社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－SLP－53　(9)
2004／10／22

公共音声情報案内システム「たけまるくん」の運用および収集発話の分析

李　晃伸[†]　山田　真士[†]　西村　竜一[††]　鹿野　清宏[†]

† 奈良先端科学技術大学院大学 情報科学研究科　〒 630–0192 奈良県生駒市高山町 8916–5
†† 和歌山大学 システム工学部　〒 640–8510 和歌山市栄谷 930
E-mail: †{ri,masasi-y,shikano}@is.naist.jp, ††nisimura@sys.wakayama-u.ac.jp

あらまし　機械に対するユーザの自然な実発話の収集と統計的な分析のために，我々は音声情報案内エージェントシステム「たけまるくん」を公共施設に設置し，2004 年 5 月までの 19ヶ月間で約 17 万発話を収集・整備した．本稿では現在のシステム構成，収集データの分析結果および雑音・不要音棄却実験の結果を報告する．全体のおよそ 30% が雑音などの非音声入力であった．音声入力のうち 81%が有効発話であり，残りは背景会話・無意味な発声・不明瞭で聞き取れない発声・発話断片・オーバフローなどの応答不能な無効発話であった．これらの無効発話に対して，入力長と GMM に基づく雑音・不要音棄却の性能を評価した．1 か月分 8,248 個のデータで実験した結果，雑音・息・咳・笑い声などの非音声入力は 99%棄却でき，叫び声や遠隔で発声された背景会話もある程度棄却できることが分かった．一方で，発話断片やドメイン外発話については音響的特徴からの弁別は難しかった．

キーワード　音声対話システム，実環境，GMM，不要音棄却，自然発話データベース

Public Speech-oriented Information Guidance System “Takemaru-kun”

— Its Long-term Operation and Collected Data —

Akinobu LEE[†], Masashi YAMADA[†], Ryuichi NISIMURA[††], and Kiyohiro SHIKANO[†]

† Graduate School of Information Science, Nara Institute of Science and Technology　Takayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan
†† Faculty of Systems Engineering, Wakayama University　Sakaedani 930, Wakayama, 640–8510 Japan
E-mail: †{ri,masasi-y,shikano}@is.naist.jp, ††nisimura@sys.wakayama-u.ac.jp

Abstract　In order to collect user’s actual utterances to a speech dialogue system on real situation, we have located a speech-oriented information guidance system called “Tekemaru-kun” at a public civil hall, and collected 177,789 inputs via 19 months’ operation. This paper will report the current system architecture, details of collected data and experimental results of invalid input rejection. As a result, non-voice (noise) inputs occupies about 30% of total input, and 81% of voice inputs are valid inputs. The rests are invalid voice inputs that includes background speech, incomprehensible voice, obscure speech, fragmented speech, level overflow and so on. Rejection of those invalid inputs has been examined based on input length threshold and GMM-based identification. Experiments on 8,248 inputs of one month showed that almost all of noise and non-verbal inputs such as breath, coughing and laughter can be rejected successfully, and distant background speech and shouts were also discriminative, whereas out-of-domain utterance, obscure speech and fragments cannot be detected only by the acoustic property.

Key words　Spoken dialogue system, utterance collection, Gaussian mixture model, invalid input rejection

1. はじめに

近年，音声言語処理の要素技術は着実な進歩を遂げている．連続音声認識のための可搬性の高い汎用ツールが共有・整備され [1]，音声対話エージェント構築のためのツールキットが開発されつつある [2] など，音声インタフェースに対する注目が高まっている．しかし，実際に音声対話システムが広く実用化された例はほとんどない．より自然で身近な音声インタフェース

を実現するためには，特に機械に対してユーザが発声するその振る舞いを統計的に分析し，自然なユーザ発声を収集・分析を行うことが必要である．

我々はこれまでに，音声情報案内エージェントシステム「たけまるくん」を公共施設に恒常的に設置し，約 2 年間にわたりフィールドテストおよびユーザ発話の収集を行ってきた [3]．システムは 2002 年 11 月の設置以来現在に至るまで，平均で一日あたり 388 発話の高頻度で一般ユーザに利用され続けており，使いやすく親しみのある音声対話システムとして広くユーザに受け入れられているといえる．本稿ではまず，本システムのハードウェアおよびソフトウェアの構成と，運用状況について詳細を述べる．

認識対象となった全ての入力音声についてシステム上に収録し，書き起こしとタグ付けを行っている．本稿では，このデータベースの整備について詳しく述べると共に，執筆時点で整備が終了している 5 月末までの約 17 万の収集データについて，非音声入力の割合等を集計した結果を報告する．

また，実環境においては，音声区間の誤検出や非音声入力によって音声認識システムが誤動作しやすく，これが音声インタフェースの入力系としての確実な動作を損ねる大きな要因となっている．特に実環境においては，周囲の雑音や背景会話，笑い声，咳などがマイクに混入することは避けられない．本システムでは，これらの無効入力をユーザの発話から弁別し棄却するために，Gaussian Mixture Model (GMM) に基づく雑音・不要音棄却 [4] [5] を行う．この機構が実際のユーザの入力に対してどの程度効果があるのかについて，検証実験を行った結果を報告する．

2. 公共音声情報案内システム「たけまるくん」

本システムは，館内施設や周辺情報の案内を行う音声対話システムである．ユーザの質問に対して，合成音声とキャラクタのアニメーション，および関連する Web ページを用いて応答する．対話戦略は単純な一問一答形式であり，対話履歴や対話状態は用いていない．このシステムは，生駒市北コミュニティーセンターのロビーに常設されている（図 1）．なお「たけまるくん」は同市のマスコットキャラクターである．

2003 年のシステム [3] からは，以下の点が改善されている．

- 年齢層別並列デコーディングによる年齢層識別の実装
- 入力長と GMM に基づく不要音棄却の実装
- 動作の安定化と高速化
- Web ベースの状態監視システム

以下，全体構成および個々のモジュールについて解説する．

2.1 システム構成

システムは複数のモジュールから構成されている．全体構成を図 2 に示す．モジュール間の連携には，黒板モデルを参考にしている．各モジュールは独立した実行可能なプログラム（プロセス）であり，TCP/IP ソケットを通じて通信を行う．全モジュールの状態は一括して “Status server” と呼ばれるプロセスに保持される．各モジュールはそれぞれのタイミングで，自らの状態変化や出力を status server に逐次書き込む．また必要に応じて status server から他のモジュールの状態を読み込むことができる．例えば，音声認識モジュールは入力待ち・音声入力開始・入力終了・棄却の有無・認識結果などの情報を音声入力の状態に合わせて適宜サーバに書き込む．agent モジュールは status server 上の上記の音声認識モジュールの状態情報を常に監視し，変化があった場合にそれに合わせてアニメーションを表示する．この様式は，モジュールの追加や単体・ダミーモジュールによる個々の動作検証などが容易に行える利点がある．

図 1　音声情報案内システム「たけまるくん」
Fig. 1　Speech information guidance system “Takemaru-kun.”

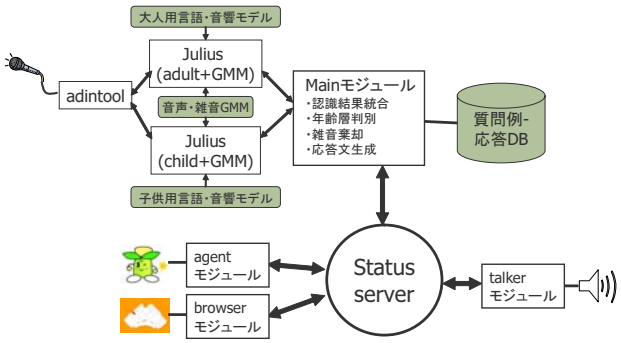

図 2　全体および音声認識部のモジュール構成
Fig. 2　Module design.

ハードウェア構成は，PC がメイン・サブの計 2 台，液晶モニタ 2 台，および周辺機器からなる．OS は Linux を使用している．メイン PC は status server および音声認識・音声合成の各モジュールとブラウザ画面の表示を担当し，サブ PC はエージェントの表示を担当する．メイン PC は，後述の年齢層別ディクテーションで複数の認識エンジンを並列動作させるため 2CPU 構成となっている．

2.2 音声認識部

音声認識モジュール（図 2 の上側）は，単語 N-gram を用いた大語彙連続音声認識を行う．また，マイク入力の監視，音声区間の切り出し，話者の年齢層判別，雑音棄却などもここで行う．音声認識エンジンには，Julius のバージョン 3.4.2 を拡張したものを使用している．

年齢層別の言語モデル・音響モデルを用いることで，特に子供の認識精度を向上させている．さらに，両認識結果の尤度を比較することで話者の年齢層判別を行う [6]．判別結果を元に，応答生成部では大人と子供の各年齢層に即した応答を行う．

また，実環境における雑音の誤検出や笑い声などへの対処として，入力長および GMM に基づく雑音・不要音の棄却 [4] を行っている．GMM の尤度計算は Julius 内に組み込まれており，音声認識処理と同時に GMM の尤度計算も平行して行われる．これについては第 3 節で詳細に述べる．

2004 年のシステム改良により，高速性と安定性が大幅に改善した．音声入力のプロセス間のパイプライン化，および GMM の尤度計算を認識処理と並列化するなど，認識から応答生成までの各部を最適化した結果，ユーザに遅延をほとんど感じさせない速度で応答を開始できている．また，棄却時や認識再開時のモジュールの相互制御を最適化した結果，メモリリークやネットワーク経由の音声入力動作の安定性が大きく向上した．システム上では計 10 個のプロセスが動作しているが，2004 年 4 月以降システムの不具合によるハングアップは生じていない．

2.3 応答文生成部

応答文生成では，認識結果をもとにあらかじめ用意された応答文テキストから候補を選択して出力する．応答文は 299 文用意されており，ホール内や周辺の施設案内が 60 文 (20%)，生駒市の施設や観光案内が 89 文 (30%)，時間・天気・ニュースなどの情報提示が 30 文 (10%)，あいさつや自己紹介が 78 文 (26%)，相づちや悪口への受け答えなどが 42 個 (14%) となっている．天気予報は Web から取得するが，応答時に情報を取得するとネットワークの遅延により応答の遅延が発生することがあるため，ここでは一定時間おきに情報をあらかじめ取得している．

応答文は，認識結果と対応する質問文のデータベースに基づいて選択される [3]．まず認識結果のうち自立語について，質問文とのマッチングを行い，類似する質問文をデータベースから抽出し，それらに対応する応答文にスコアを付与する．質問文例を用いた認識結果の整形に近く，たとえ質問がシステム設計者の想定範囲からはずれたものであっても，それに近い応答を選択できると期待できる．

年齢層判別の結果を用いて，大人と子供で異なる文章が生成される．質問文データベースは大人 3,307 文，子供 6,573 文を別々に保持している．応答文は種類と数は同じであるが，子供に対する応答はより柔らかい語尾表現を用いるなどの変更を行っている．

2.4 出 力 部

ユーザへの応答の出力は，音声合成モジュール（talker），エージェントのアニメーションモジュール（agent），および関連情報を提示する Web ブラウザモジュール（browser）からなる．talker は応答文の音声を合成し，再生する．ソフトウェアはクリエートシステム開発の日本語音声合成ライブラリを使用している．応答音声の出力とアニメーションは同期して開始される．すなわち，talker が合成音声を生成する間 agent と browser は待機し，音声ファイルが完成したのち，同時に出力を開始する．browser は，場所の問い合わせに対して地図を表示するなど，Web ブラウザを用いて応答文に応じた情報を提示する．

agent は応答文に対応するエージェントのアニメーションを表示する．Macromedia Flash を用いて作成されており，現在 71 種類の応答用アニメーションが用意されている．さらに，音声入力開始時にはエージェントがうなずくアニメーションを表示する．これにり，ユーザはシステムが自分の音声を検出したことを確認できる．なお，応答動作中に次のユーザ発話が割り込んだ場合，その応答動作は即時にキャンセルされ新たな入力が優先される．

3. 雑音・不要音の棄却

実環境下の音声対話システムでは，ユーザの入力発話以外に，物音や周囲の背景雑音などの非音声入力や，笑い声や咳などの不要音がマイクに混入する．また，ユーザのに自然な発話スタイルを許すほど，ユーザがマイクを意識せずに発話しがちなため，マイクから離れた不明瞭な入力になったり，システムが発話を正しく切り出せずに断片化するケースが増大する．このような雑音や不要音，不明瞭発話の入力は，システムの誤動作の大きな要因となる．Push-to-talk を用いることでユーザが明示的に発話タイミングを指示することも考えられるが，ユーザにとってより自然な音声対話システムを目指すには，ユーザがスイッチを意識せずに利用できることが重要であると考える．このため，これらの雑音や不要音，不明瞭発話をシステム上で識別および棄却する機能が不可欠である．

本システムでは，入力長および Gaussian mixutre model (GMM) の 2 つの基準に基づいて雑音・不要音の棄却を行っている．以下，各手法について詳細に述べる．

3.1 入力長に基づく誤検出の棄却

足音やマイクを叩く音などの突発性の雑音による誤検出を抑制するために，検出した入力区間の長さがしきい値以下であった場合に，その入力を棄却する．遠距離での背景発話などによる短い区間の誤検出抑制する働きも期待される．Julius は 2 パス探索のため，第 1 パス終了時に入力長が決定した段階で判定を行い，その長さがしきい値以下であれば第 2 パスを行わずにそこで認識を中断する．

3.2 GMM に基づく不要音棄却

認識処理と平行して，Gaussian Mixture Model (GMM) に基づく入力音識別を行い，雑音や笑い声などの不要音の棄却を行う．具体的には，本システムで収集したデータの 2002 年 11 月から 2003 年 3 月までのデータから学習した成人・子供・笑い声・咳・雑音の計 5 クラスの GMM を用いる．1 つあたり 128 混合ガウス分布からなる [4].

GMM の計算は Julius 内部に組み込まれる．尤度計算と棄却の仕組みを図 3 に示す．発話の尤度計算は認識処理の第 1 パスと平行してフレーム単位で行われる．入力終了時（第 1 パス終了時）にその入力長と GMM の識別結果を調べ，入力長が指定時間以下であるか，または最尤の GMM が非音声モデルであった場合には，続く第 2 パスの処理をキャンセルして入

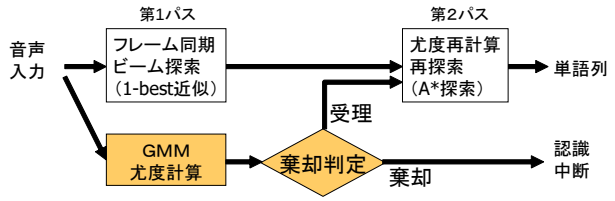

図 3　Julius における雑音・不要音棄却
Fig. 3　Invalid input rejection on Julius.

力を棄却する．また，混合分布の分布集合のうち上位 N 分布のみ計算する Gaussian pruning [7] を導入し，実質的な計算量を抑えている．また 入力が棄却された場合，Cepstral Mean Normalization (CMN) のパラメータを更新しない．

4. 運　　用

本システムは，生駒市の協力のもと生駒市北コミュニティーセンターのロビーに常設されている．2002 年 11 月 6 日の開館日以降，休館日を除くすべての日に 24 時間稼動している．実際の運用においては，初期のシステムの不具合によるハングアップやネットワークの不調によるフリーズ，ユーザのいたずらによる電源断，金属疲労によるマイクの破断等の障害が生じていた．

公共の場で監視者無しに長期間安定した運用を行うために，Web ベースの状態監視システムを構築した．システム上で 10 分おきにプロセス数と最終発話時刻を調べ，結果を Web サーバへ HTTP 経由で送信する．一定時間（現在は 30 分）以上送信が途絶えるか，あるいはシステム内のモジュールのプロセス数に変化が生じたとき，システムに不具合が生じた可能性が高いとして，警告メールをあらかじめ定められた管理者のアドレスへ送信する．これにより障害発生の検知が容易になった．さらに，発話数や雑音棄却数，話者層ごとの発話数などの収録情報を日ごとに自動集計して，サーバーへ送信する．現在のシステムの状態および過去の統計は Web ブラウザや携帯電話から見ることもできる．

5. データ収集

設置開始日より，システムに入力されたすべての音声入力を記録し，手作業による書き起こしとクラス分け・タグ付けを行っている．2004 年 7 月までの 21ヶ月間で収集したデータ数は，200,855 発話，約 93 時間，ファイル総量は約 10.8GB である．一日あたりの平均入力数は 388 である．

入力単位ごとに，発話内容の書き起こし，および性別・年齢層・有効／無効のクラス分けを行っている．性別については，物音やノイズ，息などの非音声入力は「Z:非音声」に分類する．年齢層については，特に子供について，年齢による発話様式の変化を考慮して乳幼児，低学年子供（10 歳未満程度），高学年子供（10 歳以上〜高校生程度）に分ける．有効／無効のクラス分けでは，たけまるに対して明らかに質問する意図をもって発声されている，書き起こし可能な程度に明瞭な発話には “Y” に，雑音や意味がとれない発声，明らかにたけまるに対する発話ではない発話には “N” に，どちらか判断がつけられない発話には “X” に分類する．分類クラスの一覧を表 1 に示す．

表 1　分類クラス一覧
Table 1　Classification

項目名	内容
性別	M:男性/F:女性/X:判別困難/Z:非音声
年齢層	a:乳幼児/b:低学年子供/c:高学年子供/ d:成人/e:高齢者/z:判別不能
有効発話	Y:有効 / N:無効 / X:判別困難

表 2　タ グ 一 覧
Table 2　Annotated tags

タグ名	対象となる入力
雑音	物音や足音，マイクを擦る音などのノイズ．
不要	ノンバーバルな発声．笑い声・咳・息・ 乳幼児の発声・歌など．
不明瞭	ひとり言や遠くの発話など，不明瞭で書き起こしが困難な発話．
背景会話	他の人間との雑談などが収録されたもの．
雑音混入	発声中に物音等の雑音が混入している音声
背景会話あり	他の人間の発話が混入している発話
文頭落ち	文頭が切れている発話
文末落ち	文末が途切れている発話
レベル不足	入力レベルが小さすぎて書き起こしできない音声
オーバーフロー	振幅レベルがオーバフローしている音声

作業は一貫して 3 名の作業員によって人手で行われている．クラス分けの判断は収録音声データのみから作業員の主観で行っている．どのクラスに属するか個人で判断がつかない場合，合議により決定を行うが，なおも決定できない場合は「判別困難」のクラスに分類する．男性と女性のどちらとも判断がつかない場合は「X:判別困難」，非音声など年齢層で分類できないものは「z:判別不能」のクラスに分類する．

さらに，不明瞭な発話や雑音などの入力に対してその種類を表すタグを付与する．タグの種類の一覧を表 2 に示す．ひとつの入力に複数のタグが付与されることもある．

現時点で，2004 年 5 月末までの 177,789 発話についてデータ整備が完了している．発話数の内訳を表 3 に示す．子供のユーザが全体の 58.1% を占めており，特に低学年子供の利用者が多いことが分かる．男性・女性ともにほぼ利用頻度は同じである．低年齢のユーザほど性差が音声に現れにくいため，性別の判断が行えないケースが多かった．また全体のおよそ 30% が非音声入力であり，システム上でこれらを棄却できることが求められる．

各年齢層ごとの有効／無効発話の割合を表 4 に示す．乳幼児以外は発話の 80%程度が有効発話であり，本システムが広く利用されていることが分かる．年齢別では，成人に近いほど有効発話の割合が高い傾向にあった．また，年齢層が判別不能な非音声入力のほとんどが無効発話であることが確認された．全ての年齢層で平均した有効発話の割合は 81%であった．なお，有効／無効の判別が困難な入力の多くは，マイク付近でたけまるについて他人と話している会話や，文頭落ちや遠隔発話の断片，

表 3　年齢層・性別別の収録発話数

Table 3　Overview of collected data

年齢層＼性別	男性	女性	判別困難	非音声	総数
乳幼児	6,793	8,862	4,153	4	19,812 (11.1%)
低学年子供	25,015	28,527	11,487	2	65,031 (36.6%)
高学年子供	8,655	7,552	2,303	1	18,511 (10.4%)
成人	12,874	8,620	165	13	21,672 (12.2%)
高齢者	103	204	1	6	314 (0.2%)
判別不能	3	4	16	52,426	52,449 (29.5%)
計	53,443	53,769	18,125	52,452	177,789(100.0%)

(注)2002 年 11 月〜2004 年 5 月

表 4　有効発話／無効発話の分類結果

Table 4　Classification result of valid / invalid inputs

年齢層	有効	無効	判別困難	有効の割合
乳幼児	13,535	5,534	743	68%
低学年子供	53,422	9,605	2,003	82%
高学年子供	15,605	2,548	358	84%
成人	19,193	2,024	455	89%
高齢者	244	67	3	78%
判別不能	2,214	50,034	200	4%

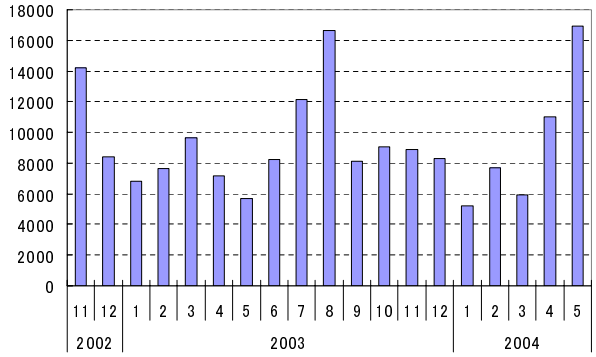

図 4　月ごとの入力数

Fig. 4　Collected utterances per month.

無意味な問いかけなど，有効か無効かの区別がつけ難いものであった．

月ごとのデータ収集数を図 4 に示す．グラフより，本システムが一般ユーザに長期間にわたって継続して利用され続けていることが分かる．2004 年 4 月以降に発話数が上昇しているのは，4 月 9 日に雑音・不要音の棄却および応答速度の改善が行われたことが関与していると考えられる．また，月ごとの年齢層別のユーザの割合を図 5 に示す．非音声入力（図中の「判別不能」）の割合は，設置当初から大きな変化は見られない．また，成人の発話は設置直後は全体の 24%あったが，その後徐々に減少し，現在は 10% 程度にとどまっている．

無効発話において，タグが付与されたデータ数を図 6 に示す．「応答不可発話」は明瞭な無効発話であり，ドメイン外発話に対応する．「不要」のタグについては，さらに笑い声・咳・息・その他に細分した．「不要：その他」は主に「あー」「わー」など発話者の意図が読み取れない声が多く含まれている．結果として，

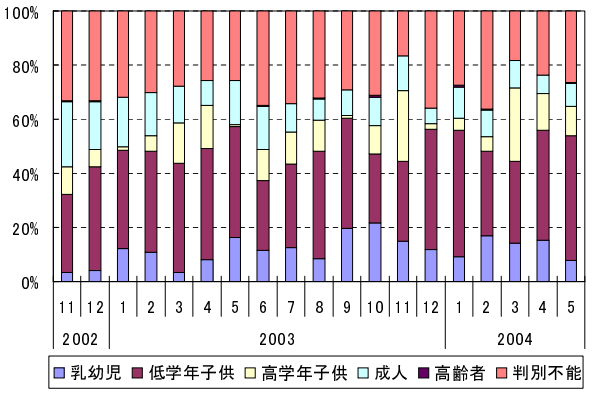

図 5　月ごとの年齢層の割合

Fig. 5　Populations of age classes per month.

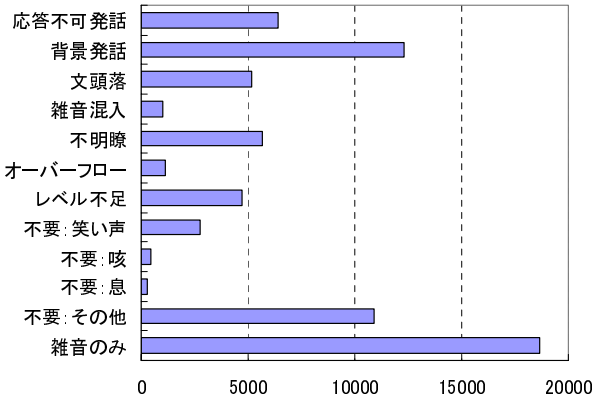

図 6　無効発話におけるタグ数

Fig. 6　Number of tags on invalid inputs.

無効発話のうち 37% が雑音のみによる誤トリガ入力であった．複数人数で喋ったものや，一人が入力している背景で会話している音声はほとんどが背景発話に属するため，背景発話の割合が多くなっている．年齢層については，オーバーフローのほとんどが子供の音声であった．また特に乳幼児において不明瞭や不要発話の数が多く見られた．

6.　雑音・不要音棄却実験

入力長および GMM に基づく雑音・不要音棄却の性能を，実際の収録データで評価した．テストデータとして 2004 年 4 月 10 日から 30 日までの 8,248 入力を用いた．

棄却処理は，まず入力長による棄却を行い，次に GMM による判別を行う．棄却用の GMM は大人発声 (adult)，子供発声 (child)，笑い声 (laughter)，咳 (coughing)，雑音 (noise) の計 5 つからなり，それぞれ 128 混合分布を持つ [4]．最尤の GMM が adult もしくは child のときに受理，それ以外の場合は棄却する．GMM 計算時には実際には上位 10 個の分布のみを計算している．また短時間入力棄却のしきい値は，先頭と末尾の無音部を含めて 0.8 秒とした．

表 5 に，入力データの内訳および棄却率を示す．ここで有効発話とは，明瞭でかつたけまるが返答可能な発話を表す．無効発話の内訳は，複数カテゴリに属する発話もあるため入力数の

表 5　発話棄却実験の結果

内容	入力数	棄却数 短	棄却数 GMM	棄却率 (%)
有効発話	4450	567	171	16.58
無効発話　全体	3654	1957	1153	85.11
ドメイン外	381	99	45	37.80
文頭落	309	178	33	68.28
雑音混じり	167	65	44	65.27
背景会話	567	244	220	81.83
内　不明瞭	244	125	67	78.69
意味不明	455	248	175	92.97
訳　雑音のみ	1139	720	419	100.00
笑い声	160	47	111	98.75
泣き声	11	0	10	90.91
息	69	11	58	100.00
咳払い	38	16	22	100.00
レベル不足	300	299	1	100.00
判断できず	144	—		

合計は一致しない．棄却判定は入力長，GMM の順で行っており，入力長で棄却されたものは GMM 判別には含まれない．

有効発話の 16.58% が誤棄却された．多くは入力長によるもので，実際には挨拶や悪口などの早口かつ短い発話が多く見られた．本来のタスクである質問発話を誤棄却するケースはほとんど見られなかった．

無効発話については，全体の 85.11% を棄却できていることが分かった．雑音や笑い声・泣き声・息・咳払いなどの非音声入力については，GMM によりほぼすべて棄却可能である．背景会話・不明瞭・意味不明・レベル不足などは，人どうしの会話などの誤検出である．これらは人の音声のため GMM での判別は容易ではないが，レベルが小さくパワーが持続しない傾向があり，入力長による棄却がある程度機能していると考えられる．またこのような発話は一般にマイクから離れた遠隔発話であることが多いため，GMM でも一定量の識別が行われている．一方，ドメイン外発話，文頭落発声（発話開始部が落ちている発声），雑音交じりの発声については，発話が比較的明瞭であるため GMM や入力時間での棄却は難しい．これらは入力全体の 1 割程度を占めており，今後の対策が望まれる．

図 7 に有効発話/無効発話ごとの入力長のヒストグラムを示す．無効発話は短い入力が多く，0.7 秒以上 0.8 秒未満の区画がピークであった．このことから，入力長によって無効発話を効率よく棄却できていることが分かる．一方で，棄却される有効発話数も少なくない．今後は GMM との棄却基準の統合が求められる．

7. おわりに

音声情報案内システム「たけまるくん」について，近年の改善点を含めたシステム構成と収集データについて述べ，不要音・雑音棄却の性能評価を行った．入力の約 30%が非音声入力であり，音声であるが入力として無効な背景会話や不明瞭な入力，笑い声などを含めると全体の 4 割以上が不要な入力であることが分かった．入力長と GMM による棄却実験の結果，無効発話の 85.1% が棄却できた．非音声については高い精度で棄却でき，背景発話や叫び声などについてもある程度判別可能であるが，ドメイン外発話などは音響的特徴からの弁別は難しい．また有効発話の 16.6%が誤棄却された．

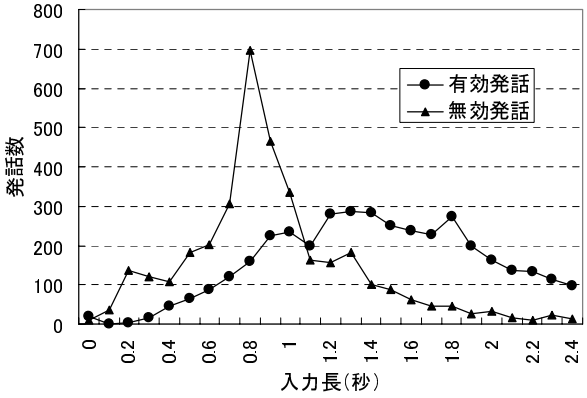

図 7　有効発話/無効発話ごとの入力長の分布
Fig. 7　Distribution of input length for valid/invalid input.

今後は，特に背景会話やドメイン外発話の検出について，言語的性質や確信度と組み合わせた検出方法を検討するとともに，GMM および入力長についても各基準を統合することにより，頑健で確実な雑音・不要音棄却法について検討する．また，収集データについては，ユーザの経年変化や発話傾向など，より多角的な分析を試みたい．

謝辞

この研究（の一部）は，文部科学省のリーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われた．本実験に全面的にご協力いただいている生駒市の関連各位に感謝いたします．また，データベース整備にご尽力いただいている技術補佐員の方々に感謝します．

文　献

[1] 河原達也他．連続音声認識コンソーシアム 2002 年度ソフトウェアの概要．情処学研報, 2003-SLP-48-1, 2003.

[2] T. Nitta et al. Activities of Interactive Speech Technology Consortium (ISTC) Targeting Open Software Development for MMI Systems. Proc. IEEE Intl. Workshop on Robot and Human Interactive Communication (Roman2004), (to come), 2004.

[3] 西村竜一他．実環境研究プラットホームとしての音声情報案内システムの運用．信学論, Vol.J87-D-II, No.3, pp.789–798, 2004.

[4] 中村敬介他．実環境音声情報案内システムにおける環境雑音および不要発話の識別．信学技報, SP2003-172, 2004.

[5] A. Lee et al. Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs. Proc. ICSLP2004, (to come), 2004.

[6] 西原洋平他．ユーザ層別並列モデルを用いた音声情報案内システム．春季音講論, 1-8-22, pp.49–50, 2004.

[7] A. Lee et al. A new phonetic tied-mixture model for efficient decoding. Proc. ICASSP2000, pp.1269–1272, 2000.