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

2006－SLP－64 (30)
2006／12／21

「たけまるくん」実環境音声案内システムのデータベース整備と「キタちゃん」へのポータビリティーの検討

トビアス・ツィンツァレク† 川波 弘道† 木田 学† 猿渡 洋† 鹿野 清宏†
西村 竜一†† 李 晃伸†††

† 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デザイン情報学科 〒640-8510 和歌山市栄谷 930
††† 名古屋工業大学大学院工学研究科情報工学専攻 〒466-8555 名古屋市昭和区御器所町
E-mail: †{cincar-t,kawanami,manabu-k,sawatari,shikano}@is.naist.jp

あらまし 本稿では、「たけまるくん」の長期間運用に伴って構築してきた実環境音声データベースを用いて、音声案内システムの性能向上を時間軸でシミュレートする。収集期間と、データ数、出現している単語の数、評価データにおける言語モデルのパープレキシティーと未知語率、音声認識性能、応答正解率との関係を分析し、システム開発におけるそれぞれの構成要素の重要性に関して報告を行う。結果として、質問応答 DB を始め、音響モデル、言語モデルの順番でシステム更新が効果的であることが判った。また、一年間の収集データに基づいて構築したシステムにおける性能はほぼ飽和し、より多くの収集データを学習に用いても、上昇は見られなかった。更に、「たけまるくん」の収集データで構築した地下鉄の駅に設置した「キタちゃん」のポータビリティーを検討する。より現実的な開発状況を想定するため、新環境で収集した 20 日間のデータのみをシステムの適応に用いる。音声認識性能に関して、性能改善が比較的に低く、たけまるの音声認識部の頑健性は高い。応答正解率は 6 割程度であり、主に質問応答データベースからなる応答生成部の開発が今後の最も重要な課題である。
キーワード 実環境音声案内システム、開発状況、音響モデル、言語モデル、質問応答 DB、ポータビリティー

The Speech-oriented Guidance System Takemaru and its Portability

Tobias CINCAREK†, Hiromichi KAWANAMI†, Manabu KIDA†, Hiroshi SARUWATARI†,
Kiyohiro SHIKANO†, Ryuichi NISHIMURA††, and Akinobu LEE†††

† Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara-ken, 630-0192 Japan
†† Faculty of Systems Engineering, Wakayama University, 930 Sakaetani, Wakayama-shi, 640-8510 Japan
††† Dept. of CS, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, 466-8555 Japan
E-mail: †{cincar-t,kawanami,manabu-k,sawatari,shikano}@is.naist.jp

Abstract The development of the speech-oriented guidance system Takemaru is simulated using the data collected during the first two-years of regular operation. The relationship between the period of data collection, the number of speech inputs, number of words, language model perplexity, OOV rate, speech recognition performance and response accuracy is investigated. Based on this analysis, the relative importance of developing each system component can be assessed. Most important are efforts regarding the Q&A database followed by the acoustic and the language model. Moreover, a maximum in system performance seems to be reached after one year of data collection and overall improvement has the tendency to stagnate even if more data becomes available. Apart from that, the portability of the Takemaru system regarding a different environment, the Kitachan system installed at a local subway station, is investigated. Assuming a more realistic setting, only the data collected during twenty days is employed for system development. While the Takemaru ASR component shows a high portability, more efforts regarding the question and answer database have to be carried out in the future.
Key words Real Environment, Guidance System, Development Simulation, AM, LM, Q&A DB, Portability

1. まえがき

著者らは、2002年11月に、音声案内システム「たけまるくん」[1]を生駒市北コミュニティーセンターに設置し、4年に渡って運用してきた。サービスのドメインとして、市民のための施設の様々な案内、天気やニュースなどの一般的な情報の提供、子供が楽しめるCGエージェントとの生き生きとした会話がある。設置施設には図書館が併設してあるため、子供を始めとして訪問者が多く、案内システムは通常頻繁に利用されており、ユーザの自然発話音声データが大量に収集されている。これらについて最初の2年間に収集した音声データの書き起こしが終了している。また、運用開始から数ヵ月のデータと、頻度2回以上のユーザ質問に対する応答文の付与を行った[2]。本稿では、構築した音声、書き起こし、質問応答データからなるたけまるデータベースを用いて、開発のシミュレーション実験を行う。実験結果に基づいて、開発におけるシステムのそれぞれの構成要素の重要性を議論する。システム設計に欠かせない大規模な実環境データベースは、構築におけるコストが非常に高く、新たなシステムを開発する度に新規構築することは困難である。既存の案内システムを少量の収集データと人手によるわずかな作業のみで、別な環境に適応することが目標である。そのために、著者らは2006年3月に、たけまる案内システムの基本的な構成を拡張し、卓上型案内システム「キタちゃん」とロボット型案内システム「キタロボ」を設計した。完成後、近鉄の学研北生駒駅に設置した[3]。キタちゃんタスクに対して、2年のデータで構築したたけまるシステムのポータビリティーはどの程度であるか、1ヵ月のキタちゃん運用で収集したデータに基づいて検証する。

2. 音声案内システムの構成

2.1 たけまるくん

たけまる音声案内システムの構成を図1に示す。発話区間検出を行ってからまず音声か雑音に分類する。不要入力として棄却されなかった場合は、10ベストの認識仮説を出力し、音響尤度に基づいて年齢層を判定してから、応答生成を行う。応答生成は、年齢層別に、質問応答データベースから認識仮説に最も近い用例質問を見つけ出し、対応している応答文を応答としてユーザに返す。発話区間検出から認識仮説までの処理はJulius[4]で行われ、システム制御、応答生成、応答出力などは

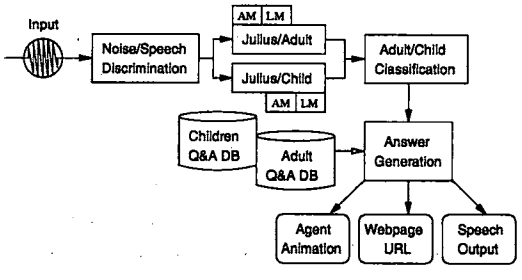

図1 たけまる音声案内システムの構成

表1 たけまる、キタちゃん、キタロボ（両キタ）の収集データ

分類	たけまるくん 入力数	たけまるくん 時間	両キタ 入力数	両キタ 時間
書き起こしあり	273,698	121.2	21,222	10.2
幼児	27,535	14.3	1,573	0.8
低学年子供	106,797	57.7	5,565	2.9
高学年子供	31,402	15.8	2,499	1.2
大人，高齢者	31,100	14.1	6,919	3.1
雑音など	76,864	19.3	4,716	2.2
書き起こしなし	388,249	174.1	35,164	14.5
合計収集データ	661,947	295.3	56,386	24.7

独自に開発したバックエンドとフリーソフトウェアに基づいている。最終的な目標は、ユーザが満足できる、応答正解率の高いシステムを構築することである。音声認識率と応答正解率の間には密接な関係があると予測される。音声認識性能を上げるために、音響モデル(AM)と言語モデル(LM)の開発が重要な課題である。また、ユーザの幅広い質問に十分対応できるように、充実した質問応答データベース(Q&A DB)を構築することが欠かせない。

2.2 キタちゃんとキタロボ

キタちゃんとキタロボの音声案内システムは、たけまるシステムに対して複数のドメインの拡張を行ったものである。キタロボは駅名に対して該当する路線を表示する。キタちゃんは、駅名に対しては路線探索を示し、音声ウェブナビ機能も持つ。更に、両システムは周辺案内を提供する。Juliusに基づく大語彙認識と並列に、ネットワーク文法で路線探索と周辺案内に関する発話の認識を行う。これらの新しいドメインは多対一の質問応答ペアの集合として見なすことができる。そこで、本稿では両システムの構成を統一した上で、たけまるのキタちゃんへのポータビリティーを調査する。

3. 音声と質問応答データベース

3.1 収集したデータ

たけまるくん、キタちゃんとキタロボで稼働開始から現在までに収集したデータを表1に示す。以後、キタちゃん、キタロボを示す場合、「両キタ」と総称する。たけまるくんは、最初の2年の間に、両キタは、最初の一ヵ月の間に、収集した各発話に対して話者ラベル、書き起こし、雑音タグの付与が人手によって行われた。

3.2 実験において使用するデータ

たけまるの長期間運用と音声認識部の開発に関する検討は、有効で背景会話と物音を含まない音声データに限定した（表2の前半を参照）。たけまるのキタちゃんへのポータビリティーの研究では、全有効発話を使用した（表2の後半を参照）。

応答生成のために使用する質問応答データベース(QADB)の内容を表3に示す。たけまるの評価は、全体の収集期間に対する頻度2回以上の全質問応答ペアと、積極的にシステム開発を行った稼働開始からの時期の一度しか出現していない用例をQADBの学習に用いる。両キタ初期QADBは、たけまる

表 2　音声認識部の開発のために使用したデータ

分類	収集期間	大人	子供
たけまる学習用	22ヵ月	16,332	75,315
たけまる評価用	1ヵ月	1,085	6,568
両キタ初期	24ヵ月	23,417	120,671
両キタ適応	20 日間	2,858	4,733
キタちゃん評価用	10 日間	631	565
キタロボ評価用	10 日間	1,130	1,446

表 3　質問応答データベース

質問応答ペア	用例質問文数				応答
データベース	大人	子供	頻度 1 回	2 回以上	文数
たけまる学習用	2,891	9,237	5,933	5,671	322
両キタ初期	2,974	5,553	-	-	274
両キタ追加	1,729	2,830	3,562	659	75
両キタ学習用	4,255	7,625	-	-	349

QADB の一部をキタちゃんの環境に合わせた修正することで作成されていた。適応のために、20 日間の収集データを追加した。

4. たけまるの定期的更新と長期間運用のシミュレーション

4.1 実験条件

2002 年 12 月からそれぞれの月末までの収集データ (2003 年 8 月の評価データは除く) を用いて、一ヵ月単位で音響モデルと言語モデルを構築する。質問応答データベースは、該当期間の応答正解ラベルが付与されているデータを使用する。それ以外の実験条件は、表 4 に示す。

表 4　実験条件

学習ソフトウェア	AM: HTK 3.2 [5], LM: SRILM 1.5.0 [6]
初期音響モデル	2,000 状態, PTM, 8,256 ガウス分布
音響特徴量	12 MFCC, 12 Δ MFCC, Δ E
音響モデル学習	Baum-Welch, 3 回
音響モデル適応	MLLR-MAP, 3 回, 回帰クラス数 : 256
言語モデル	3-gram, Kneser-Ney 平滑化
認識エンジン	Julius 3.5

4.2 音声認識部と認識精度

図 3〜図 9 に音声案内システムの開発シミュレーションの結果を示す。全グラフの横軸は、学習に使用したデータの月数を示す。実際に各期間において使用した発話数は図 2 に示す。まず、音声認識部に関して評価を行い、結果を解説する。

収集したデータで出現している異なり単語 (形態素) の数は、2 年経過後も増加する傾向にある (図 3 参照)。一ヵ月の評価データに対する未知語率は、最初の一年で大幅に減少するが、2 年後でも 1 %以上である (図 4 参照)。言語モデルのパープレキシティーは 8 ヵ月後に既に下限を達しているようであるが (図 5 参照)、未知語率が高いことと、トライグラムの確率の推定に必要な学習データが少ないことから、ウェブ収集データを用いた言語モデルの補間が欠かせない。

以下、音声認識性能が音響モデルと言語モデルの一ヵ月単位での更新によってどの程度に向上するかを検討する。図 6 (大人) と図 7 (子供) は、音響モデルを固定にした場合 (初期、中間、最終モデルの)、言語モデル更新による性能変化を示す。最初の 5 ヵ月の間に認識精度が著しく向上するが、6 ヵ月以上の書き起こしデータを学習に用いても、性能はほとんど改善せず、飽和する傾向にある。音響モデルの学習期間が変わっても、同じ結果が得られている。統計的な n-gram 言語モデルの学習は信頼できる確率推定において常に数百万〜数千万文の学習テキストを必要とするので、最終的な判断を下すには根拠が不十分ではあるが、ウェブ収集データによる補間を行っても、数年か数ヵ月の収集音声の書き起こしの効果は恐らく同程度である。

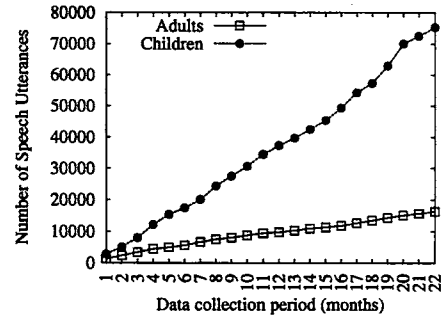

図 2　使用発話数 (たけまる)

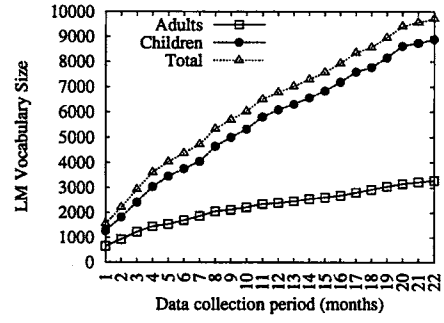

図 3　言語モデルの語彙数 (たけまる)

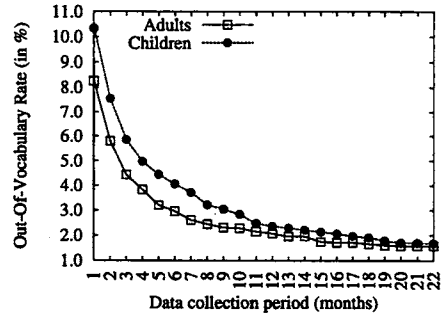

図 4　言語モデルの未知語率 (たけまる)

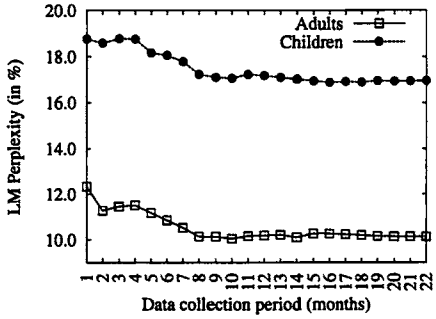

図 5　評価データにおける言語モデルのパープレキシティー

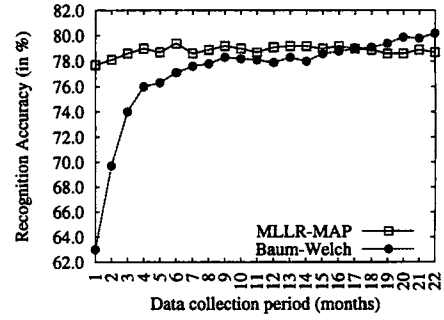

図 8　大人の音響モデル、認識精度（たけまる）

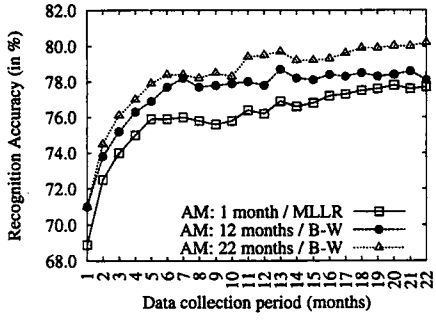

図 6　大人の言語モデル、認識精度（たけまる）

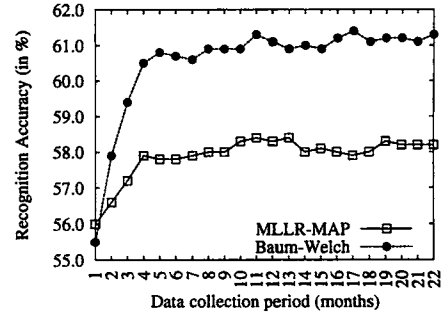

図 9　子供の音響モデル、認識精度（たけまる）

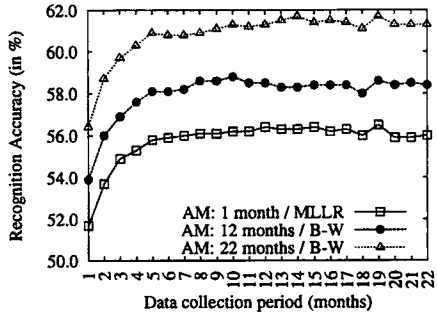

図 7　子供の言語モデル、認識精度（たけまる）

続いて、言語モデルを固定にした場合の音響モデル更新の効果を図 8（大人）と図 9（子供）に示す。言語モデルは固定であり、22ヵ月の書き起こしデータで学習したものである。大人の音声データが比較的少ないので、18ヵ月以上のデータがなければ、MLLR-MAP 適応の方が Baum-Welch 学習より効果的である。また、初期モデルは大人の音声に基づいて構築したものなので、再学習でも適応でも性能は殆んど変わらない。同じ理由から、子供音響モデルの認識性能は最初の数ヵ月の更新以降、学習データが少なくても Baum-Welch 学習によって認識精度の方が向上する。しかし、半年の収集データで既に性能の頂点が達成されているように見える。音響モデルのパラメータ数が割と少ないことも、飽和の理由の一つだと考えられる。

4.3　応答生成部と応答正解率

アプリケーションはディクテーションではなく、ユーザに負担の少ない自然な音声インタフェースを介した案内なので、システムの評価値としては、音声認識精度より、応答正解率の方が適切である。たけまるの応答生成は質問応答ペアに基づいている。現時点では、全てのユーザ発声に対する正解応答文の付与がまだ完了していないので、一部のデータで予備的な評価を行う。まず、データベースの更新に使用した用例質問文と応答文をそれぞれ図 10 と図 11 に示す。用例質問文数の増加が最初の 6 ヵ月と最後の一ヵ月の間に顕著である。これは、該当期間については表記上一度しか出現していない質問文もデータベースに加えたからである。それ以外の期間は、2 回以上出現した質問文しか追加していない。応答文の数に関しても同じ傾向が見られる。

応答正解率の評価結果は図 12 と図 13 に示されている。黒いダイアモンド曲線は、全構成要素を期間毎に更新した場合である。その他の三つの曲線は、それぞれ音響モデル（AM）、言語モデル（LM）、質問応答データベース（QADB）の更新による性能変化を示すが、他の構成要素については既に全学習データで更新した状態のものを用いている。この評価結果から、各構成要素の学習において必要な収集データ量と、開発の重要性が判る。データ量に関して、音声案内の音声認識と応答正解の性能は 1 年の収集データを使用した場合に、既に頂点を達しているように見える。言語モデルの更新は、未知語率が高く学習データが少ないにもかかわらず、応答正解率に対する影響が比

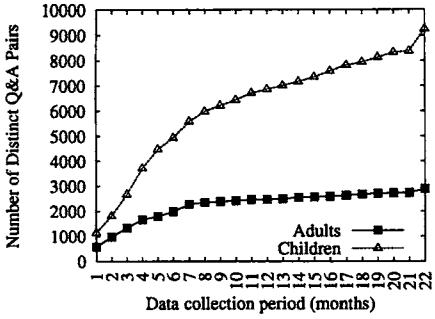

図 10　用例質問文の種類（たけまる）

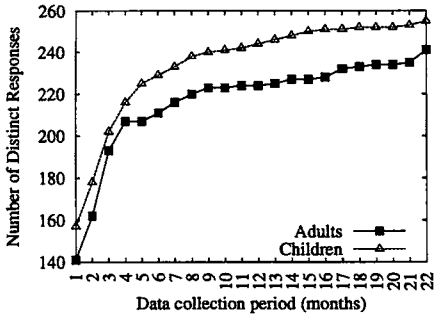

図 11　応答文の種類（たけまる）

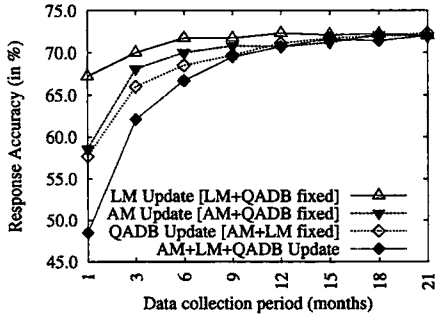

図 12　大人の応答正解率（たけまる）

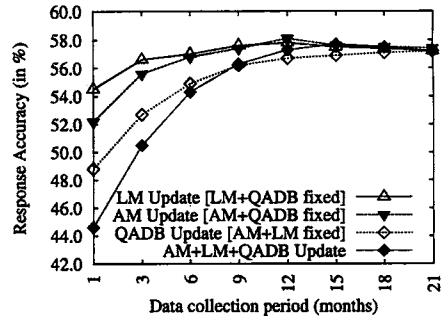

図 13　子供の応答正解率（たけまる）

較的に小さい。一ヵ月の学習データだけで言語モデルを構築したとしても、応答正解率の上限からの下落は最大５％であることに対して、音響モデルは１ヵ月のみでの学習では５～10％程度、質問応答データベースの１ヵ月のみの学習では10～15％も性能が劣ってしまう。

5.　たけまるのキタちゃんへのポータビリティー

前節の結果から、音声案内システムの開発コストは非常に高く、それなりの性能を達するシステムを構築するのには半年～１年程度の収集データが必要であると判る。たけまるの設置環境は市民センターである。異なる環境に音声案内システムを設置するために、再び長期間に渡ってデータ収集を行うことは非現実的である。音声案内システムの実用化にあたり、既存システムや学習データを活用する上では、少量の収集データに基づく適応が必然である。ここでは、たけまる音声案内システムを基盤にして、地下鉄の駅に設置したキタちゃんとキタロボの案内システムの適応を検討する。たけまるシステムを基本的にそのままに利用する場合と、短期間の収集データで構成要素を更新した場合の性能評価を行う。以下、各要素のポータビリティーを検証する。

5.1　音声認識部と認識精度

たけまるのベースラインシステムは収集データ２年分で学習したものである。新たにキタちゃんとキタロボで収集した20日間のデータを音響モデルと言語モデルの更新に用いる。音響モデルはMLLR-MAPで適応した。言語モデルは、収集したデータを既存のたけまる学習データに追加し、言語モデルの再構築を行った。音声認識実験の結果を表５に示す。たけまるの大人、子供の音響、言語モデルを同時に更新したとしても、認識精度の改善は１～２％程度である。言語モデルの更新は音響モデルの更新より僅かに効果的ではある。この結果から、たけまるの音響モデルのポータビリティーは非常に高いことが判る。言語モデルは、キタちゃんの収集データを加えたとしても、パープレキシティーと未知語率がかなり高い水準にあり（表６参照）、たけまるの２倍程度である。従って、他の収集源からキタちゃんのドメインに近い学習データを収集することが言語モデルの増強に欠かせない。

表５　音声認識精度（読み、％）

評価データ	キタちゃん		キタロボ	
年齢層	大人	子供	大人	子供
たけまる	68.9	54.4	76.6	59.3
AM 更新	69.8	54.2	77.7	60.0
LM 更新	70.3	54.8	77.4	60.0
両更新	71.5	54.6	78.0	60.7

表６　言語モデルのパープレキシティーと未知語率

評価データ	キタちゃん				キタロボ			
年齢層	大人		子供		大人		子供	
LM の評価	OOV	PP	OOV	PP	OOV	PP	OOV	PP
たけまる	3.0%	27	3.0%	39	2.6%	17	1.6%	23
キタ適応	1.9%	25	2.7%	38	2.0%	16	1.2%	22

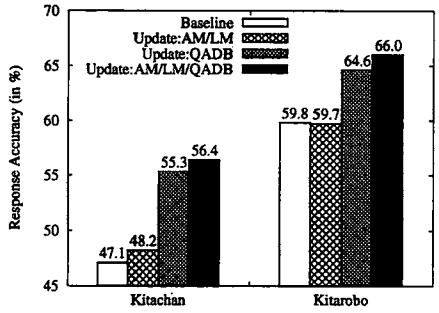

図 14　大人の応答正解率（キタちゃんとキタロボ）

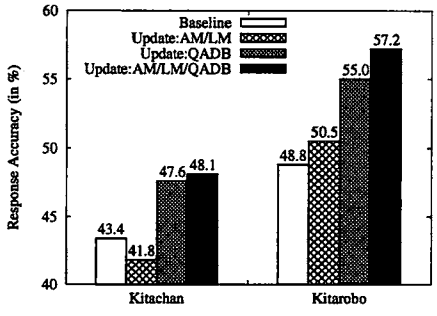

図 15　子供の応答正解率（キタちゃんとキタロボ）

5.2　応答生成部と応答正解率

最後に、予備的な評価において、キタちゃんとキタロボの応答正解率を報告する。図 14 と図 15 はそれぞれ大人と子供の応答正解率を示す。設置環境が異なっているため、最低限の質問応答データベースの人手による修正以外、たけまるシステムをそのまま駅の環境で使用すると、応答正解率がたけまるに比べ比較的に低い。音声認識部の適応を行っても、性能が殆んど改善されない。20 日間の収集データでの質問応答正解データベースの更新が一番効果的である。同時に音声認識部のモデルも更新すると、場合によって若干の性能向上が見られる。システムの全構成要素の更新によって、応答正解率は 5 ～ 8 %上昇する。

ここまでの評価はオープンである。評価データの質問応答ペアも学習データとして質問応答データベースに追加した場合の傾向を図 7 に示す。オープン評価に比べ、大人も子供も応答正解率は大幅に向上し、ほぼ 20 %にも達している。これは、学習と評価期間の間、ユーザ発話内容に大きな差があることが判る。従って、これからの主な課題は少量の収集データから、ユーザの要求を予測し、予め他源（たとえばウェブ）のデータ収集によって充実した質問応答データベースを構築することである。たけまるの長期間運用に関しても、システム開発において質問応答データベースが最も重要である。

6.　まとめと今後の課題

本稿では、二つの音声案内システムの開発状況を検討した。たけまるの 2 年間の実環境音声データベースに基づいて、長期間運用シミュレーションを行った。システム開発において半年～ 1 年の収集データしか用いなくても、認識精度と応答正解率が既に飽和する。音声案内システムの構成要素の内で、開発が最も重要である順番は、質問応答データベース、音響モデル、言語モデルであるという結果が得られた。言語モデルの学習に用いた書き起こしデータの量は信頼性のあるパラメータ推定において不十分であるため、ウェブ収集データで補ってから再び検討を要する。市民センターのために開発した音声案内システムをそのまま駅の環境に設置した場合、音響モデルに関して、高い頑健性が見られた。しかしながら、評価データに対する未知語率とパープレキシティーが比較的に高いことから、他収集源を利用して言語モデルの学習データを補う必要がある。また、応答正解率は、20 日間の収集データを適応に用いた場合、5 ～ 6 割である。たけまるベースラインシステムに比べ 6 %程度に上昇する。評価データの質問応答ペアを全部応答生成部に追加した場合との性能差は 15 %程度である。従って、ポータビリティーにおいて、ユーザが知りたい情報の予測とそれに合った質問応答データベースを少量の収集質問データで構築することが最も重要な今後の課題である。

表 7　システム稼働開始、収集データで更新した後、評価データの質問応答ペアも追加した質問応答データベースを使用した場合のキタちゃんとキタロボの応答正解率（%）

質問応答	キタちゃん		キタロボ	
データベース	大人	子供	大人	子供
両キタ初期 (open)	47.7	41.8	59.7	50.5
両キタ更新 (open)	56.4	48.1	66.0	57.2
両キタ更新 (closed)	76.5	62.3	82.7	68.4

文　献

[1] 西村 竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: "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", 電子情報通信学会論文誌, Vol.J87-D-II, No.3, pp.789-798, 2004.

[2] 木田 学, 川波 弘道, 猿渡 洋, 鹿野 清宏: "音声情報案内システムにおける質問応答データベース最適化手法の検討", 情報処理学会研究報告, 2006-SLP-62-15, pp. 81–86, July 2006.

[3] 川波弘道, 木田学, 早川直樹, トビアス ツィンツァレク, 北村任宏, 加藤智之, 鹿野清宏, "駅構内音声案内システム「キタちゃん」「キタちゃんロボット」の開発", 電子情報通信学会技術研究報告, Vol.106, No.123, SP2006-14, pp.19-24, June 2006.

[4] Julius, an Open-Source Large Vocabulary CSR Engine - http://julius.sourceforge.jp/.

[5] HTK Speech Recognition Toolkit http://htk.eng.cam.ac.uk/.

[6] Andreas Stolcke, SRILM - An Extensible Language Modeling Toolkit, Proc. of ICSLP, pp. 901-904, 2002.