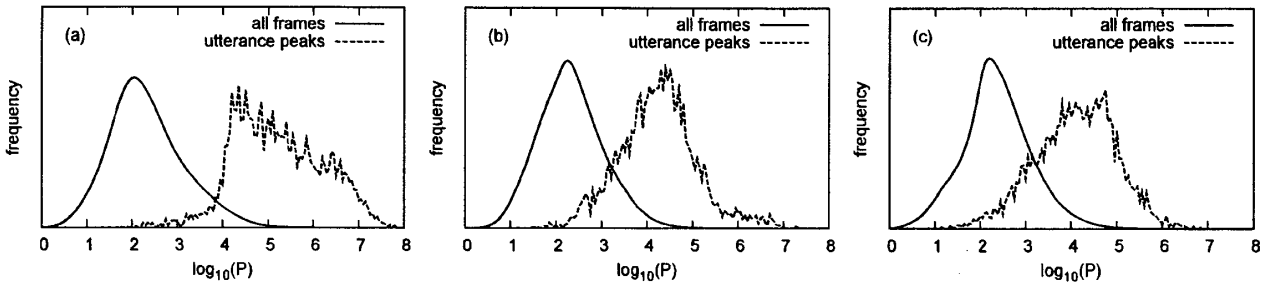

図 1: 短時間対数パワーの分布 (a)HITHOME07-リモコン (b)HITHOME07-天井 (c)HITHOME08-天井

表 2: 語彙内発声に対する音声認識率

DB	発話単位	機能単位
HITHOME07	68.8 %	69.5 %
HITHOME08	44.3 %	49.7 %

3 音声認識率維持のための改良

HITHOME07 収録時に TV 操作用に使用した音声認識のログデータの解析の結果，不明瞭ないし極端に早い発話，言いよどみ，前後への不要語の付加，背景雑音などが誤認識の原因となっていることがわかった．これらは実環境かつオペレータが同席しない収録では不可避の問題であり，この条件で天井マイクによる音声認識を行った場合，認識率が更に大きく劣化し，被験者の振る舞いが不自然になる恐れがある．そこで，HITHOME08 収録に先立ち，音声認識率維持のため以下の改良を行った．

(1) コマンドの複合語化による音素数増加
(2) 反響畳み込みデータによる音響モデル再学習
(3) DCN[3] による特徴量補償
(4) 区間長上限値導入による強制区間検出
(5) 履歴参照による認識語彙絞り込み
(6) ビデオ教材による音声認識方法事前教示

これらの対策の効果を見るため，HITHOME07/08 における音声認識対象データ及び認識結果ログの中から 500 発話相当をそれぞれ無作為抽出し，さらに無発声および語彙外発声を取り除いた上で音声認識の正解率を確認した．その結果を表 2 に示す．発話単位での音声認識率に加えて，TV の動作がユーザの意図にあっていれば良いとする機能単位の認識率も合わせて示した．このとおり，HITHOME07 で約 70%あった認識率が HITHOME08 では 50%弱まで低下しているが，収録そのものが成り立たないような劣化には至らずにすんでいることがわかる．

4 発話スタイルの比較

図 1 に，各マイクのデータを 20ms 幅のフレーム (フレームシフト 10ms) に分けて求めた短時間対数パワーの分布のヒストグラムを示す (全体の積分値が 1 になるよう正規化してある)．ボタン押下後 5 秒以内のフレームの短時間パワーの最大値だけを集めた分布を，全体の分布と比較している．全体平均に対するピーク平均の相対強度を求めると，(a)29.0dB (b)19.1dB (c)17.3dB となり，リモコンマイクと天井マイクでは大きな差があるのに対し，天井マイク同士ではほとんど差が無い．即ち，認識にどのマイクを使うかは，ユーザの発話スタイルに大きな影響を及ぼしていないことがわかった．

5 まとめ

生活環境下での長時間音声データベースを構築し，リモコンマイク使用時と天井マイク使用時の差異を解析した．今後は生活環境に適した音声区間検出方式を開発し，本データベースによる評価を経て，高精度なボタンレス音声認識の実現を目指す．

謝辞

有益なご助言をいただいた，東京工業大学古井貞熙教授，早稲田大学小林哲則教授に感謝いたします．なお本研究は，経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発/音声認識基盤技術の開発」(早稲田大学受託) の再委託を受けて実施しました．

参考文献

[1] S. Itahashi, "On recent speech corpora activities in Japan," *J. Acoust. Soc. Jpn (E)*, Vol.20, No.3, pp.163-169, 1999.

[2] Y. Obuchi, et al., "Always listening to you: Creating exhaustive audio database in home environments," *Proc. INTERSPEECH2007*, Antwerp, Belgium, 2007.

[3] Y. Obuchi, et al., "Normalization of time-derivative parameters for robust speech recognition in small devices," *IEICE Trans.* Vol.E87-D, No.4, pp.1004-1011, 2004.