2008－SLP－73 (5)
2008／10／24

浴室向け音声コントローラに用いる音響モデルの開発

奈木野豪秀†，小笹詩織†，馬場朗‡，日比谷新平‡，竹原清隆‡，庄境誠†

†旭化成株式会社 音声ソリューションビジネス推進部
‡パナソニック電工株式会社 新規商品創出技術開発部

E-mail: {nagino.gb,kozasa.sb,shozai.mb}@om.asahi-kasei.co.jp
E-mail: {baba.a, hibiya, takehara_kiyotaka }@panasonic-denko.co.jp

あらまし 浴室のような高残響でかつ様々な非定常雑音が発生する空間では，音声認識性能の劣化だけでなく，非定常雑音による誤動作に対しても対処が必要となる．本稿では，浴室向け音声コントローラの実現を目指し，性能向上のための音響モデル開発として，雑音及び残響適応音響モデルの開発及び浴室内で発生する非定常雑音を棄却するための雑音棄却音響モデルの開発を行った．また，音響モデルの開発では，音声コントローラが受け付ける語彙に特化した音響モデルの作成手法として，学習データに含まれる音素コンテキストに注目したデータ選択手法の検討を行った．音声認識実験では，雑音適応，残響適応，語彙適応を施すとことで，86.77%から 92.24%と約 50%のエラー削減率を達成した．また，雑音棄却実験では，雑音棄却モデルを用いることで雑音棄却率が 47.43%から 93.19%と大幅な改善率を実現した．

キーワード 音声認識，浴室内，音響モデル，非定常雑音，残響，語彙

Development of Acoustic Model for Speech Interface for Bathroom

Goshu Nagino†, Shiori Kozasa†, Akira Baba‡, Shinpei Hibiya‡,
Kiyotaka Takehara‡, Makoto Shozakai†

† New Business Development, Asahi Kasei Corporation
‡ New Product Technologies Development Dep., Panasonic Electric Works Co., Ltd.

Abstract

In the bathroom, there are many problems for using automatic speech recognition system. For instance, echo, noise and non-stational noise. In this case, importance issue is not only to improve speech recognition performance but also to reduce false operation. In this paper, noise and echo adaptation techniques and noise rejection modeling are proposed. Additionaly, a data selection technique based on frequency of phoneme context in target task and training data set is proposed. In speech recognition experiment, a perfomance is improved from 86.77% to 92.24%. In noise rejection experiment, a rejection perfomance is improved from 47.43% to 93.19%.

Key words Speech recognition, Bathroom, Acoustic model, Non-stational noise, Echo, Vocabrary

1. はじめに

本稿では，浴室を統合制御するコントローラへの音声認識機能の搭載事例について，音声認識システムにおける音響モデルの性能改善手法とその効果について報告する．

近年，浴室空間において，ジェットバスやテレビ等の機能が付加され，より快適な空間が提供され始めている．これらの機能の操作は現状ボタンによるものがほとんどであるが，音声による操作が実現されればより快適な空間となることが期待される[1].

浴室空間での音声認識では，残響と雑音が大きな課題と言える．特に雑音源としては，ジェットバスやテレビ等の機能付加に付随し発生するものだけでなく，利用者が浴室内で発生させる様々な非定常雑音が考えられる．この場合，音声認識性能の劣化だけでなく，誤動作が頻発する恐れがあるため，音声認識性能だけでなく，誤動作をできるだけ軽減するための試みが必要となる．

本稿ではこれらの課題をふまえ，雑音及び残響適応音響モデルの開発及び，浴室内で発生する非定常雑音を棄却するための雑音棄却音響モデルの開発を行ったので，その結果を報告する．また，目的とするタスクの語彙を考慮した音響モデルの開発に関する検討結果もあわせて報告する．

2. 実験条件

本章では，音声認識性能評価用音声データと雑音棄却性能評価用雑音データについて説明を行う．また，ベースラインとなる音響モデル及び評価用ネットワークについても説明を行う．なお，導入として，音声認識処理の概要図を図 1 に示す.

2.1 音声認識性能評価用音声データ

評価音声データは実際に発話者が浴室内で発話した音声が収録されている．浴室のサイズ（m）は 1.6×2.4×2.2 である．話者数は女性 10 名男性 9 名である．また，話者には高齢者及び子供も含まれている．収録は雑音環境が 3 通り（暗騒音，ジェットバス作動時（2 種））となっており，話者の顔向きと話者位置は図 2 に示す通り 3 通り（浴槽内で正面を向いた状態，マイクを向いた状態，洗い場でマイクを向いた状態）である．よって収録の条件は全 9 通りで行われており，各話者は 40 発話の音声コントローラ操作用語彙を各条件下で発話している．発話総数は 2987 である（発話間違い等があるため，話者によっては，発話数は 40 に満たない）．音声のサンプリング周波数は 11kHz である.

2.2 雑音棄却性能評価用雑音データ

実際の家庭の浴室で収録された 16 軒分の実入浴雑音データの内，6 軒分を無作為に選択し，評価用データとして用いており，残りの 10 軒分は学習データに用いている．一回の入浴時間は家庭により異なるがおおよそ 30 分程度である．なお，浴室のサイズや材質は各家庭で異なるため，伝達特性は一致していない.

2.3 ベースライン及び評価用ネットワーク

実験では音声認識エンジンに VORERO（Ver.7）を用いており[2]，ベースラインとなる音響モデルには VORERO に含まれている標準音響モデル（音素 HMM）を用いる．また，キーワッドスポッティング及びリジェクション用のガーベージ音響モデル（GB モデル）も

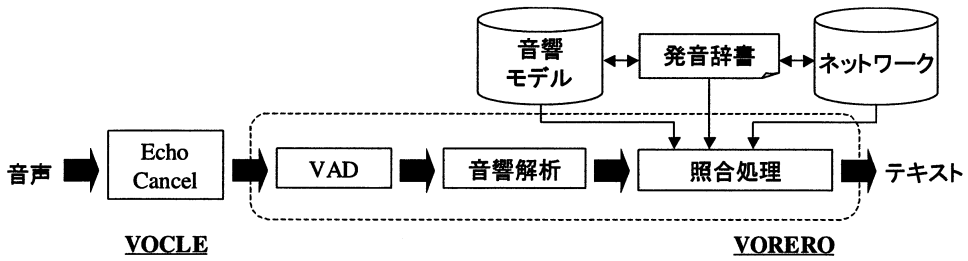

図 1. 音声認識処理概要図

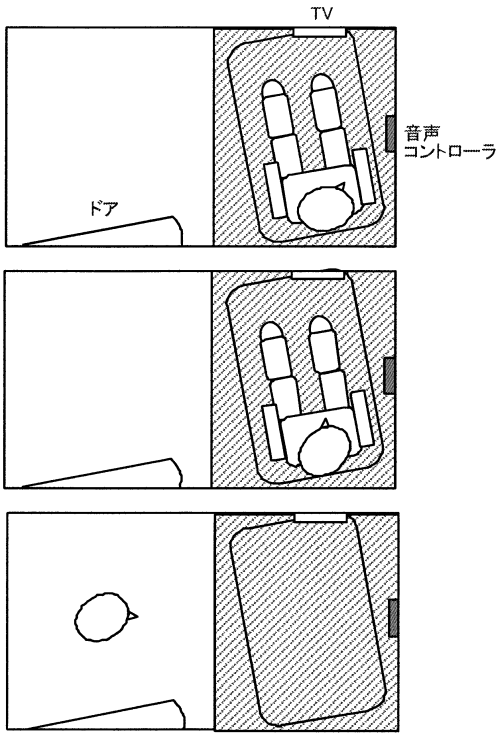

図 2. 話者位置　（図上部より，湯船につかり顔はコントローラ向き/湯船につかり顔はテレビ向き/洗い場位置で顔はコントローラ向き）

同様に VORERO 専用のものを用いる．評価用ネットワークの構造は評価用音声データと整合する 40 単語が並列に並べられており（孤立単語認識），単語の前後及び並列に GB モデルが配置されている．但し，GB モデルを通らない経路も考慮されている．ネットワーク構造を図 2 に示す．

2.4 ベースライン評価

前述の評価データ及びベースライン音響モデル，ネットワークを用いて評価を行った結果を表 1 に示す．音声認識性能は 2.1 節で示した全ての条件における平均の単語正解精度となっており，なお，GB モデルのパラメータの値として，単語の前後に配置されている GB モデルには推奨値である 4.5 を，単語に並列に配置されている GB モデルには推奨値である 2.5 に対し±0.5 の値を設定し，実験を行っている．

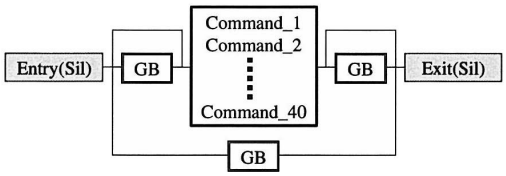

図 2. ネットワーク構造

表 1. ベースライン性能

	GB パラメータ		
	2.0	2.5	3.0
単語正解精度[%]	82.48	86.77	86.77
雑音棄却率[%]	74.51	47.43	31.5

3. 音響モデルの開発

本章では表 1 の性能を改善すべく，まず雑音及び残響適応音響モデルの開発を行う．次に，目的とするタスクの語彙を考慮した音響モデルの開発を行う．最後に，雑音棄却音響モデルの開発を行う．なお，音響モデル作成用の学習音声データには，ベースラインの音響モデル作成用の学習データの Clean セット（データセット A）を用いる．また，サンプリング周波数は 11kHz である．

3.1 雑音及び残響適応音響モデルの開発

まず，耐雑音のために，ベースラインの音響モデルの代わりに，VORERO の Ver8 以降で提供されている非定常雑音環境用標準音響モデル（以下，耐非定常雑音音響モデル）を用いた．この音響モデルには，データセット A に多種の非定常雑音を含む雑音データを重畳することで作成される[3]．また，評価音声データを収録した浴室のインパルス応答を複数の位置で収録し，予備実験で最も性能の良かった位置でのインパル応答を前述の耐非定常雑音音響モデルの学習データに対し畳み込むことで残響適応音響モデルを作成した．

3.2 語彙適応音響モデルの開発

一般に，目的タスクにおいて最も高い精度を示すためには，目的タスクに依存した学習データセットを構築する必要がある．タスク依存性には，雑音，残響などの背景環境に起因する特徴や，発話様式，年齢，性別等の話者性に起因する特徴，発話内容（語彙）等のドメインに起因する特徴などが挙げられる．いずれも重要な要因であるが，本節ではドメインに起因するタス

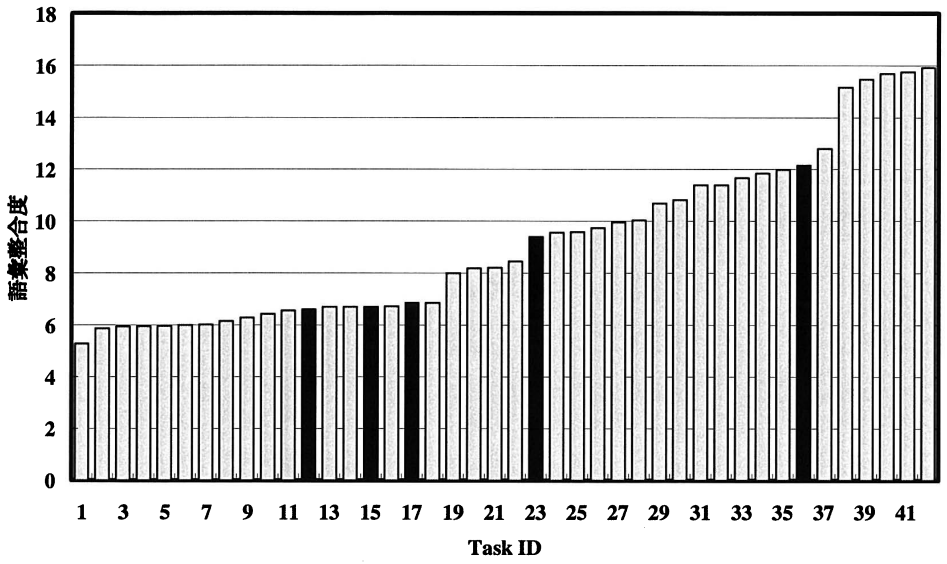

図3. 目的タスクに対する既存タスクの語彙整合度

ク依存性の問題を軽減するために，目的タスクの音素コンテキストの出現頻度を考慮した音響モデルの開発プロセスに関する検討を行う．ドメインAとBの2つのドメイン間の音素コンテキスト出現頻度の整合度（以下，語彙整合度）を次式で定義する．

$$D_{AB} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\{f(s_r^A) - f(s_r^B)\}^2} \quad (1)$$

$$f(s_r) = \frac{g(s_r)}{\sum_{r=1}^{R} g(s_r)} \quad (2)$$

ここで，Rは音素の総数，sは音素，s_rはr番目の音素，$g(s_r)$はr番目の音素の出現頻度である．

データセットAは5つのタスクのデータから構成されている．データセットAに含まれるタスク及びその他に保有する既存タスクのデータセットに対して，語彙整合度を調査した結果を図3に示す．なお，データセットAは図3中の黒色で塗りつぶされたタスクから構成されている．図3より，データセットAに含まれるタスクには，目的タスクの語彙と語彙整合度の低いドメインが含まれていることが分かる．そこで，図3より，語彙整合度の高いドメインを含むタスクを上位 N 個選択することを考える．本稿では，予備実験で最適でった上位 10 個のドメインを含むタスクのデータを学習データとし，前述の耐非定常雑音処理，残響適応処理を施し，音響モデルを作成した．この音響モデルを語彙適応音響モデルと呼ぶ．なお，選択された学習データセットのサイズは，データセットAのサイズを1とした場合に，約0.77に相当する．

3.3 雑音棄却音響モデルの開発

実際の家庭の浴室で収録された 16 軒分の実入浴雑音データから，評価用の 6 軒を除いた 10 軒分のデータを用い，雑音棄却音響モデルを作成する．雑音棄却音響モデルには一般にGMMが用いられることが多いが，計算量の観点や，浴室内で発生する非定常雑音の特徴として時間長が比較的長いことから（残響の影響もある），雑音データをその特徴毎に分類し，複数の状態からなる HMM 音響モデルを作成することを検討する．なお，実入浴データに含まれる雑音の例としては，ドアの開閉音，桶を使う音，浴槽内にはられたお湯に触れる音（チ

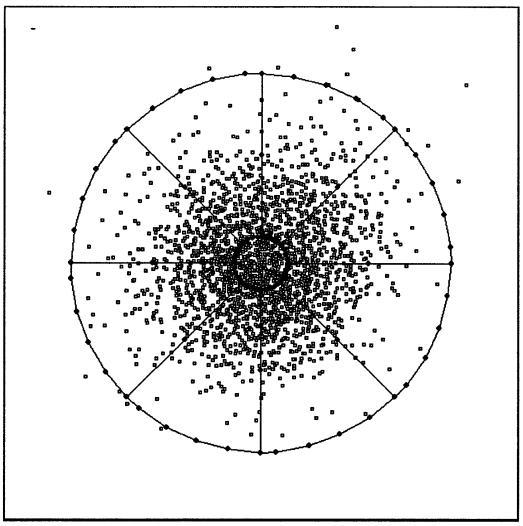

図 4. 雑音 COSMOS Map

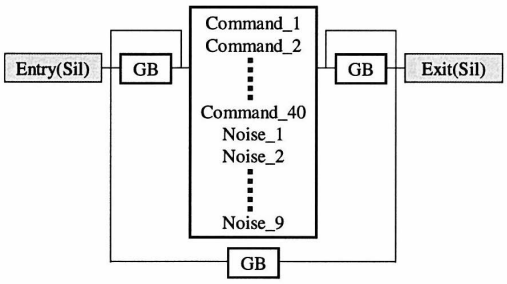

図 5. ネットワーク構造

ャプチャプ音）等がある．

雑音棄却音響モデルの作成手順について説明する．まず，学習用の 10 軒分の実入浴データに対し，VORERO の発話区間検出器により，音声と誤検出されたデータの特徴量系列を求め，簡易 COSMOS 法[4]により 2 次元空間に写像する．なお，時系列長が異なる 2 つの特徴量系列間の距離計算には DTW 法が適用されている．写像結果を図 4 に示す．各点はそれぞれ誤検出された雑音データに相当し，位置が近いデータの特徴は似ており，中心付近には比較的時間変動の小さい雑音データが集まっており，中心から遠ざかるほど，時間変動の大きい突発性雑音データが含まれる割合が多い．音の特徴に合わせ，図 4 に示されるようにデータを 9 つに分割し，それぞれの区域毎に HMM 音響モデルを作成した．状態数は各区域の雑音データの時間長から，区域毎に定め，状態毎の分布の数は 1 とした．なお，作成した音響モデルは雑音棄却音響モデルとして，図 1 の評価ネットワーク上の単語グループに追加している（図 5 参照）．

4. 実験結果

3 章で作成した各音響モデルを適用した場合の音声認識性能及び雑音棄却性能を表 2 及び表 3 に示す． なお，表 2 の結果は雑音棄却音響モデルを含まない，図 2 のネットワークを用いた場合の性能を示している．また，表 3 の結果は，ベースラインの評価には図 2 のネットワークを用い，雑音棄却音響モデル使用時（表中の「棄却音響モデル」）の評価には図 5 のネットワークを用いており，音響モデルは共にベースライン音響モデルを用いている．

表 2 より，各取り組みにより性能が改善するだけでなく，GB パラメータの違いによる性能変動が小さくなり，安定性の高い音響モデルとなっていることがわかる．また，表 3 より，雑音棄却音響モデルを導入することで，棄却性能が大幅に向上し，同様に，GB パラメータの違いによる性能変動が小さくなり，安定性が向上することが分かる．なお，語彙適応音響モデル（耐非定常雑音処理及び残響適応処理適用済み）に雑音棄却モデルを導入した場合の音声認識性能及び雑音棄却性能の劣化は僅かである．

なお，GB パラメータ=2.5 とした場合の，話者位置別（図 2 における，湯船につかり顔はコントローラ向き/湯船につかり顔はテレビ向き/洗い場位置で顔はコントローラ向き，をそれぞれ P1/P2/P3 とする）及び雑音環境別（暗騒音/ジェットバス 1/ジェットバス 2 をそれぞれ N1/N2/N3 とする）の各音響モデルの音声認識性能を表 4 及び表 5 に示す．

表 2. 音声認識性能 [%]

音響モデル	GB パラメータ		
	2.0	2.5	3.0
ベースライン	82.48	86.77	86.77
耐非定常雑音	85.40	87.87	87.87
残響適応	88.94	89.25	89.25
語彙適応	92.14	92.24	92.31

表 3. 雑音棄却性能 [%]

雑音棄却手法	GB パラメータ		
	2.0	2.5	3.0
ベースライン	74.51	47.43	31.5
棄却音響モデル	96.59	93.19	89.27

表 4. 音声認識性能
（話者位置別，GB パラメータ=2.5，雑音環境：N1）

音響モデル	話者位置		
	P1	P2	P3
ベースライン	89.73	85.58	85.18
耐非定常雑音	89.78	84.93	83.45
残響適応	92.9	89.55	89.98
語彙適応	95.81	93.4	92.54

表 5. 音声認識性能
（雑音環境別，GB パラメータ=2.5，話者位置：P1）

音響モデル	雑音環境		
	N1	N2	N3
ベースライン	89.73	82.38	81.84
耐非定常雑音	89.78	80.55	82.18
残響適応	92.9	88.42	90.56
語彙適応	95.81	91.94	93.03

5. まとめと今後の方針

本稿では，浴室向け音声コントローラに用いる音響モデルの開発を行った．音響モデルの性能改善方策として，耐雑音，残響適応，語彙適応による音響モデルを作成し，浴室内で発生する非定常雑音による誤動作軽減策として雑音棄却モデルを作成することで，音声認識実験では，86.77%から 92.24%と約 50%のエラー削減率を達成し，雑音棄却実験では，47.43%から 93.19%と大幅な改善率を実現した．

冒頭でも述べたように，浴室での音声コントローラを用いる動機は「より快適な環境」を適用することにある．この場合，使用者にはよりリラックスした状態であることが望ましく，発話様式もその「リラックス度」に応じて変動することが予想される．そのため，今後の方針として，使用者の「慣れ」，「リラックス度」に起因した発話様式の変動にも対応する必要があると考えられる．

6. References

[1] 馬場，日比谷，奈木野，小笹，庄境，竹原，“浴室向け音声コントローラの開発，”SLP, Oct, 2008.

[2] 庄境，“組み込み向け音声認識ミドルウェアＶＯＲＥＲＯの開発，”音講論集，1-8-13, pp.31-32, Mar, 2004.

[3] M. Shozakai and G.Nagino, “Improving Robustness of Speech Recognition Performance to Aggregate of Noises by Two-Dimensional Visualization,” EUROSPEECH, pp.921-924, 2005.

[4] 柴崎，庄境，“鉄道分野における音声・音響センサの応用，”電気学会論文誌 (E)，vol.127，no.11, pp.493-498, 2007.