音声対話コーパスに基づく あいづち生成タイミングの検出とその評価

神谷　優貴[†]　　大野　誠寛[‡]　　松原　茂樹[†]

[†]名古屋大学大学院情報科学研究科　　[‡]名古屋大学大学院国際開発研究科

kamiya@el.itc.nagoya-u.ac.jp,{ohno,matubara}@nagoya-u.jp

1　はじめに

近年，音声対話システムの実用化が進みつつある．なかでも車内音声対話システムは，最も普及している音声アプリケーションの一つであり，その多くは，ナビゲーションや情報検索を主要なタスクとしている．これまで，タスクを確実に遂行できるシステムの実現を目的に，音声認識や発話理解，対話制御などの研究開発が進められ，車内音声対話の技術は大いに向上している．今後は，単にタスクを達成できるだけでなく，ドライバがより快適に対話を進められるように，システムの応答性を高めることが重要となる．

システムの応答性を高めるための一つの方法は，ユーザの発話中においてもシステムが何らかの反応を返すことにより，システムによる認識や理解の状態を適宜開示することである．このような開示は，人間同士の対話の場合，頷きや表情，身振りや手ぶり，あいづちなどの行為を通して遂行される．しかし，実走行車内においては，ドライバの視線は音声対話システムにないため，システムによるドライバへの応答としては，音声による応答，すなわち，あいづちによらざるを得ない．また，ドライバにとって快適な車内対話を実現するためのシステムの応答戦略としては，ドライバの発話中に積極的にあいづちを打つことが望まれる一方，適切なタイミングであいづちが生成される必要がある．

そこで本論文では，高い応答性を備えた車内音声対話システムの実現を目指し，対話コーパスを用いたあいづち生成タイミングの検出とその評価について述べる．あいづちの発生タイミングに関する研究は既にいくつか存在しており [2, 7, 10, 11, 13]，人間の間で遂行された対話中のあいづち発生タイミングに基づいて，あいづち位置の分析や推定が行われている．しかし，これらの研究で利用されてきた現存する音声対話コーパスは，あいづち発生タイミングの揺れが大きいため，実践的に利用するのに適したデータとはいえず，あいづち生成タイミングの推定において，十分な精度は達成されていない．本研究では，既存の車内対話データに対して，あいづちの生成に適した位置に網羅的にタグ付けすることにより構築した，あいづちコーパスを用いて，適切なあいづち生成タイミングを高精度に検出する．あいづち生成タイミングの検出実験を行い，本手法は，人手による検出結果を若干下回る程度の推定性能（適合率 78.6%，再現率 64.7%）を達成した．また，被験者 6 名による主観的評価を実施した結果，本手法により生成されたあいづちの 95.4%が半数以上の被験者により「不自然ではない」と判定されており，本手法の有効性を確認した．

2　あいづちコーパス

あいづちタイミングについては，人間による対話を調査した研究がいくつかあるものの [2, 7, 10, 11, 13]，これらの知見だけでもって，あいづちを生成する機構を実現することは容易ではない．というのも，あるタイミングで生成されたあいづちの適切さとは，音響あるいは言語的な諸要因の複合によるものであり，それを統一的に体系化することは困難であるためである．これに対する一つの解決法として，タイミングの適切さを，大規模データに基づいて判定することが考えられるが，現存する音声対話コーパスに収録されたあいづちの発生タイミングは，対話の環境や聞き手の心理状態などによる揺れが大きく，上述の目的に直接利用するためのデータとして適さない．

そこで著者らは，上述の問題を解決し，より実践的なデータを整備するために，以下の 4 つの方針を設けてタグ付け作業を実施し，タグ付けの安定性を備えたあいづちコーパスを構築した [5]．なお，タグ付け作業は，CIAIR 車内音声対話コーパス [6] に収録されているドライバ発話に対して，1 名の作業者により実施した．

1）網羅的なタグ付け: 作業者は，不自然でないあらゆるタイミングにあいづちタグを付与する．

2）オフライン環境でのタグ付け: 作業者は，付与対象の音声を一回以上聞いた上で，ドライバ音声の書き起こしテキストに対してタグ付けする．

3）あいづち発生タイミングの離散化: 対話ターンを，時間軸上に連続した形態素区間または無音区間（以下，**基本区間**）からなる列であるとし，作業者は，基本区間ごとに，あいづち生成タイミングとして適切であるかを判断し，適切であればタグを付与する．なお，無音区間については，200 ミリ秒を超える場合には，200 ミリ秒の無音区間 (sp) とそれ以外の無音区間 (pause) とに分割し，それぞれを基本区間とした．

4）あいづち合成音の再生による推敲: 作業者がタグを付与したタイミングであいづちが発生する対話音声を自動生成し，作業者はその音声を再生することによって推敲する．本研究はあいづちの発生タイミング

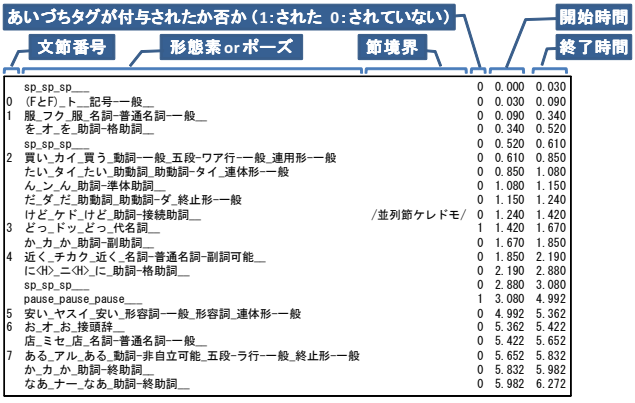

図 1: あいづちコーパスの例

表 1: あいづちコーパスの規模

話者	346
発話ターン	11,181
節	14,643
文節	43,723
形態素区間	94,030
無音区間	19,142
あいづち	5,416

に焦点をあてるため，いくつか存在する，あいづちの種類のうち，理解・同意を示す最も一般的な様式として「はい」を採用し，その合成音を HITACHI 製の音声合成ソフトウェア HitVoice を用いて生成した．また，あいづち音声は，タグ付けされた基本区間の開始から 50 ミリ秒後のタイミングで発生を開始することとした．

図 1 に構築したあいづちコーパスの例を示す．各行は，基本区間を意味しており，ドライバ発話における形態素区間または無音区間の情報が表記されている．また，開始終了時間，ならびに，あいづちタグが付与されたか否かの情報を与えている．さらに，形態素の場合は，形態素情報や文節境界情報，節境界情報を付与している．ここで，形態素情報は ChaSen[9] + Unidic[12] を，節境界情報は CBAP[14] を，各基本区間の開始・終了時間は連続音声認識システム Julius[8] を用いて自動的に付与した．また，文節境界は CIAIR 車内音声対話コーパスに付与されているものを利用した．なお，表 1 に構築したあいづちコーパスの規模を示す．

本研究では，構築したあいづちコーパスにおけるタグ付けの安定性とタグ付け位置の妥当性を評価するため，(1) 4 名（コーパス作成者と被験者 A,B,C）によるタグ付け結果の　値 [4] の測定と，(2) タグ付け結果に基づいて生成したあいづち音声の自然さに関する主観的評価，を実施した．その結果，(1) では，　値が usable quality（.67 <　< .80）[1] を示し，(2) では，全体の 98.5%のあいづちが自然であると被験者により判断されており，本タグ付け作業によって，高い安定性と自然さを備えたあいづちコーパスが構築できることを確認した [5].

表 2: SVM で用いた素性

文節境界に関する素性	
1.	m_j(直近の形態素区間) が文節の最終形態素であるか否か
2.	1 が真の場合，m_j の品詞
3.	1 が真の場合，m_j が属する文節内に，以下の 4 分類のうち，いずれの形態素が存在するか．（名詞，動詞，形容詞，それ以外）
節境界に関する素性	
4.	m_j が節の最終形態素であるか否か
5.	4 が真の場合，m_j が属する節の種類
無音区間に関する素性	
6.	m_i が sp であるか否か
7.	6 が真の場合，m_i のポーズ長が以下の 3 分類のいずれであるか．（0.17 秒未満，0.17 秒以上 0.20 秒未満，0.20 秒）
8.	m_i が pause であるか否か
母音の引き伸ばしに関する素性	
9.	m_j を構成する最後の 1 モーラの時間長が以下の 3 分類のいずれであるか．（0.17 秒未満，0.17 秒以上 0.25 秒未満，0.25 秒以上）
発話速度に関する素性	
10.	m_j の発話速度が平均発話速度より遅いか否か
11.	10 が真の場合，m_j の発話速度と平均発話速度の差が以下の 3 分類のいずれであるか．（3 モーラ/秒未満未満，3 モーラ/秒以上 9 モーラ/秒未満，9 モーラ/秒以上）
ピッチ，パワーに関する素性	
12.	ピッチ変動パターン
13.	パワー変動パターン
直前のあいづち生成からの経過時間に関する素性	
14.	直前のあいづち生成時間から m_i の終了時間までの時間長が δ(秒) が以下の 3 分類のいずれであるか．（1.2 秒未満，1.2 秒以上 2.9 秒未満，2.9 秒以上 5.0 秒未満，5 秒以上）

3　あいづち生成タイミングの検出

本研究では，1 対話ターンの基本区間列 $m_1 \ldots m_n$ 中の基本区間が連続して入力されることを想定し，基本区間が 1 つ入力されるごとに，その入力基本区間の直後に対して，あいづちを生成できるか否かを Support Vector Machine(SVM) を用いて推定する．

表 2 に，ある基本区間 m_i の直後にあいづちを生成するか否かを決定する際に利用した素性を示す．これらの素性は，先行研究による知見に基づいたものであり，デベロップメントデータを用いて実験的に決定した．また，すべて基本区間列 $m_1 \ldots m_i$ から得られる素性である．ここで，m_j とは，直近の形態素区間のことであり，m_i が形態素の場合は m_i，m_i が sp の場合は $m_{i\ 1}$，m_i が pause の場合は $m_{i\ 2}$，となる．素性 10 と 11 の平均発話速度は，そのドライバがその形態素より前に発話した全形態素に対する平均発話速度である．素性 12 と 13 における変動パターンは，文献 [7] と同様の方法で求める．素性 14 は，直前にあいづちが生成されていない場合は使用しない．なお，各基本区間の言語情報や音響情報，時間情報は，その基本区間の入力が終わり次第得られるものとする．

基本区間	もう	ガソリン	が	ない	ん	で	給油	し	たい	ん	です	けど<H>	sp	pause	(Fえんと F)	sp	今	ねえ	sp	pause	これ	は	吹上	の	ほう	に	いる	ん	だ	けど	何	か	ない	か	なあ
正解							1						1							1											1				
実験結果							1						1							1											1				

図 2: あいづちタイミング推定の成功例

4 検出実験

構築したあいづちコーパス [5] を用いてあいづち生成タイミングの検出実験を実施した.

4.1 実験概要

実験は交差検定により実施した. 表 1 に示すデータを 10 分割してグループに分け, そのうちの 1 グループをテストデータとし, 残りの 9 グループを学習データとして使用した. ただし, 10 グループのうち 1 グループは素性決定のためのデベロップメントデータとして使用したため, 評価データから取り除き, 残りの 9 グループに対する実験結果に基づいて評価した. 実験のための SVM のツールとして LibSVM[3] をデフォルトのオプションのまま使用した.

評価には以下の指標を用いた.

$$\text{再現率} = \frac{\text{正しく生成されたあいづち数}}{\text{正解のあいづち数}}$$

$$\text{適合率} = \frac{\text{正しく生成されたあいづち数}}{\text{生成されたあいづち数}}$$

コーパス構築作業者のタグ付け結果を正解データとし, その正解データとの間であいづち生成箇所が一致すれば正しく生成されたと判定した.

4.2 実験結果

評価データに対する本手法の再現率, 適合率及びこれらの調和平均である F 値は, それぞれ, 72.9% (3,147/4,317), 64.6% (3,147/4,870), 68.5% であった. また, 本手法による検出結果と正解データが全ての基本区間で一致した対話ターンは 8,009 ターン存在し, 全対話ターンの 80.4%を占めた. ただし, 正解データにおいてあいづちタグが 1 回以上付与された対話ターンは 3,162 個あり, そのうち, 本手法が全ての基本区間で正解した対話ターンは 47.1% (1,489/3,162) であった. 1 対話ターンの全基本区間であいづちタイミング推定が成功した例を図 2 に示す. 1 行目は話者音声の基本区間列を, 2 行目は正解データ上のタグ付けを, 3 行目は本手法の推定結果をそれぞれ表す. 1 が記された基本区間の開始直後にあいづちを生成していることを表す. この例では, 基本区間「給油」「sp」「pause」「何」の開始直後 (「で」「けど <H>」「sp」「けど」の終了直後) へのあいづち生成が成功している様子がわかる.

比較のために, 評価データの一部に対して, 被験者 3 名 (A, B, C) によるあいづちタグ付け作業を実施し, それぞれの適合率と再現率を測定した. これらの結果と同一データに対する本手法の検出結果を表 3 に示す. 本手法は, 適合率において最も高い値を示している一方で, 再現率は最も低い値となっており, 適切なタイミングであいづちを生成できるものの, 人間と比べ, あいづちを生成可能なタイミングを検出し損なう傾向にあることが分かる. F 値においては, 本手法は, 被験者によるタグ付け結果を若干下回る程度の検出性能を達成している.

表 3: 人間による推定結果との比較

	適合率	再現率	F 値
本手法	78.6% (66/84)	64.7% (66/102)	71.0
被験者 A	73.2% (82/112)	80.4% (82/102)	76.6
被験者 B	77.6% (83/107)	81.4% (83/102)	79.4
被験者 C	71.2% (74/104)	72.6% (74/102)	71.8

表 4: 不自然ではないと判定されたあいづちの割合

被験者	あいづちの割合
1	81.5% (53/65)
2	83.1% (54/65)
3	90.8% (59/65)
4	92.3% (60/65)
5	87.7% (57/65)
6	95.4% (62/65)

4.3 主観評価

上記では, 正解データとの比較によってあいづち生成タイミングの検出結果を評価した. しかし, 正解データと一致していないあいづち生成箇所が必ずしも不適切なタイミングであるとは限らない. そこで, 本手法のあいづち検出結果に対して, 被験者による主観的評価を実施した. 被験者 6 名が, ドライバ音声に対して本手法が生成したタイミングであいづち音声が挿入された音声を聴取し, 各あいづちについて, その生成タイミングが不自然でないか否かを主観的に評価した. 評価には, 検出実験において本手法があいづちを生成した対話ターンからランダムに 50 対話ターンを抽出し, この中に含まれる 65 個のあいづちを使用した. なお, コーパス構築時と同様に, あいづち音声は HITACHI 製の音声合成ソフトウェア HitVoice を用いて生成し, あいづちが生成されると推定された基本区間の開始時間から 50ms 後をあいづち音声の開始位置とした.

表 4 に, 各被験者が「不自然ではない」と判定したあいづちの割合を示す. 6 人の平均で 88.4%のあいづちが「不自然ではない」と判定された. 図 3 は,「不自然である」と判定した被験者数により各あいづちを

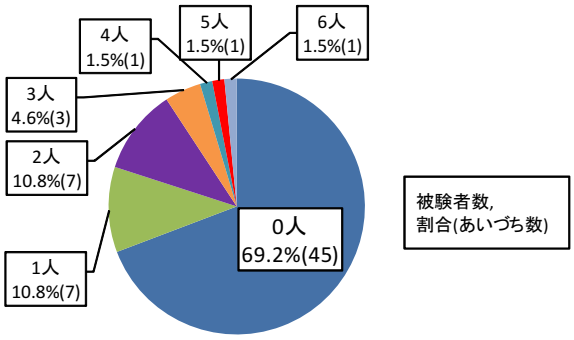

図 3: 「不自然である」と判定した被験者数による，あいづちの分類とその頻度割合

基本区間	家族	で	わいわい	食事	し	たい	ん	だ	けど	騒い	で	も
実験結果										1		

あんまり	迷惑	なら	ない	お	店	って	あり	ます	か<H>
1									

図 4: 不自然であると判定されたあいづち

分類し，各分類クラスの頻度割合を示したグラフである．6 人全員により「不自然である」と判定されたあいづちが 1 個存在した．このあいづちを図 4 に示す．このうち，2 つ目のあいづち，つまり「も」の終了直後 (「あんまり」の開始直後) に生成されたあいづちが被験者全員に「不自然である」と判定されたあいづちである．1 つ目のあいづち（「けど」の終了直後のあいづち）が生成されてすぐにあいづちが生成されており，人間であれば生成しないあいづちであると考えられる．このようなあいづちの生成を防ぐことが必要となる．一方，6 人のうち 3 人以上の被験者により「不自然ではない」と判定されたあいづち（すなわち，3 人以下の被験者によってのみ「不自然である」と判定されたあいづち）は，全体の 95.4% (62/65) を占めており，本手法がかなり高い確率で「不自然ではない」あいづちを生成できていることがわかる．

5 おわりに

本論文では，タグ付けの安定性を備えた大規模あいづちコーパスを用いて，統計的手法により，あいづちの生成タイミングを検出する手法を提案した．検出実験の結果，被験者によるあいづち生成タイミングの検出結果を若干下回るものの同程度の F 値（本手法：71.0，被験者の平均：75.9）を達成した．また，被験者 6 名による主観的評価を実施した結果，本手法により生成されたあいづちの 95.4%が半数以上の被験者により「不自然ではない」と判定されており，本手法の有効性を確認した．

今後は，誰もが不自然だと感じるあいづちの生成を防ぐことを検討したい．また，実際にリアルタイム環境であいづちを生成することを試みる予定である．

謝辞 本研究は一部，科研費挑戦的萌芽研究（No. 21650028）「音声対話システムの個性化に関する基礎的研究」による．

参考文献

[1] J. Carletta. Assessing agreement on classi cation tasks. *Computational Linguistics*, Vol. 22, No. 2, pp. 249–254, 1996.

[2] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *Proc. of EACL2003*, pp. 51–58, 2003.

[3] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, pp. 37–46, 1960.

[5] Y. Kamiya, T. Ohno, and S. Matsubara. Coherent back-channel feedback tagging of in-car spoken dialogue corpus. In *Proc. of SIGDIAL2010*, pp. 205–208, 2010.

[6] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura. CIAIR in-car speech corpus – influence of driving status–. *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, pp. 578–582, 2005.

[7] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Journal of JSAI*, Vol. 20, No. 3, pp. 220–228, 2005.

[8] A. Lee, T. Kawahara, and K. Shikano. Julius – an open source real-time large vocabulary recognition engine. In *Proc. of EUROSPEECH2001*, pp. 1691–1694, 2001.

[9] Y. Matsumoto, A. Kitauchi, T. Yamashita, and Y. Hirano. *Japanese Morphological Analysis System ChaSen version 2.0 Manual*. NAIST Technical Report, NAIST-IS-TR99009, 1999.

[10] S. K. Maynard. *Japanese conversation : self-contextualization through structure and interactional management*. Ablex, 1989.

[11] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, Vol. 32, pp. 1177–1207, 2000.

[12] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–122, 2007.

[13] 堀口純子. 日本語教育と会話分析. くろしお出版, 1997.

[14] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.