自然言語処理　143－14
音声言語情報処理　36－9
（２００１．６．１）

superwordモデルに基づく話者交替関連表現の抽出と予測力評価

森　大毅　　粕谷 英樹

宇都宮大学工学部

音声対話システムのための言語モデルとしてsuperwordモデルを提案しており、パープレキシティの点で優れていることがわかっている。本報告では、音声対話システムの応答タイミングの高度な制御を目的として、superwordに基づく話者交替の予測モデルを提案する。話者交替／非交替のキューとなる表現の抽出のため、superword確率から計算されるキューの強度を定義した。キューの強度に従って抽出したsuperwordには、話者交替に関係があると思われる表現が多く含まれていた。また、一部のタスクに対してはキューの強度分布が実際の話者交替／非交替によって異なることから、提案した予測モデルの有効性が示された。

Extraction of Turn-Taking-Related Expressions and Evaluation as Predictors Based on the Superword Model

Hiroki Mori　　Hideki Kasuya

Faculty of Engineering, Utsunomiya University
7-1-2, Yoto, Utsunomiya-shi, 321-8585 Japan

The superword model is a data-driven framework for dialogue modeling and its superiority was shown in our previous works. In this report, we propose a superword-based turn-taking prediction model for precise control of responce timing of spoken dialogue systems. First, cue intensity is defined with superword probability in order to extract cue expressions for turn-taking or turn-holding. Extracted superword set is shown to include a lot of relevant expressions to turn-taking. Finally, the effectiveness of the proposed prediction model for some tasks has been revealed by showing the difference of cue distribution according to actual turn-taking / turn-holding.

1 はじめに

人間は音声中の様々な情報を無意識下に活用することで対話を円滑に行っている。韻律情報はその代表的なものであり、話者の意図や態度を伝えるだけでなく、メタコミュニケーションの形成にも寄与している [1, 2]。韻律情報を利用することにより、音声対話システムにおいては人間と機械の円滑で自然な会話を可能とするような対話管理も可能となろう。その一例が、ユーザに対する応答タイミングの制御である。岡登ら [3] は、韻律テンプレートを用いた相槌箇所の予測実験を行い、58%の精度のとき 15%の検出率が得られたと報告している。この結果は韻律情報の有効性を裏付けてはいるが、またそれと同時に分節音情報のような他の情報をも利用すべきことを示唆している。

この意味においても、対話の円滑な進行に大きな役割を果たしていると考えられる語を整理することは重要である。しかしながら、特に日本語の場合には、対話の円滑な進行に寄与すると考えられる表現はしばしば語彙的でなく、また話者間の親密さや社会的関係に依存する様々な発話様式によってその種類や役割は変化する。さらに、/e:to/ という表現がしばしば /eQto/ や /eto/, /to/ などと変化するように、この種の表現は対話音声特有の変形の影響を強く受ける。このため、これらの表現とその談話における役割を記述的にモデル化するのは困難である。

一方我々は、対話音声の言語モデルとして superword モデルと呼ぶ枠組を提案している。superword モデルは単語 n-gram のスーパーセットとして定義されるが、単語辞書を必要とせず対話データベースのみから学習できるという特長を持っている。日本語の対話音声は、単語の定義が曖昧であるばかりでなく前述したような特性を持っているため、事例を効果的に反映できる枠組は有望である。

本報告では、音声対話システムの応答タイミングの高度な制御を目的として、自発的 (spontaneous) な対話音声から発話権交替または発話継続のキューとなる表現を見出し、また発話交替のタイミングを予測する手法について述べる。さらに、これらトリガとなる表現と実際の発話交替／非交替との関係を観察し、本手法を評価する。

2 superword モデル

日本語のように単語境界が明確でない言語で単語 n-gram を構築する場合には、形態素解析などの前処理が必要になる。superword モデルは、単語境界を曖昧にしたまま n-gram を構築するためこのような処理は必要としない。対話音声に対しては、モーラ n-gram に比べて低いパープレキシティを示すことが過去の研究からわかっている [4]。

superword とは単語などの文字列を一般化したものであり、訓練テキスト中の任意の文字列を含み得る。ただし、言語モデルとして意味を持つために次の条件を満たす文字列を superword と定義する.

- 訓練テキスト中に最低 2 回出現する

または

- 長さ 1 の文字列である

これにより、全ての列は少なくとも 1 通りの superword の系列として表現できることが保証される. superword n-gram 確率 $P(w_i|w_{i-(n-1)}\cdots w_{i-1})$ は, 直前に $n-1$ 個の superword の列 $w_{i-(n-1)}\cdots w_{i-1}$ が生起したと仮定した時の superword w_i の条件付き生起確率である.

以下、何らかの発話内容を表すモーラ列を単に発話と呼ぶことにする。与えられた発話 $\boldsymbol{C} = C_1C_2\cdots C_k$ が superword の列 $w_1w_2\cdots w_l$ に分割できるとき、$w_1w_2\cdots w_l \in \boldsymbol{C}$ と書く。superword n-gram モデルは、$\boldsymbol{C}$ の全ての可能な分割に関して計算した superword n-gram 確率の積の総和をもって $\boldsymbol{C}$ の発生確率を推定するものである。すなわち, その確率を次式で与える。

$$P(\boldsymbol{C}) = \sum_{w_1\cdots w_l \in \boldsymbol{C}} \prod_{i=1}^{l} P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (1)$$

ここで $n = 1$ の時、すなわち superword unigram モデルは、発話全体の生起確率がそれぞれ独立な superword の生起確率の積で表されるとする

ものであり、multigram[5] と呼ばれる可変長単語列に基づく言語モデルと同一のものである。

定義より、superword モデルは HMM の一種となり、その確率分布は Forward-Backward アルゴリズムにより学習できる。例えば、superword bigram 確率の再推定式は次により与えられる。

$$\tilde{P}(w_i|w_{i-1}) = \frac{\sum_u \sum_t \alpha_{t-1}(w_{i-1}) P(w_i|w_{i-1}) \beta_t(w_i)}{\sum_u \sum_t \alpha_t(w_i)\beta_t(w_i)}, \quad (2)$$

ただし、u は発話、$\alpha_t(w)$ は発話の先頭から時刻 t で superword w を発生するまでの累積確率、$\beta_t(w)$ は時刻 t で superword w を発生してから発話の末尾に至るまでの累積確率である。

superword モデルの学習は、(1) コーパス中の文字列出現頻度統計による superword 集合の獲得、(2)Forward-Backward アルゴリズムによるパラメータ最適化、の 2 段階で行う。タスクに依存しない包括的な対話音声のモデルを獲得するため、superword 集合の獲得プロセスは以下の手順により行った。

1. 対話コーパスをタスクにより分類する
2. 各タスクについて superword 集合を求める
3. ある 1 つの集合だけに属している superword を捨てる
4. 全ての集合を合併する

この操作により、モデルが特定のタスクへ特化するのを防ぐ。

3 話者交替のキュー

3.1 仮説

superword に基づく話者交替の予測モデルを考えるに先立ち、自然な音声対話に関して次のような仮説を立てた。

1. 話者交替のキューとなる表現の他に、話者非交替 (発話継続) のキューとなる表現が存在する
2. 話者交替／非交替のキューは発話末に現れるだけでなく、発話全体に分布する

ある種の表現が話者交替を引き起こすと考えることは妥当であるが、1 番目の仮説はそれに加えて、発話権を渡したくない場面では「ジャ」「デ」といった表現を含む発話がなされるであろう、ということを述べたものである。

また、発話末には話者交替に関係する表現が来やすいと言ってよいが、2 番目の仮説を認めるならば、例えば bigram のような局所情報だけでは発話交替を予測するには不十分で、もっと遠距離の制約を利用する必要があることになる。

3.2 話者交替の予測モデル

前節で述べたように、学習によって高い確率を付与された superword の中には談話構造に関係していると思われるものがいくつかある。そこで、もし特定の superword と談話におけるイベントの関係がわかれば、発話中にその superword を見ることによってイベントの発生を予測することができる。これは、談話におけるイベントのキューとなる表現を superword の中から探すことであると言い換えることができる。

superword の発生は陽には知ることはできないため、キューとなる表現は発話中に分布して存在すると考えなければならない。そのように分布している多数の superword から話者交替を予測することがここでの課題となる。

これはトリガモデル [6] と同じく多数の手がかりから 1 つのイベントを予測する問題であるが、ここではまず superword w に関する話者交替確率のモデル $P(\text{turntake}|w)$ を定義することにする。この値は、ある発話の中に superword w が現れたとき、その発話の直後に話者交替が起こる確率を意味する。

ある superword モデルに対し、上述の話者交替確率を次のように定義する。

$$P(\text{turntake}|w) = \frac{\sum_{u \in U_\text{T}} \sum_t \alpha_t(w)\beta_t(w)}{\sum_u \sum_t \alpha_t(w)\beta_t(w)} \quad (3)$$

ただし U_T は直後に話者交替がある発話の集合

である。superword の発生は陽には観測されないが、この確率はデコーディング中の任意の時点での superword 仮説に対して計算することができる。

3.3 キューの強度

1 発話の中にはかなり多くの superword が含まれる可能性があるため、これらには適当な重み付けをする必要がある。本報告では、上記の話者交替確率から求められる情報量を基に、重み付け関数に superword 確率の事前確率を用いた値、および発話 u に関する事後確率を用いた値の 2 種類によるキューの強度を定義する。

- 事前確率に基づくキューの強度

$$P(w)I(\text{turntake}; w) \tag{4}$$

- 事後確率に基づくキューの強度

$$P(w|u)I(\text{turntake}; w) \tag{5}$$

ただし $I(\text{turntake}; w)$ は話者交替または非交替に関する相互情報量で、

$P(\text{turntake}|w) > P(\text{turntake})$ のとき

$$I(\text{turntake}; w) = \log \frac{P(\text{turntake}|w)}{P(\text{turntake})} \tag{6}$$

それ以外

$$I(\text{turntake}; w) = \log \frac{1 - P(\text{turntake}|w)}{1 - P(\text{turntake})} \tag{7}$$

式 (4) はある superword に固有の重要度であり、モデルそのものに対する評価に用いる。式 (5) はある superword が特定の発話の中で占める重要度であり、未知のテストセットに対する評価に用いる。事後確率 $P(w|u)$ は式 (2) 右辺の u と t を固定することにより求められる。式 (6) を基に算出された強度が大きい superword は話者交替のキューとなる表現と考えられる。同様に、式 (7) を基にした強度が大きい superword は話者非交替のキューとなる表現と考えられる。

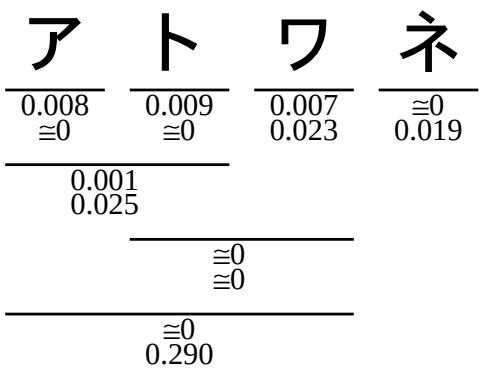

図 1: 上段:事前確率に基づくキューの強度 (ビット)。下段:事後確率に基づくキューの強度 (ビット)。この例では全ての superword が話者非交替のキューとなっている。

例として図 1 に、「アトワネ (後はね)」という発話に対する superword とその強度を示す。この発話には「ア」「ト」「ワ」「ネ」「アト」「トワ」「アトワ」の 7 個の superword が含まれ、特に「アトワ」の話者非交替のキューとしての強度が大きい。

4 評価

4.1 音声対話コーパス

学習および評価用の音声対話書き起こしデータは重点領域「音声対話」の対話音声コーパス [7] に含まれる「秘書システム」「スケジュール調整タスク」「スケジューリング会話」「クロスワードパズル」「テレフォンショッピング」「間違い探し」から選んだ 44 対話を基に、人手で音素表記にした後モーラ表記に変換したものを用いた。このうち各タスクからの最低 1 対話を含む 8 対話を評価用セットとし、残りを学習に用いた。あいづちは話者交替に直接関与しないものと考えコーパスから除いた。フィラーや言いよどみはそのまま残した。学習によって高い確率を付与された superword には、「ハイ」「ウン」のようなあいづち、「ワ」「ニ」などの格マーカ、「デ」「ジャー」「アノ」などの談話マーカ、「エート」「ナンカ」などのフィラーがあった。

データはコーパス中の句読点によらずポーズ

ハイ	0.491	カイテ	0.872
デス	0.987	シチ	0.915
デスカ	0.962	ソー	0.496
リ	0.928	オオネガイシマス	0.999
デショーカ	0.998	トユーコトデ	0.994
ハチ	0.502	ヒト	0.983
オネガイシマス	0.890	ヨネ	0.999
ビ	0.772	モノ	0.851
エン	0.943	シテ	0.999
ナナ	0.537	タイ	0.982
ジカン	0.937	シマス	0.991
キュー	0.731	ソーデス	0.827
ウチアワセ	0.970	ゴザイマ	0.999
デスネ	0.489	イーデスカ	0.857
ヨー	0.958	スカ	0.955
ナイ	0.557	ター	0.999
ホー	0.999	ボー	0.560
ッテル	0.985	サンカ	0.986
デヨロシーデショーカ	1.000	カイギシツ	0.000
チガウ	0.993	ミッツ	0.750
セ	0.609	テイタダキマス	1.000
オシ	0.844	ゾ	0.984

図 2: 話者交替のキューとなる superword

ガ	0.006	センセー	0.000
ノ	0.001	キ	0.000
ト	0.001	チュー	0.000
デ	0.015	ウン	0.202
エー	0.003	ド	0.168
ア	0.007	イチ	0.071
ワ	0.069	ダ	0.001
ニ	0.002	ヒ	0.002
オ	0.001	モン	0.005
サン	0.000	ツ	0.063
エ	0.000	ッテ	0.048
ク	0.009	ラ	0.000
ー	0.028	ム	0.000
モ	0.000	ゴー	0.000
エート	0.016	ヤ	0.003
ミ	0.000	ジャー	0.038
カ	0.101	ニー	0.000
ス	0.034	テ	0.010
ゴ	0.013	コ	0.009
ジュー	0.000	エット	0.000
シ	0.002	ナ	0.009
ロク	0.000	デワ	0.024

図 3: 話者非交替のキューとなる superword

節を単位として分割した。ポーズが与えられていないデータに対しては聴取によりポーズ位置を与えた。その後、各ポーズ節に対し、直後に話者交替が起こっているか否かに関するラベルを人手で付与した。総計 8117 ポーズ節のうち約 3 分の 1 に対し直後に話者交替が起こっているというラベルが付与された。

4.2 話者交替／非交替のキュー抽出実験

図 2 に、得られた話者交替のキューのうち式 (4) の強度が上位の superword を、図 3 に同じく話者非交替のキューのうち強度が上位の superword をそれぞれ示す。確率計算は superword unigram 確率に基づいた。数字は推定された話者交替確率を示す。 話者交替のキューの多くは文末表現で、明示的な談話マーカの他に「ッテル」(「うん光ってる」など)「トユーコトデ」(「木曜日の 1 時から 3 時ということで」など) といった、それ自身が持つ意味とは無関係に実際上発話末によく現れる表現が得られている。話者非交替のキューは話者交替のキューに比べて断片的で、「ガ」「ノ」「ト」「ワ」「ニ」といった格マーカ、「エー」「ア」「エ」「エート」といったフィラー、「デ」「ジャー」「デワ」といった談話マーカが見られる。

4.3 話者交替／非交替予測力評価実験

superword の話者交替予測力を評価するため、評価用セット中の superword に対して式 (5) の話者交替／非交替のキューの強度を求めた。確率計算は superword unigram 確率に基づいて行い、評価はタスク毎に個別に行った。

キューの強度と実際の話者交替／非交替との関係を図 4 および図 5 に示す。図 4 は「秘書システム」タスク中の 5 モーラから 15 モーラの長さのポーズ節に対するキューの強度分布、図 5 は同じく「クロスワードパズル」タスクに対する分布である。各点はそれぞれ評価セット中の superword を示す。このうち□はその superword が含まれるポーズ節の直後に話者交替が起こっているもの、×はその superword が含まれるポーズ節の直後に話者交替が起こっていないものを表す。横軸は各 superword の終端のポーズ節末から数えた位置であり、右端がポーズ節末に相当

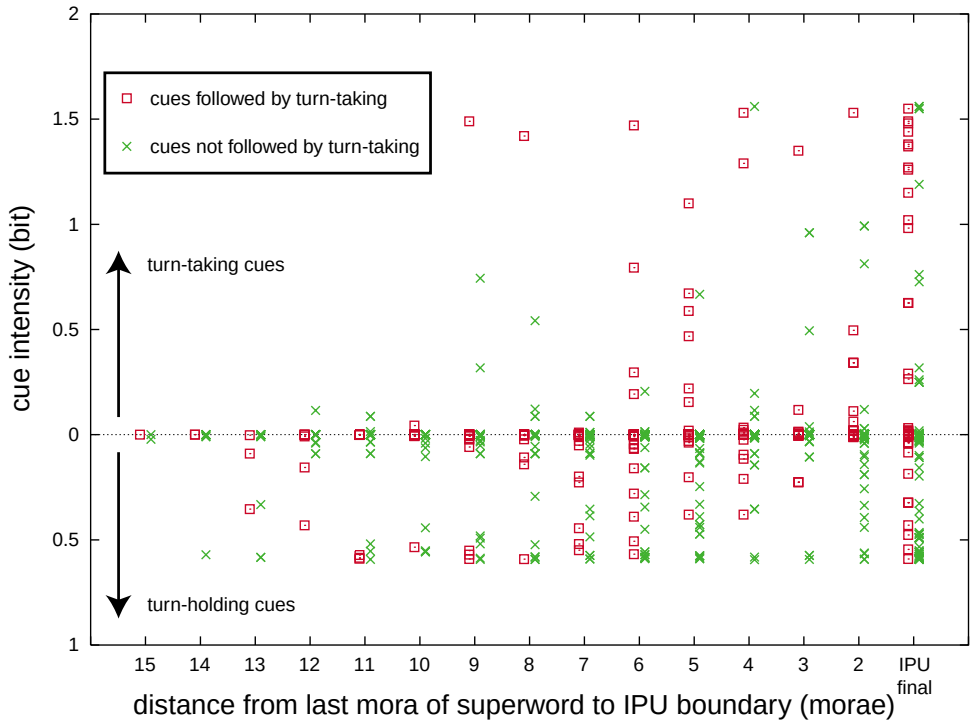

図 4: 「秘書システム」タスクに対するキュー強度分布

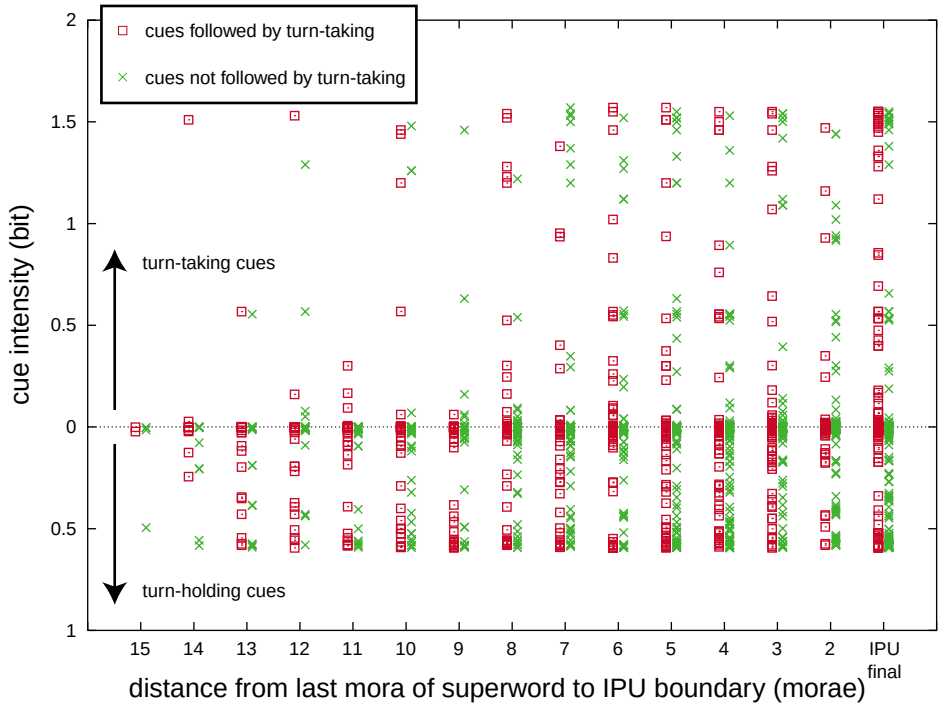

図 5: 「クロスワードパズル」タスクに対するキュー強度分布

する。縦軸はキューの強度を表す。ただし、話者交替のキューは上の領域に、話者非交替のキューは下の領域に描いてある。これらのキューにより話者交替／非交替が予測できるならば、□点と×点は異なる分布をしなければならない。また、3.1 節の仮説 1. が成り立つならば×点は図中下部の領域に集中すると考えられる。さらに、仮説 2. が成り立つならば横軸の右端 (ポーズ節末) だけでなくもっと左の部分でも上部に□点が、下部に×点が多く現れるはずである。

図 4 からは次のことがわかる。

1. 直後に話者交替が起こらないポーズ節中の superword の多くが図中下部の領域に集中している。つまり、ほとんどが話者非交替のキューである。この結果は 3.1 節の仮説 1. を裏付けるものとなっている。
2. 直後に話者交替が起こるポーズ節では、ポーズ節末 (図中右端) だけでなくもっと早い時点で話者交替のキューが多く現れている。この結果は仮説 2. を裏付けるものとなっている。
3. 話者交替／非交替に関わらず、早い時点 (ポーズ節末から数えて 4 モーラ以前) で話者非交替のキューが多く現れている。

よって、「秘書システム」タスクにおいては話者交替／非交替のキューに実際のイベントを予測する能力があり、またその分布が当初の仮説にほぼ沿ったものとなっていると言える。他に、「スケジュール調整タスク」においても概ね同様の傾向が見られた。

しかしながら、図 5 に示す「クロスワードパズル」タスクにおいてはこのような明確な分離は見られなかった。「スケジューリング会話」「テレフォンショッピング」「間違い探し」タスクに関しても、図 4 ほど明確ではなかった。

このような違いが見られた理由の 1 つとして、発話様式の違いが考えられる。「秘書システム」タスクにおける発話は比較的 well-formed であり、自発性がやや低いように感じられる。それに比べ、「クロスワードパズル」タスクにおける発話は自発的かつ自由である。後者において話者交替／非交替の予測が困難になった原因には、学習に用いたデータの大部分が占める発話様式と個々のテストデータの様式が異なっていた可能性が考えられる。学習データの拡充が必要であるが、同時に発話様式に対する superword モデルの適応化を行うことにより、不整合を解消し予測性能を上げることができると思われる。

5 おわりに

本報告では、superword モデルに基づく話者交替の予測法を提案し、自発音声の対話データベースを用いた評価を行った。キューの強度と話者交替の関係を調べた結果、いくつかのタスクに対してはキューの強度分布はその発話の直後に話者交替が起こるか否かにより変化し、キューに関するいくつかの仮説が裏付けられた。

しかしながら、特に他のいくつかのタスクに対しては偽の話者交替キューが多く観察された。この主たる原因として、発話中の実質語が細かい superword に不適当に分割されていることが挙げられる。superword モデルの音声認識システムにおける使われ方としては、ネットワーク文法などにより記述された文節文法によりスポッティングを行う際の背景モデルとしての利用が考えられる。この時は、superword は主に実質語以外の部分にマッチするため、上述したような悪影響を及ぼすことは少なくなると思われる。また、出現回数を考慮するなどしてキューの強度の定義を見直すことにより改善が図れる可能性もある。

提案した発話交替のキューを実際の音声対話システムにおける対話制御に利用するためには、ユーザの発話に含まれる複数のキューからシステムが応答すべきかどうかを判定する必要がある。今後の課題として、このような判定を行う関数を定義し評価を行うこと、および実際の音声対話システムを用いた評価を挙げる。

参考文献

[1] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y.: An Analysis of Turn-Taking and Backchannels Based

on Prosodic and Syntactic Features in Japanese Map Task Dialogues, *Language and Speech*, Vol.41, Nos.3–4, pp.295–321 (1998).

[2] Ward, N.: Using Prosodic Clues to Decide When to Produce Back-channel Utterances, *Proc. ICSLP 96*, pp.1728–1731 (1996).

[3] 岡登 洋平, 加藤 佳司, 山本 幹雄, 板橋 秀一: 韻律情報を用いた相槌の挿入, 情報処理学会論文誌, Vol.40, No.2, pp.469–478 (1999).

[4] Mori, H. and Kasuya, H.: Automatic Lexicon Generation and Dialogue Modeling for Spontaneous Speech, *Proc. ICSLP 2000*, pp.577–580 (2000).

[5] Deligne, S. and Bimbot, F.: Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams, *Proc. ICASSP 95*, pp.169–172 (1995).

[6] Rosenfeld, R.: A maximum entropy approach to adaptive statistical language modelling, *Computer Speech and Language*, Vol.10, pp.187–228 (1996).

[7] Itahashi S., Yamamoto M. and Kawahara, T.: Speech Corpus by "Spoken Dialogue" Project, *Proc. Oriental-COCOSDA Workshop*, pp.156–161 (1998).