リアルタイム発話継続/交替予測システムの構築

Real-time Turn-taking Prediction System

小川 翼 [1*] 伊藤 敏彦 [1]

Tsubasa Ogawa [1] and Toshihiko Itoh [1]

[1] 北海道大学 大学院情報科学研究科

[1] Graduate School of Information Science and Technology, Hokkaido University

Abstract: In this study, we attempt to overcome a barrier between humans and machines by developing a system that can predict end-of-utterances and turn-taking with good accuracy. This system predicts them using two machine-learning classifiers. Experimental results indicate that the system can effectively predict four classes, i.e., no turn-taking, turn-taking, speaking, and silence, at an interval of 100 ms with f-measures of 0.51, 0.74, 0.99, and 1.00, respectively. The system suppresses typical user dissatisfaction because the system can detect an end-of-utterance at that a user expects it to respond.

1 はじめに

1.1 背景

近年，音声対話システムの性能は飛躍的に向上してきており，多くの研究がなされている．しかし，現在のところ人―機械間のコミュニケーションが人間同士のものと同等であるとはいえない．ユーザがシステムに発話をしてから応答が返ってくるまでに長い時間がかかることが原因の一つである．従来，人―機械間のコミュニケーションは音声認識や言語理解の精度が重要であると考えられてきたが，近年の研究では対話リズム（対話のテンポ）や，システム発話の韻律情報が十分でないと人は機械と自然な対話を行えないことが示されている[1]．対話リズムは円滑なコミュニケーションを行う上で必要不可欠なものであり，人間はこのリズムを良好にするため発話を無意識的にコントロールしているものと考えられる．

そこで，適切なリズムの発話を生成することにより人間らしさを感じられる音声対話システムの実現を目指す．しかし，そのためには無音で区切られたユーザの発話終了を前もって予測しそこが発話継続なのか発話交替なのかを判別することによって適切な話者交替[2]を行い，さらにシステムの内部状態である発話意図や感情などに応じて発話タイミングも制御する必要があると考えられる．加えて，ユーザとの同調作用やオーバーラップなども行えるようにする必要もあり，課題は多い．

我々はまず研究の第一段階として「発話継続/交替の判別を行うシステム」の開発を行い，発話継続/交替について 6 割程度の F 値で判別できることを示した[3]．本稿は，このシステムに改良を加え発話継続/交替だけでなく発話区間や発話終了も高精度に判別できるシステムについて述べる．

1.2 関連研究

発話終了検出に関する研究として，Ferrer ら[4]，Hariharan ら[5]や Edlund ら[6]によるものが挙げられる．一方発話継続/交替の分析として，大須賀ら[7]は F0（基本周波数），パワー，時間長といった情報のみから発話交替か否かを判別し，発話継続/交替を予測可能とするような情報が韻律特徴量にも表出しているという可能性を示した．木村ら[8]はより正確に F0 を予測するために F0 モデルを導入し，発話継続/交替の判別精度を 2%程度向上させ，オープン実験（2 クラス判別）で約 7 割の判別精度を示した．

音声対話システム上の動作を想定した発話継続/交替判別手法として，Raux & Eskenazi[9]による 6 状態を想定したもの，Sato ら[10]による 2 状態の判別を行うものなどがある．Sato らは判別に有効である素性として文末表現やシステムの理解結果などを挙げ，高精度な判別は言語的情報に依存していることを示した．また，Nishimura & Nakagawa[11]は音声対話コーパスを学習し，100 ms ごとに応答するかどうかの判定を決定木で行った．彼らの手法はシステム

*〒060-0814 札幌市北区北 14 条西 9 丁目
E-mail: {togawa, t-itoh}@media.eng.hokudai.ac.jp

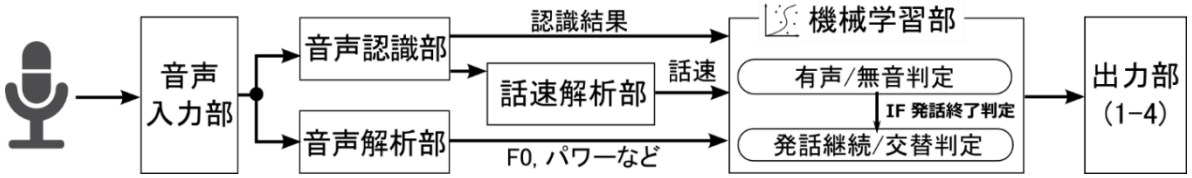

図 1: 本システムの構成

表 1: クラスと発話状態

クラス	状態
1	発話継続
2	発話交替
3	発話中
4	無音

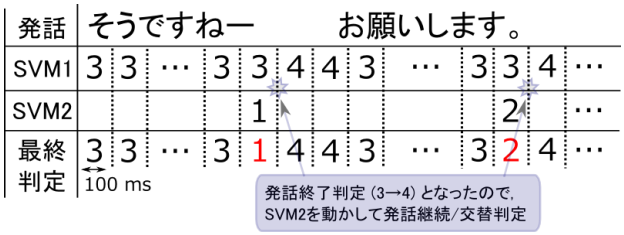

図 2: 最終的な判定の算出方法

側の内部状態を考慮せず，学習器の出力結果に依存する形となる．本手法は発話タイミングでなく発話状態を予測するため，発話タイミングはシステム側が決定する形となる．特に，ユーザの期待する発話タイミングは会話の文脈により大きく異なる[12]ことからも，システムがユーザの意図や感情をきちんと認識してタイミングを決定させる必要がある．

2 本研究の手法

2.1 システム構成と動作の流れ

本システムの構成は図 1 の通りである．音声入力部から音声データ（16 kHz, 16 bit, 1 ch）と音声認識結果が得られ，システムはこれらのデータを用いて各解析部で判別を行うための素性を計算する．クラス判別には Support Vector Machine（SVM）を用いソフトウェアとして LibSVM[13]を，また音声入力/認識システムとして Julius[14]を利用した．

提案手法では表 1 のように発話の状態を 4 つのクラスに分け，SVM を 2 段階で利用することで高精度に 4 値判別を行う（図 2）．1 段階目では音声区間かどうかを判定し**発話中**と**無音**のどちらかのクラスを出力する．ここでは仮に発話継続/交替地点であっても発話中となる．そして，発話中から無音に切り替わったところで 2 段階目の SVM 判別が行われ，**発話継続**と**発話交替**のどちらかのクラスが出力される．判別間隔は 100 ms である．

2.2 素性

阪田・広瀬[15]は文末に向かい話速は遅くなる傾向があるという知見を示した．また，大野ら[16]によると相手の発話がゆっくりになったときにあいづちが打たれやすいと述べている．あいづちでは発話権の譲渡がなされないため，聞き手があいづちをうった箇所では発話継続しやすいと考えられる．

ワード[17]，大須賀ら[7]，そして小磯ら[18]は，F0 やパワーは発話終了や発話継続/交替の予測に有効であると示している．また小磯らは発話継続/交替と密接に関係している品詞があると主張している．例えば，発話末における動詞や終助詞では発話交替となる場合が多く，副詞や接続詞に関しては発話継続となる場合が多い．よって文末の品詞を調べることで発話継続/交替の判別に寄与すると考えられる．

発話長や（Δ）MFCC も予備実験で精度向上の効果があったため素性として用いた．なお，発話長と発話終了（継続/交替も同様）には非線形な関係があると考えられるため，発話長は平方根の結果を利用した．

以上より，本手法では次に示す 48 個の素性を用いることとする．素性の値はすべて正規化が施されている．

- 話速，F0，パワーの変化量（現在値，100ms 過去値）
- 話速，F0，パワーを最小二乗法で線形近似した傾き（1 秒，500ms，300ms 過去値）
- 過去 200ms 間における F0 とパワーの変化パターン（パターンの種類については文献[18]に準拠している．）
- 音声認識結果末 2 形態素分の品詞
- 発話長の平方根
- MFCC，ΔMFCC（12+12 次元）

表 2: 作成したデータセット

対話数	36
人数	21 (重複あり)
発話数	1,201
時間	約 35 分
クラス 1: 発話継続	473
クラス 2: 発話交替	728
クラス 3: 発話中	43,465
クラス 4: 無音	80,584

素性の計算について, 話速と品詞情報は MeCab[†]と Julius 単語・音素セグメンテーションキットを用いた. F0 とパワーは WaveSurfer[19]の解析結果を用いた. 解析パラメータはフレームシフト 10ms, フレーム長 25ms（F0）・200 points（パワー）である. また, MFCC の計算はデフォルト設定の HTK[‡]を用いた.

2.3 データセット

機械学習に用いるデータセットの作成には音声対話コーパスを用いた. コーパスは RWC コーパス[20]を利用した. RWC コーパスには「海外旅行計画」と「車の購入」という 2 種類のタスクが設定されており, 片方のチャンネルに 1 人ずつステレオの WAVE ファイルとして提供されている. 学習の際にはステレオの音声を分割して両方とも利用した.

コーパスには音声データの他に専門家によって付与されたラベリングデータが付属しており, ラベリングデータ中の発話終了時刻を発話継続/交替の位置に用いた. ラベルを決めるにあたっては**発話権の交替が起きなかったら「発話継続」, 起きたら「発話交替」**というルールに従って決定した.

これらの素性と, 前節のラベル情報を用いてデータセットを作成した. 加えて, 学習における悪影響を抑えるために以下の処理を行った. 完成したデータセットの詳細を表 2 に示す.

- あいづちは発話権を持たない発話であると考えられることから, あいづちのみの発話データはデータセットから除去した.
- 無音のうち, 判別が容易なもの（すべての素性値が 0 またはそれに準じた値）は除去した.

本手法では SVM を 2 段階で用いるため, 前節で作成したデータセットを 2 つ用意し, 片方は発話終了判定のために発話継続クラスと発話交替クラスをそれぞれ発話中クラスに書き換えることで発話中か否かの 2 値判別ができるようにした.

もう片方については, 逆に発話中クラスと無音クラスをデータセットから除くことで発話継続/交替の 2 値判別ができるようにした.

3 評価実験

本稿では, リアルタイム環境における判別の実現可能性を調べるためまず分析精度がより高いオフライン環境による実験を行った. 評価方法として**発話単位の** 10-fold cross validation を用いた. フレームをランダムに分割する通常の cross validation を用いると, 判別精度が高く出る傾向があるからである. 評価についてはフレーム単位での正誤がカウントされ, 適合率, 再現率, F 値の 3 指標で表される.

以下, 発話状態については SVM で判別する際にそれぞれ設定したクラス番号で述べる（発話継続: (1), 発話交替: (2), 発話中: (3), 無音: (4)）.

3.1 1 段階目: 発話終了検出実験

本手法は SVM 1 段階目, 発話終了検出の精度に大きく依存している. 言いかえれば, 適切な位置で(3)→(4)と判別されなければいけない. このクラスの変化位置がずれてしまえば(1) or (2)と判別される位置もその分ずれ, 結果として不自然な間のシステム応答が生成されてしまう. よって, まずは発話終了の判別精度を調べるため「発話中/無音判別実験」「発話終了検出実験」の 2 つの実験を行った. 前者はフレーム単位で 2 値の区別ができるかの評価実験である. この実験は, 全（音声）区間としたもの, 発話終了の 2 フレーム: (3)→(4)のみを対象としたものをそれぞれ評価データとして用いて行った. 一方で, 発話終了検出実験は(3)→(4)の 2 フレームを**1 組とみなして**評価する実験である（すなわち, この 2 フレームが一致してはじめて正解となる）.

今回の実験では(1)と(2)の判別は必要ないため, 1 段階目で用いるデータセット（(3)と(4)の 2 値のみ）を使った. また, 過学習の防止と正確な発話終了検出に特化した学習モデルとするため「(3)→(4)と切り変わった点から前後 4 フレームずつ」のみを学習データの対象とし, これを用いてモデルを作成した. 学習パラメータはコスト $C = 4$, $\gamma = 2^{-9}$ である.

両実験の結果を図 3, 図 4 にそれぞれ示す. 図 3 の「全区間対象」より発話中と無音の区別がきちんとできていることがわかる. しかし図 4 の 0 ms の結果から, 発話終了位置という単位で見てみると約 0.5 の F 値しか得られなかった. この原因は図 3 の

[†] https://code.google.com/p/mecab/
[‡] http://htk.eng.cam.ac.uk/

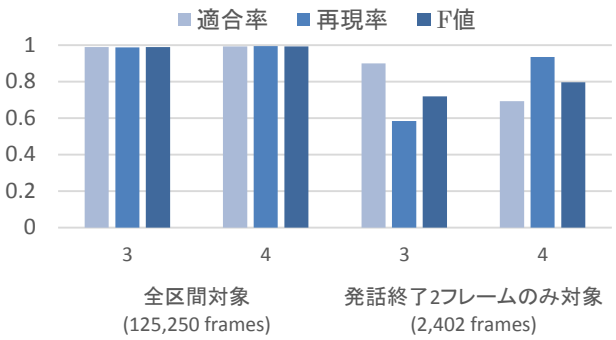

図 3: 発話中/無音判別実験の結果

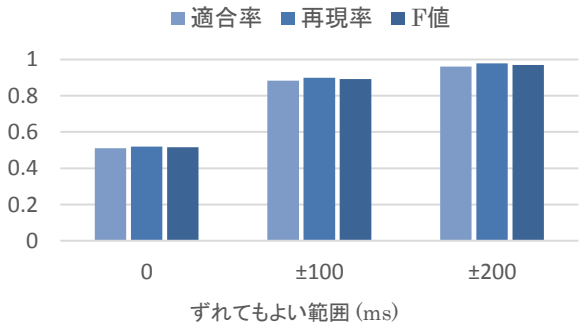

図 4: 発話終了検出実験の結果

「発話終了 2 フレームのみ対象」から考察できる．(3)よりも(4)の再現率が比較的高いことから，発話終了位置の(3)を(4)と予測する傾向がある，すなわち，システムは 1 フレーム（以上）早く発話終了判定する傾向があると考えられる．

この結果から改めて考えてみると，これまで行ってきた評価は 1 フレーム（100 ms）ごとの正誤判定になるため非常にシビアな評価であるといえる．そこで，±100 ms，±200 ms の「ずれ」も正解に含めるという条件で再実験を行った結果を図 4 の±100 ms, ±200 ms にそれぞれ示す．±100 ms 程度のずれを許せば 0.9, ±200 ms ではほぼ 1.0 の F 値を得ることができた．

3.2 2 段階目: 発話継続/交替判別実験

続いて 2 段階目にあたる発話継続/交替の判別実験を行った．データセットは 2 段階目の SVM で用いるもの（(1)と(2)の 2 値）を用いた．なお，今回の実験では章の初めで述べたような精度向上の危険性はないため，通常の cross validation で評価した．学習パラメータは $C = 512$, $\gamma = 2^{-7}$ である．

実験結果を表 3 に示す．結果から分かる通り，両者の区別がある程度できていることがわかる．したがって，仮に前節の発話終了判定が 100%の精度で行えれば，7 割～8 割の精度で発話状態の判別が行えることとなる．

表 3: 発話継続/交替判別実験の結果

	適合率	再現率	F 値
1	0.739 (322/436)	0.680 (322/473)	0.708
2	0.803 (614/765)	0.843 (614/728)	0.823

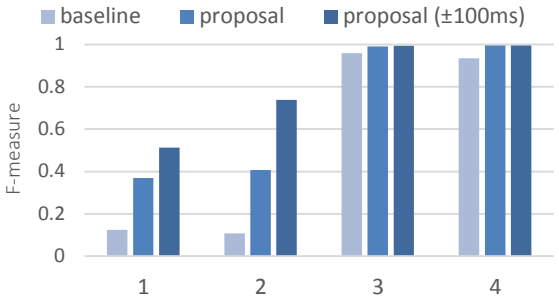

図 5: 4 値の判別結果

3.3 4 値判別実験

これまでの実験をふまえ，本手法による最終出力結果の精度をみてみる．2.1 節で述べたように SVM を 2 つ用いて(1)～(4)の 4 値判別を行った．

ベースラインと本手法による結果を図 5 にそれぞれ示す．ここでのベースラインとは 2 段階式を用いず 1 つの SVM のみで 4 値を判別する学習データを用いたときの結果である．ただし，表 2 からわかる通り学習データは(1)(2)と(3)(4)のサンプル数に大きな差がある不均衡なデータ[21]であり学習に悪影響が生じているため，最もサンプル数の少ないクラス(2)以外のデータを無作為に削除することでサンプル数を 4 クラスとも 473 に揃えた．

図 5 より，ベースラインは(1)と(2)の判別ができておらず F 値が約 0.1 と低い水準の精度であった．原因として適合率の値が極めて低い（両者とも 0.1 未満）ことから，(1)や(2)の不必要な湧き出し（False positive）が大量に発生していると考えられる．

一方，提案手法はベースライン以上の精度を示している．発話終了判定がされたとき**のみ**(1)や(2)の判別を行うという枠組みにより，湧き出しを効果的に減らすことができたと考えられる．しかしながら，図 4 で示したようにフレーム単位の評価では発話終了判定の精度が完璧ではないため，表 3 の結果ほど高精度の判別はできていない．

表 4: 4 値の判別結果（±100 ms のずれを許可）

	適合率	再現率	F 値
1	0.578 (218/377)	0.461 (218/473)	0.513
2	0.687 (581/846)	0.798 (581/728)	0.738
3	0.995 (42943/43146)	0.988 (42943/43465)	0.992
4	0.994 (80368/80889)	0.997 (80368/80584)	0.995

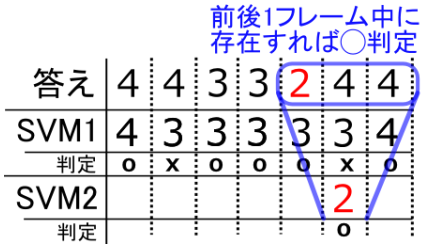

図 6: ±100 ms の探索判定

そこで，(1)と(2)の評価は 3.1 節のように±100 ms までずれを許可することとする（図 6）．(3)と(4)の判別についてはこれまで通りフレーム単位の評価で変わらない．この方法で再評価した実験結果を図 5 の「proposal (±100 ms)」と表 4 に示す．低かった(1)の F 値が 0.513 となり，各クラスが実用的な精度で検出されているといえる．また，表 4 より(2)の再現率が約 80%と高いことから，(システムの応答を期待している）ユーザの発話終了を高精度でキャッチできることを示している．したがって「発話権を譲渡したけれども反応がこない」といった代表的な *User Experience* の低下原因を抑制することが可能である．

4 考察

4.1 発話終了の判別精度

図 4 より，ずれの許容範囲を±200 ms に広げることで 1.0 に近い発話終了検出の F 値を得ている．言いかえれば，フレーム単位のシビアな条件（±0 ms）においてもミスの原因はほとんどが±200 ms 以内の判定ずれである．

そこで，正解の位置から±200 ms 以上ずれているミス: 湧きだしや検出漏れ（True negative）を調べたところ，湧きだしが 32 個（予測総数の 2.6%．そのうち(1)と判別した数が 28 個，(2)と判別した数が 4 個）で，検出漏れが 9 個（答え総数の 0.7%．そのうち(1)を漏らした数が 6 個，(2)を漏らした数が 3 個）であった．湧き出しや検出漏れの数は(1)の方が多い．しかし，実際のシステム動作を考えれば(1)の判定ミスは対話の破綻につながらない一方で，(2)のミスは不適切な発話や応答につながると考えられるため，本手法はほとんどの場合で適切な発話終了を検出できるといえる．

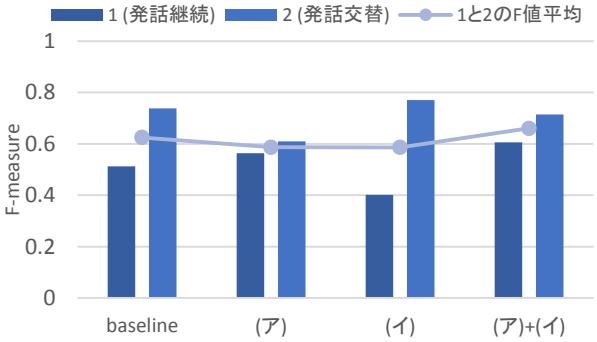

図 7: 条件を変えて実験した 1 と 2 の結果（±100 ms のずれを許可）

4.2 さらなる判別精度向上のために

3.1 節の結果より，システムは実際よりも 1 フレーム早く発話終了と判定しやすいことがわかった．そこで，

(ア) 発話終了判定の 1 フレーム後で発話継続/交替判定を行う

という条件で追加実験を行った．これまでのように±100 ms のずれを許可する条件で，表 4 の結果（baseline とする）と追加実験の結果を図 7 にそれぞれ示す．なお，(3)と(4)の結果についてはフレーム単位の評価であることからほとんど差が生じないため割愛している．

結果を見ると，(ア)の条件を用いることで(1)の F 値は上がったがその分(2)の F 値が大きく下がり，全体の結果を見ればベースラインの方が高性能であった．

性能向上のための他の手段として，学習データの増強を図った．3.1 節で述べたように SVM の 1 段階目で用いる学習データは「(3)→(4)と切り替わった点から前後 4 フレームずつ」を対象としているが，これを

(イ) (3)→(4)と切り替わった点から前後 10 フレームまで拡大する

ことにし，より多くのデータを学習させた．学習パラメータは $C = 32$，$\gamma = 2^{-7}$ である．この学習モデル

で同様の実験を行った．また，(ア)と(イ)を組み合わせた方法による結果も含めて図 7 に示す．

(イ)の結果を見ると，(ア)のアプローチとは逆に(2)の F 値がさらに上がったが，(1)の F 値は大きく減少し全体的な性能向上には至らなかった．一方(ア)と(イ)を組み合わせた方法では，(2)の F 値を大きく変えることなく(1)の F 値を 0.1 程度上げることに成功した．しかし，(ア)＋(イ)の結果は(イ)で正解していた箇所の 27.1%（217/800）が不正解へ転じてしまった．これは見方を変えれば，発話継続/交替判定の位置が 1 フレーム異なるだけで大きく精度向上/精度低下するという知見を得たこととなる．原因として，発話終了判定のフレームシフトが 100 ms と大きすぎることにより正確な位置検出を妨げていることがあげられる．よって，10 ms などのさらに短いフレームシフトで判定を行うことにより，さらなる精度向上が見込める可能性がある．

5 まとめと今後の課題

機械学習器の SVM を 2 つ用いて音声の発話終了と発話継続/交替判別をリアルタイムに行える手法を提案した．SVM の 1 段階目では発話中か無音かどうかを判別し，発話終了地点であったとき 2 段階目の発話継続/交替判別を行う．

非リアルタイム環境におけるフレーム単位の評価実験を行った結果，発話継続，発話交替，発話中，無音の 4 クラス判別でそれぞれ F 値 0.51, 0.74, 0.99, 1.00 を得た．特に発話交替は高い再現率を示しており，ユーザの発話権譲渡を高精度にキャッチできる．

今後の課題として，2 段階目の SVM ではより時間的にロングレンジな素性を使い，さらに正確な区別ができるようにしたい．

なお，本手法は現時点の発話状態を判別するだけでなく，理論上は未来の発話状態の予測も可能である．データセット中のクラスのみを n フレーム分過去にずらせば，n フレーム分先の判別ができるからである．これにより，1.1 節で述べたような発話の先読み予測が可能となり，オーバーラップなどの人間らしい対話が実現できる．本手法で先読み予測実験も行い，判別精度を見てみたい．

最終的に，我々が過去の文献[3]で示したようなリアルタイムシステムを構築し，実システムでの挙動や性能評価も行いリズミカルな対話システムを作るための橋渡しにしたい．

6 参考文献

[1] 伊藤敏彦, 山田真也, 荒木健治: 音声認識・言語理解性能や状況の違いによるタスク指向音声対話の言語的･音響的特徴の比較, 日本音響学会誌, Vol. 63, No. 5, pp. 251-261 (2007).

[2] Sacks, H., Schegloff, E. A. & Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. *Language*, pp. 696-735 (1974).

[3] 伊藤敏彦, 小川翼: 機械学習を用いたリアルタイム発話継続､発話終了予測システム, 人工知能学会研究会資料. SIG-SLUD-B301-04, pp. 15-20 (2013).

[4] Ferrer, L., Shriberg, E. & Stolcke, A.: A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 608-611 (2003).

[5] Hariharan, R., Hakkinen, J. & Laurila, K.: Robust end-of-utterance detection for real-time speech recognition applications. In *Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 249-252 (2001).

[6] Edlund, J., Heldner, M. & Gustafson, J.: Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pp. 576-587 (2005).

[7] 大須賀智子, 堀内靖雄, 西田昌史, 市川熹: 音声対話での話者交替・継続の予測における韻律情報の有効性, 人工知能学会論文誌, Vol. 21, pp.1-8 (2006).

[8] 木村太郎, 堀内靖雄, 西田昌史, 市川熹: F0 モデルを用いた日本語対話における韻律と話者交替の分析, 電子情報通信学会技術研究報告. SP, Vol. 107, No. 282, pp. 25-30 (2007).

[9] Raux, A. & Eskenazi, M.: A finite-state turn-taking model for spoken dialog systems. In Proceedings of Human Language Technologies. In *the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 629-637 (2009).

[10] Sato, R., Higashinaka, R., Tamoto, M., Nakano, M. & Aikawa, K.: Learning decision trees to determine turn-taking by spoken dialogue systems, *INTERSPEECH* (2002).

[11] Nishimura, R. & Nakagawa, S.: A spoken dialog system for spontaneous conversations considering response timing and response type. *IEEJ Transactions on Electrical and Electronic Engineering*, No. 6, Vol. S1, pp. S17-S26 (2011).

[12] Funakoshi, K., Nakano, M., Kobayashi, K.,

Komatsu, T. & Yamada, S.: Non-humanlike spoken dialogue: a design perspective. In *The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 176-184 (2010).

[１３] Chang, C. C., & Lin, C. J.: LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27 (2011).

[１４] Lee, A., Kawahara, T. & Shikano, K.: Julius—an open source real-time large vocabulary recognition engine (2001).

[１５] 阪田真弓，広瀬啓吉: 対話音声の韻律的特徴の分析と合成，電子情報通信学会技術研究報告. SP, Vol. 95, No. 42, pp. 55-62 (1995).

[１６] 大野誠寛，神谷優貴，松原茂樹: タグ付けの安定性を備えた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，電子情報通信学会技術研究報告．音声, Vol. 110, No. 357, pp. 19-24 (2010).

[１７] N.ワード: 発話の中にピッチが低い領域があったらあいづちを打つ．情報処理学会研究報告．SLP, Vol. 96, No. 55, pp. 7-12 (1996).

[１８] 小磯花絵，堀内靖雄，土屋俊，市川熹: 先行発話断片の終端部分に存在する次発話者に関する言語的･韻律的要素について，電子情報通信学会技術研究報告. NLC, Vol. 95, No. 600, pp. 25-30 (1996).

[１９] Sjolander, K. & Beskow, J.: Wavesurfer-an open source speech tool. *INTERSPEECH*, pp. 464-467 (2001).

[２０] 田中和世，速水悟，山下洋一，鹿野清宏，板橋秀一，岡隆一: RWC 計画における音声対話データベースの構築，情報処理学会研究報告. SLP, Vol. 11, No. 55, pp. 37-42 (1996).

[２１] Chawla, N. V., Japkowicz, N., & Kotcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explorations Newsletter*, Vol. 6, No. 1, pp. 1-6 (2004).