Vol.2009-SLP-79 No.14
2009/12/22

音声対話システムの対話履歴N-gramを利用したユーザ満足度推定手法

原　直[†1]　北岡教英[†1]　武田一哉[†1]

本論文では，音声対話システムの対話履歴を N-gram によってモデル化を行い、その尤度に基づいてユーザ満足度の推定を行う手法を提案する．実験では音声対話による楽曲検索システムを利用した実環境音声データに収録されているユーザとシステムの対話履歴を利用して各ユーザの満足度を N-gram の尤度を用いて推定する．それぞれの満足度レベル毎の N-gram モデルは，ユーザの対話行動とシステムの対話行動を符号化した系列を利用することで作成される．本論文では，作成したモデルを用いて満足度の推定を行いその分類性能を評価した．実験結果より，提案手法は高い分類性能を示しており，特にタスク達成者とタスク未達成者の分類では全てのタスク未達成者を検出する際の誤検出率を 10%に抑えることができた．

Estimation method of user satisfaction based on dialog history N-gram for spoken dialog system

Sunao Hara,[†1] Norihide Kitaoka[†1]
and Kazuya Takeda[†1]

In this paper, we propose an estimation method of user satisfaction for a spoken dialog system using an N-gram likelihood based on dialog history model. For the experiment, we use spoken dialog histories between users and the system,which is contained in an real environmental speech data that was collected by a music retrieval system with a spoken dailog interface, and we esimate their satisfacion level based on the N-gram likelihood. An N-gram model is trained from encoded sequences that consist of users' dialog acts and the system's dialog acts for each user satisfaction level. Experimental results show that our proposed method achieved good classification performance, especially the false detection rate of 10% in the experiment of a classification into dialogs with task completion and those without task completion.

1. はじめに

性能推定は音声対話システムの設計における重要な課題である．設計時のコストの観点から，主観的な評価よりは客観的な指標に基づいた自動性能評価が好まれる．さらに，性能のオンライン推定を実現することで，性能に悪影響を与える対話が起こりにくくなるようなシステム対話戦略の自動選択も可能になる．

コールセンターに配備される自動音声応答 (IVR) システムにおける異常対話の検出に着目した研究はこれまでに多く報告されている．Walker ら[6] は性能の指標として “音声言語理解の成功” に着目し，成功した対話を正常、失敗した場合を異常とした上で、自動的に抽出された複数の特徴量を利用した異常対話の推論手法を提案している．Kim[4] はオンライン推定手法を行うための N-gram に基づいた応答対話品質の監視システムを提案している．このシステムではユーザの発話系列を 5-gram によりモデル化を行い，最終的に IVR システムへの最初から 5 発話だけを利用するだけで 83%の精度で異常対話を検出できると報告されている．また，Herm ら[2] によってシステムログと感情認識の結果を統合したモデルが提案されており，最初から 4 発話だけを利用することで異常対話と正常対話の識別について 79%の識別精度が得られたという報告もなされている．

本研究の目標は音声対話システム利用時のユーザ満足度推定モデルの構築である．ユーザの視点から考えると，音声対話システムとの対話においてユーザが観測できるのはシステムからの指示や応答などのシステムの出力結果だけであり，システムの内部状態は観測できない．従って，システムの出力結果はユーザの感情，すなわちユーザの満足度に強く関係していると考えられる．本研究では，N-gram による対話履歴モデルを利用した音声対話システムのユーザ満足度推定モデルを提案する．ユーザとシステムの対話行動を対話意図を考慮したシンボル系列へと変換し，それぞれのユーザ満足度毎にシンボル系列を N-gram によってモデル化を行う．そして、入力対話シンボル系列に対して尤度などを規準として満足度などを推定する。

本論文は以下のように構成されている．第 2 節では，音声対話コーパスのフィールドテストとデータ収集について概略を述べる．第 3 節では，対話データと N-gram モデルの定式化を示す．第 4 節では，対話コーパスから N-gram モデルの構築を行い，実際の対話行動

†1 名古屋大学 情報科学研究科
Graduate School of Information Science, Nagoya University

系列からユーザ満足度の推定と評価を行う．最後に，第 5 節で結論を述べる．

2. 音声対話コーパス

2.1 *MusicNavi2*: 音声対話による楽曲検索システム

音声対話によってユーザの PC に存在する音楽ファイルを検索し再生することができる，楽曲検索システム *MusicNavi2*[1] のフィールドテストによってデータ収集が行われた．このシステムはインターネット上のサーバプログラムを介して，ユーザの PC にダウンロード・インストールされる．*MusicNavi2* には入力音声やシステム動作ログをユーザの識別情報とともにサーバに送信する機能が備えられているため，大量の音声データを自動的に集めることができる．音声インタフェースは文法駆動の音声認識エンジンを利用して実装されており，音声認識辞書としてはいくつかの制御コマンド用の単語とユーザ PC に保存されている音楽ファイルの楽曲名，アーティスト名およびアルバム名だけに制限された辞書が自動的に作成されている．音声認識エンジンには Julius[3] を利用した．

フィールドテストに際して被験者らには，システムを利用して 20 回以上の発話を行うか 40 分以上の利用によって，5 曲以上を再生する，という課題を与えている．その後，インターネット上のアンケートによってシステムへの感想やユーザの属性に関する質問に回答している．アンケートでは，音声対話システムに対するユーザの満足度を 5 段階評価（5:満足，4:どちらかといえば満足，3:どちらともいえない，2:どちらかといえば不満，1:不満）によって選択している．

これらの実験データは実環境における主観的なユーザビリティ評価を含んだ大規模な音声対話を含んだ *MusicNavi2* データベースとして整備されている．この実験には総計 1369 名のユーザが参加しており，その合計利用時間は約 488 時間である．

2.2 *Musicnavi2* データベースの予備評価

実験データとして，まずはタスクを達成していた被験者 449 名のデータを利用する．彼らはアンケートで回答した満足度に従って 1, 2, 3, 4, 5 のラベルがつけられている．さらに，対話の失敗によってタスクを達成できなかった被験者 69 名に ϕ というラベルをつけて実験に利用する．従って，表 1 に示したように，6 クラスに分類される合計 518 名の被験者を実験に利用した．

2.3 ユーザ・システム発話から対話行動への変換

ユーザの発話とシステムの発話はそれぞれ 19 のユーザ行動シンボルと 22 のシステム行動シンボルに変換される．本研究ではユーザ・システム対話行動を自動収集可能な特徴によって定義する．従って，手作業の含まれる書き起こしデータではなく音声認識結果を利用して，ユーザ行動シンボルへの変換を行った．ユーザ行動シンボルは *MusicNavi2* の音声認識辞書の認識単語にあらかじめタグとして付与してあるため，容易に認識結果からの変換が行える．システム行動シンボルも同様にシステムからの指示や応答内容を示すものであるので、自動的に取得できるように実装されている．

図 1 に対話例と各発話に対応する行動シンボルを示す．

表 1 満足度に従って 6 クラスに分類された被験者の人数
Table 1 Population of the subjects classified into 6 classes according to their satisfaction level.

Class	ϕ	1	2	3	4	5
# of subjects	69	38	102	107	155	47

3. 対話行動系列の N-gram モデル

対話行動系列は各ユーザ毎にシステム行動系列とユーザ行動系列を時間順に整列することで作られる．対話行動系列 $\mathbf{x}$ は次式によって示される，

$$\mathbf{x} = \{x_1, \ldots, x_t, \ldots, x_T\} \tag{1}$$

ここで，t は対話ターンの番号である．

本研究では，ユーザの満足度は現在の対話行動だけではなく，数ターン前からの対話行動系列によっても影響されていると仮定する．そこで，対話行動系列 $\mathbf{x}$ を満足度 s 毎の N-gram モデル $\mathcal{M}_s$ によってモデル化を行う．

$$\mathcal{M} = \{\mathcal{M}_s; s = \phi, 1, 2, 3, 4, 5\} \tag{2}$$

ここで，ϕ は前節で述べた課題未達成のユーザを意味する満足度である．N-gram モデルは SRILM[5] によって Witten-Bell ディスカウント手法を適用して作成した．

4. ユーザ満足度の推定実験

各ユーザの対話行動系列から満足度を推定する実験によって提案モデルの評価を行う．518 名の被験者データを leave-one-out 交差検定によって評価した．すなわち 1 名の対話行動系列を評価に用いて，残り 517 名のデータを学習用に用いる，という試験を 518 名分行った．本研究ではユーザ満足度毎に 1-gram から 8-gram までのモデルを作成して評価した．

また利用する対話行動系列として，システムの対話行動系列のみを用いた場合（SYS），ユーザの対話行動系列のみを用いた場合（USR），両方を用いた場合（SYSUSR）の 3 条件についての比較も行った．

	システムの応答文 / ユーザの発話（認識）文	対話行動シンボル
USR:	「こんにちは」	USR_CMD_HELLO
SYS:	『こんにちは』	SYS_INFO_GREETING
USR:	「サイモンアンドガーファンクル」	USR_REQUEST_BYARTIST
SYS:	『“Simon and Garfunkel” の曲を検索しますか？』	SYS_CONFIRM_KEYWORD
USR:	「はい」	USR_CMD_YES
SYS:	『“Simon and Garfunkel” の曲を検索します』	SYS_INFO_SEARCHBYARTIST
SYS:	『60 曲見つかりました』	SYS_INFO_SEARCHSUCCESS
SYS:	『“I am a rock”』	SYS_INFO_SONGTITLE
SYS:	『“明日に架ける橋”』	SYS_INFO_SONGTITLE
USR:	「その曲」	USR_CMD_THATSONG
SYS:	『“Simon and Garfunkel” の “明日に架ける橋” を再生します』	SYS_PLAY_SONG
USR:	「停止」	USR_CMD_STOP
SYS:	『停止します』	SYS_INFO_STOPPED
:	:	:

図 1 対話例と対応するシンボル系列の例
Fig. 1 Example of a dialog and its corresponding encoded symbols.

表 2 SYS 条件の 3-gram モデルによる満足度推定の混同表
Table 2 Confusion matrix of the user satisfaction estimation result by 3-gram model of SYS condition.

		Estimated $\hat{\mathcal{M}}$					
		ϕ	1	2	3	4	5
Actual $\mathcal{M}$	ϕ	43	5	7	5	6	3
	1	0	7	8	9	11	3
	2	1	8	31	16	35	11
	3	0	9	22	23	45	8
	4	0	8	34	29	66	18
	5	0	4	5	6	24	8

4.1 6 クラスの満足度識別実験

一様な事前分布 $P(\mathcal{M})$ を仮定して，6 クラスの分類問題に対して次式に示す最尤推定を適用した，

$$\hat{\mathcal{M}} = \underset{\mathcal{M}_s}{\operatorname{argmax}} \{P(\mathbf{x}|\mathcal{M}_s)\} \tag{3}$$

ここで， $\mathbf{x}$ は入力系列，$\hat{\mathcal{M}}$ は識別結果である．

図 2 の左図に識別精度を示す．SYS 条件の 3-gram モデルが他のモデルに比べて最も高い精度である 34.4%を達成している．この 3-gram モデルによる識別結果についての混同表を表 2 に示す．この表から，完全に一致した推定結果を得るということは困難だが，大きく外れた推定が行われることは少ないことがわかる．特に課題未達成ユーザ（$s = \phi$）と課題達成ユーザ（$s \neq \phi$）の識別は適切に行われていることが読み取れる．そこで，ある満足度の課題達成ユーザを異なった満足度の課題達成ユーザと混同した場合でも正解と見なして再集計した結果を図 2 の右図に示す．図より，3-gram モデルが最も高い精度である 94.7%を達成していることがわかる．

また，全体的にユーザの発話だけを使った場合 (USR) やシステムとユーザの発話の両方を

Vol.2009-SLP-79 No.14
2009/12/22

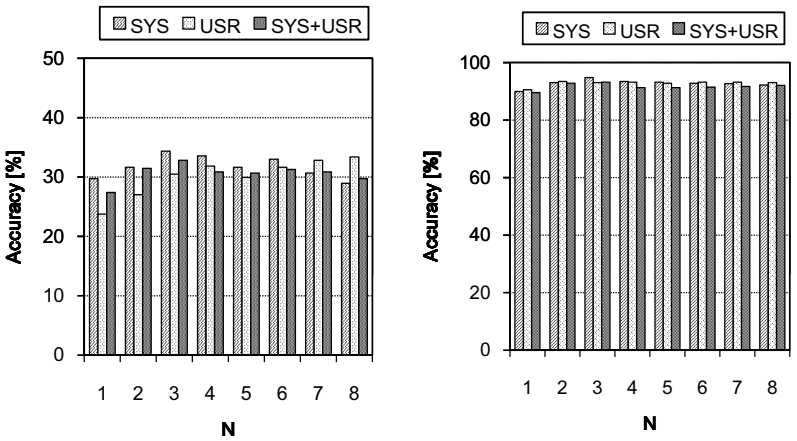


図 2 6 クラス (ϕ, 1, 2, 3, 4, 5) の分類精度（左）と 2 クラス (ϕ and the others) とみなして再計算した分類精度（右）
Fig. 2 Classification accuracies of the 6-class(ϕ, 1, 2, 3, 4, 5) experiment(left) and recalculated accuracies as 2-class(ϕ and the others) experiment(right)

利用した場合 (SYSUSR) の結果はシステムの発話だけを使った場合 (SYS) よりも性能が落ちている．この原因として，ユーザの発話に認識誤りが含まれている音声認識結果を利用していることから，モデルの精度が落ちていると考えられる．

4.2 2 クラス分類問題としてのユーザ満足度推定実験

識別性能をより詳細に調査するために，2 クラス識別手法による検討を行う．本論文では，次の 2 種類の識別器の性能に着目する．a) 課題達成ユーザ（$s \neq \phi$）と課題未達成ユーザ（$s = \phi$）の識別，b) 満足したユーザ（$s = 5$）と満足していないユーザ（$s = 1$）の識別，である．すなわち，テストユーザが “課題未達成” もしくは “満足していない” クラスに属するかどうかを検出する識別器を構築する．

次式の事後オッズ識別器を用いる．

$$\hat{\mathcal{M}} = \begin{cases} \mathcal{M}_j & \text{if } \frac{P(\mathcal{M}_j|\mathbf{x})}{P(\mathcal{M}_i|\mathbf{x})} > 1, \\ \mathcal{M}_i & \text{otherwise}. \end{cases} \tag{4}$$

(4) にベイズ則を適用すると，

$$\frac{P(\mathcal{M}_j|\mathbf{x})}{P(\mathcal{M}_i|\mathbf{x})} = \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_i)} \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} = \frac{1}{\alpha} \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} \tag{5}$$

表 3 タスク未達成ユーザ検出結果の ROC 曲線下面積 (AUC)
Table 3 Area Under the Curve(AUC) of the detection of task incomplete users

	SYS	USR	SYSUSR
1-gram	0.901	0.873	0.927
2-gram	0.948	0.929	0.977
3-gram	0.989	0.954	0.993
4-gram	0.995	0.952	0.997
5-gram	0.993	0.954	0.995
6-gram	0.989	0.951	0.995
7-gram	0.988	0.946	0.995
8-gram	0.987	0.936	0.994

ここで，α は事前オッズの逆数である．最終的に，次の識別器が得られる，

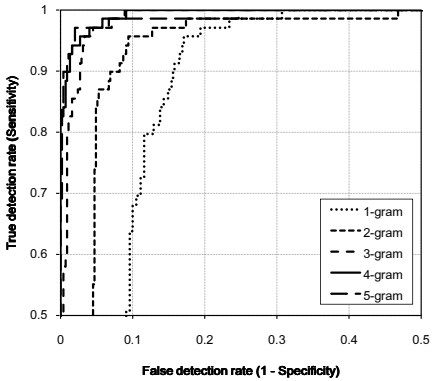
$$\hat{\mathcal{M}} = \begin{cases} \mathcal{M}_j & \text{if } \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} > \alpha, \\ \mathcal{M}_i & \text{otherwise}. \end{cases} \tag{6}$$

パラメータ α を変化させて，ROC(Receiver Operating Characteristic) 曲線の表示を行い，ROC 曲線下面積 (Area Under the Curve; AUC) によって性能の評価を行う．

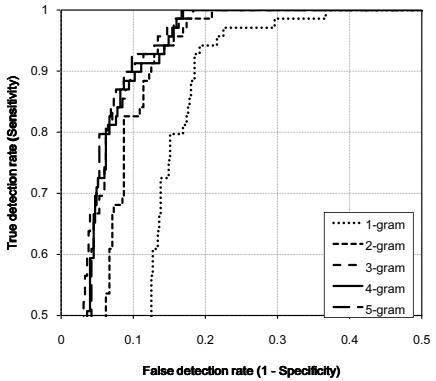
ユーザがタスクが達成できたか否かの識別結果を図 3 に示す．図より，タスク達成の識別能力としては高い性能を示していることがわかる．表 3 に示した ROC 曲線下面積によると，SYSUSR 条件における 4-gram モデルが 0.997 と最も高い性能を示している．また，3 条件の比較をすると 6 クラスの識別実験結果と同様に，ユーザの対話行動系列だけを利用した場合の識別性能の低下が見られる．

一方で，ユーザが満足していないか満足しているかの識別結果を図 4 に示す．図より，まだ十分とは言えないが、SYSUSR 条件の 2-gram モデルでは検出率 80%に対して誤検出率 28%と、ある程度の識別ができていることがわかる。表 4 に示した ROC 曲線下面積の値からも，現在より前の対話系列を用いない 1-gram モデルよりも対話系列をモデル化した 2-gram モデルの方がより高い性能を示していることがわかる．これらより、対話の履歴が満足度に対して影響を及ぼしていることが示唆される。また、システムの対話系列だけ（SYS 条件）やユーザの対話系列だけ（USR 条件）を用いるよりも、両者を用いた場合（SYSUSR 条件）の性能が高いことから、システムとユーザのやりとりが満足度に影響することが示唆される。

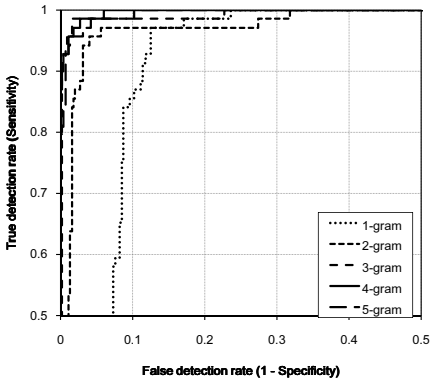
Vol.2009-SLP-79 No.14
2009/12/22



(a) SYS condition

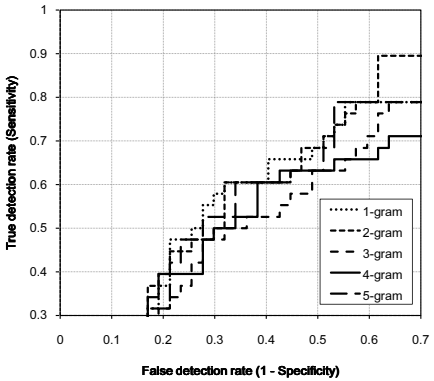


(b) USR condition

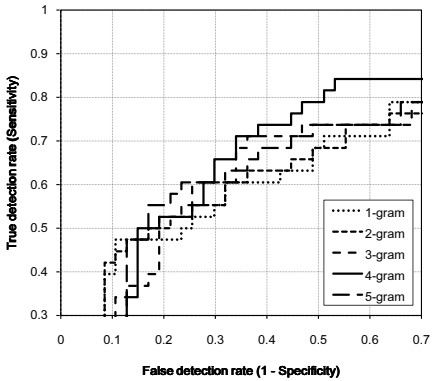


(c) SYSUSR condition

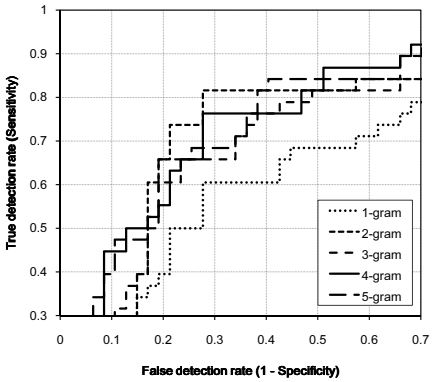
図 3 タスク達成ユーザ ($s \neq \phi$) とタスク未達成ユーザ ($s = \phi$) の識別結果に対する ROC 曲線
Fig. 3 ROC curve for the test classified into whether the user completed the task ($s \neq \phi$) or not ($s = \phi$).



(a) SYS condition



(b) USR condition



(c) SYSUSR condition

図 4 満足ユーザ ($s = 5$) と不満足ユーザ ($s = 1$) の識別結果に対する ROC 曲線
Fig. 4 ROC curve for the test classified into whether the user satisfied($s = 5$) or not ($s = 1$).

表 4 不満足ユーザ検出結果の ROC 曲線下面積 (AUC)
Table 4 Area Under the Curve(AUC) of the detection of unsatisfied users

	SYS	USR	SYSUSR
1-gram	0.611	0.638	0.619
2-gram	0.628	0.644	0.724
3-gram	0.591	0.651	0.704
4-gram	0.583	0.681	0.739
5-gram	0.629	0.662	0.739
6-gram	0.632	0.639	0.761
7-gram	0.604	0.633	0.765
8-gram	0.592	0.622	0.756

5. まとめと今後の課題

N-gram モデルを利用して，音声対話による楽曲検索システムのフィールドテストによって得られた音声対話システムのユーザ満足度を推定する手法について検討した．本研究では，N-gram によってモデル化されたユーザとシステムの対話行動系列を利用した推定手法を提案した．実験結果より，システムの対話行動系列をモデル化することで，ユーザがタスクを達成できたか否かの識別で高い識別性能が得られた．また，ユーザの満足度を推定する場合には，システムとユーザ両者の対話行動系列を利用することで高い性能が得られる事が示唆された．

実験により提案モデルの有効性は示されたが，いくつかの課題が残されている．満足度の推定性能を向上させるためにはさらなる検討が必要である。N-gram モデルの分析を行うことでユーザ満足度に影響を及ぼす重要な対話行動およびその前後関係の調査や，単語誤り率と推定性能の関係の検討も今後の課題として挙げられる．

謝辞 本研究の一部は総務省戦略的情報通信研究開発推進制度 (SCOPE) によるものである．

参 考 文 献

1) Hara, S., Miyajima, C., Itou, K., Kitaoka, N. and Takeda, K.: Data collection and usability study of a PC-based speech application in various user environments, *Proceedings of Oriental COCOSDA 2008*, pp.39–44 (2008).
2) Herm, O., Shmitt, A. and Liscombe, J.: When calls go wrong: How to Detect Problematic Calss Based on Log-Files and Emotions?, *Proceedings of INTERSPEECH 2008*, pp.463–466 (2008).
3) Kawahara, T., Lee, A., Takeda, K., Itou, K. and Shikano, K.: Recent progress of open-source LVCSR engine Julius and Japanese model repository; Software of continuous speech recognition consortium, *Proceedings of ICSLP 2004*, Vol.2, pp. 3069–3072 (2004).
4) Kim, W.: Online call quality monitoring for automating agent-based call centers, *Proceedings of INTERSPEECH 2007*, pp.130–133 (2007).
5) Stolcke, A.: SRILM – An extensible language modeling toolkit, *Proceedings of ICSLP 2002*, pp.901–904 (2002).
6) Walker, M.A., Langkilde-Geary, I., Hastie, H.W., Wright, J. and Gorin, A.: Automatically Training a Problematic Dialogue Predictor for a Spoken Dialogue System, *Journal of Artificial Intelligence Research*, Vol.16, pp.293–319 (2002).