日常会話における発話タイミング分析

芦村 和幸[1]　　ニック・キャンベル[1,2]　　武田 一哉[3]

[1] 科学技術振興事業団 CREST
[2] 株式会社国際電気通信基礎技術研究所
[3] 名古屋大学

あらまし　本稿では, 日常会話において, 発話の生成時に生じる話者交代を分析するためのフレームワークを提案する. このフレームワークは, 人=コンピュータ間の音声対話インタフェースに必要なアーキテクチャを説明するものであるとともに, 対話において見られる発話開始の遅延や, 発話の overlap について, 各話者の発話単位ごとに定義されるダンピング要素を用いて説明するためのメカニズムを提供する. 我々は, まず, 電話対話収録から抽出したデータの分析から得た最初の結果を示した上で, その分析結果から得た知見を「会話分析」の理論と関係づける.

Analysis of utterance timing in everyday conversation

Kazuyuki ASHIMURA [1]　　Nick CAMPBELL [1,2]　　Kazuya TAKEDA [3]

[1] CREST, Japan Science and Technology Corporation(JST)
[2] Advanced Telecommunications Research Institute International
[3] Nagoya University

Abstract　This paper presents a framework for the analysis of turn-taking in speech timing. It describes an architecture for a human-computer spoken-dialogue interface, and presents a mechanism to account for the delays or overlaps in speech turns by means of a single damping factor per speaker-utterance unit. We present initial results from an analysis of data taken from recorded telephone conversations and relate the findings to current theories of conversation analysis.

1　はじめに

SF の世界では, ロボットが遵守すべき大原則として, アイザック・アシモフの「ロボット工学の三原則」が定義されてきた.

我々は, 今のところ, コンピュータを「いわゆるロボット」としてではなく, 「大量の情報を管理・検索するためのシステム」ととらえているが, コンピュータのインタフェースを改善していくために, まず, ロボット工学三原則を参考に, 「対話型インタフェースの三原則」を定義した上で, これにもとづいてコンピュータインタフェースの改善方法について検討を進める.

対話型インタフェースの三原則

第一条: 人間が必要とする情報を, 時, 場所, 状況に応じて, 的確なタイミングで提供する.
第二条: 人間同士のように, 楽しく気楽なやりとりにより操作できる.
第三条: 人間に対して, 必要以上にくどくなったりし過ぎず, 適度な節度を保つ.

今までにも, コンピュータのインタフェースをより親しみやすいものとするために, 音声をインタフェースに用いた対話的情報検索システムが提案されてきているが, 音声認識技術を利用する場合, 目的タスクに依存した認識用モデルが必要なこともあり, 利用者ひとりひとりの状況や要求に合わせたシステムの構築は困難である [1].

本稿では, 音声インタフェースを利用した情報検索システムを「ひとりひとりにとって使いやすくすること」を目標に, 日常会話における発話タイミングの分析を行ない, 今まで見過ごされがちであった発話のタイミングについて考察する. これを通して, 話者交代を円滑に行ない, 情報をよりわかりやすく提示するためのモデルおよびシステムを提案する.

2 日常会話の発話タイミング

話し手の発話には, 言語情報の他に, 感情, 意図, 態度などが含まれており, それらは文字上の意味を越えた心理的な意味合いを伝達している. 特に, 日常会話においては, 発話の内容のみならず, 発話が生成されるタイミングが, 大きな意味を持っていると思われる.

2.1 発話のタイミング

本稿では,「発話のタイミング」という用語を, 日常の会話における,「話し手の発話開始時刻」と「聞き手 (=次の話者) の発話開始時刻」との相対的かつ時間的な位置関係としてとらえる. この関係は, 大きく, 以下の 2 種類に分けられる.

発話間に overlap がない場合: 話し手の発話と, 聞き手の発話の間にポーズが知覚される.

発話間に overlap がある場合: 話し手の発話の途中で, 聞き手の発話が始まる.

2.2 発話タイミングを考慮した対話型インタフェース

「対話型インタフェースの三原則」に示したような改善を行なうにあたっては, 発話タイミングを考慮した対話型インタフェースを構築する必要がある. そのために, 実際の日常会話データを分析しタイミング制御のモデルを作成するとともに, データの蓄積・分析・検索を統合的に扱うための, 以下のようなツール群を作成する必要がある.

発話タイミング制御: コンピュータとの対話が, 日常会話同様に自然なものとなるよう,「発話タイミング制御モデル」にもとづいて発話タイミングを制御.

ハードリアルタイム制御: 入力音声波形から, リアルタイムに発話と無音を弁別し, 発話開始時刻を取得. また, 最終的な出力音声波形を, 対話の環境条件に応じた適切な時刻に出力.

音声ジュークボックス: 音声データを話題や内容に応じて分類した上で, さまざまな付加情報とともに蓄積. 付加情報を利用しながら, データ探索範囲を動的に切替え, 目的や状況に応じた最適データを正確に検索.

音声特徴量抽出: 入力された音声データから, 検索のキーになる特徴量を抽出.

回答作成: 音声ジュークボックスにより, 利用者から要求された情報を検索. 出力タイミング情報を, 発話タイミング制御へフィードバック.

音声再合成: 検索結果を, 一続きの音声データとして再合成.

上記ツール群の構成イメージを図 1 に示す.

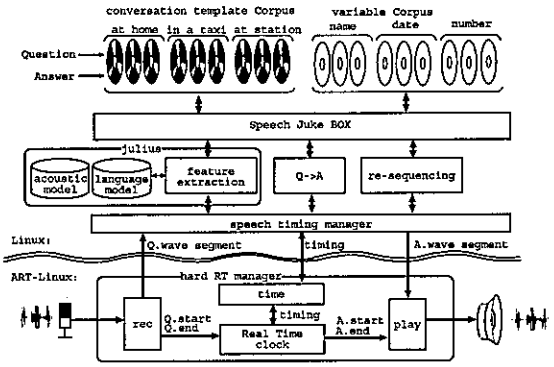

図 1: システム構成イメージ

本稿では, 上記のうち,「発話タイミング制御モデル」, および「ハードリアルタイム制御」について,「4 実データ分析とシステム化」および「5 発話タイミング制御モデル」にて詳述する.

3 発話タイミングのモデル化

3.1 「会話分析」の分析装置

実際に音声対話コーパスを作成し, 対話データを分析していくにあたっては, 以下のような問題について検討する必要がある.

- 音声対話コーパス中の音声データに対して, どのような情報を付加すればよいか.
- タスクや話題に応じて, 情報検索時の探索範囲を動的に切替えるためには, どのようなデータ構造や探索アルゴリズムが必要か.

そのために,「会話分析」の基本的な分析装置である, 隣接ペアと話者交代規則の概念を導入する. 以下では, 隣接ペアと話者交代規則について簡単に説明する.

3.1.1 隣接ペア

隣接ペアとは, 以下の 4 条件を満たすような 2 つの発話 X, Y の連続である [2].

1. X と Y は隣接した位置にある.
2. X と Y は異なる話者が産出する.
3. 隣接ペアの第 1 部分 X は, 第 2 部分 Y に先行する.
4. X は Y を特定化する (X は決まった型の Y を要求する (質問に対する返答など)).

隣接ペアの例 (挨拶に対する挨拶))

X: 久しぶりですー
Y: そーですね

隣接ペアの第 2 部分における発話タイミングのずれは,「条件的な不適切性」[1] を示していると考えられるため, 聞き手の返事のタイミングが遅過ぎる場合, 話し手にとって「聞き手の返答の不在」として否定的な含みを持つ [2].

3.1.2 話者交代規則

話者交代は, 以下の規則に支配される.

1. 現在の話者 C のターン[2] T について, 最初の TRP [3] x で, 以下の (a)~(c) を適用する.
   (a) C が, 次の話者として N を選ぶように T を構成していたならば,N は次にしゃべる権利と義務を得, x においてターンの移行が生じる.
   (b) C が,(1a) を使わなかったならば, C 以外の誰でも, 次の話者として名乗りをあげることができる. 最初に口を開いた人が, 次にしゃべる権利を得, x においてターンの移行が生じる.
   (c) C が,(1a) を使わず, かつ, 他の誰も (1b) を使わなかったならば, C はしゃべり続けることができる.
2. x において (1a) も (1b) も適用されず,(1c) によって C がしゃべり続けたならば,T 中の次の TRP x' に対して 1 を繰り返す.

これを話者の移行が実行されるまで繰り返す.

[1]「条件的な適切性」とは, 隣接ペアの第 2 部分 Y が, 第 1 部分 X にとって要求される型の発話として適切, かつ予期できるということである.

[2]話者交代の基本単位. 一人の話者がしゃべり続ける.

[3]移行適格場所. ターン構成単位 (TCU) の終了地点.TCU は話者がターンを構成する際に選択する単位で, 文脈に応じて, 文・節・句・語などにより構成される.

3.2 分析装置にもとづく対話のモデル化

3.2.1 基本モデル: 対話の局所的構造

隣接ペアは, 「質問と返答」など, 対話の参与者にとってもっとも基本的な相互行為を達成する発話対であり, 話者交代を中心として, 隣接ペアの第 1 部分と第 2 部分により, 一つの話題に関する局所的な構造が構成されていると考えられる [2].

第 1 部分-< 話者交代 >-第 2 部分

図 2: 対話の局所構造

隣接ペアの条件的適切性や, 話者交代規則を考慮しながら人とコンピュータの対話を構成していく場合, 「いつ, いかなる状況において, 発話を開始, 継続, 終了するべきか」という発話タイミングの問題を明確化する必要がある. そのために, 人間同士の対話において, どのような場合に発話を開始してよいか, を見極める必要がある.

3.2.2 構文論から見た次話者行動パターン

日本語のような SOV 型言語 (トムが りんごを 食べる.) においても, 英語のような SVO 型言語 (Tom eats apples.) と同様に,「発話内容の構文的な構造」に着目し発話内容を予測することによる話者交代が可能であり, 次話者による発話開始の行動パターンモデルとして, 図 3 のようなモデルが提案されている [3].

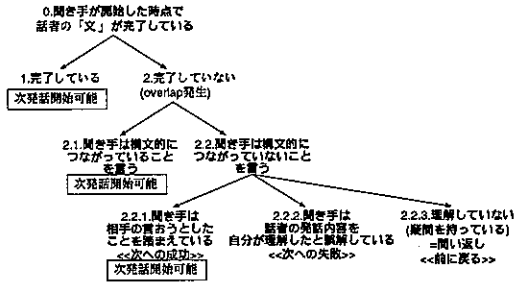

図 3: 次話者行動パターンモデル ([3] 試案を修正)

また, 構文的に可能な割り込み位置として, 表 1 のような位置が提案されている [3].

表 1: 構文的に可能な割り込み位置 ([3] をもとに作成)

割り込み位置	例
複文の接続詞や接続助詞の前後	接続助詞 (~/から/~,etc.)
	接続詞 (~/けれど/~,etc.)
題述構造の前項と後項の間	名前は/花子だ, ボチは/犬だ
ターン末	連続する終助詞の途中 (です/ねえ)
	「です」,「ます」の直前 (描いたん/です)
呼応の核要素の間	みとめ方のモダリティ(何も/ない)
	テンスのモダリティ(そのあと/なった)
	述語が予測可能 (歌が/出てくる)
	述語の前までで構文が完成 (300 人以上/いた)
倒置された述語の後	復習・予習なんですよ/順番から

※注意: 例中の「/」が割り込み可能位置.

4 実データ分析とシステム化

4.1 対話音声データの分析

4.1.1 収録条件

実際に収録した対話音声データをもとに, 書き起こしテキストの作成および分析を行なった.

我々は, 現在, 話者の違いや収録時期に応じたさまざまなバリエーションの対話データを収集することを目標に, 2001 年 12 月より 10 週間の計画で, 対話音声を収録中である. 毎週の収録においては, 10 人の被験者を組合わせることにより, 30 分の対話を 12 組収録している. 今回のデータ分析にあたっては,「収録にある程度慣れたが, まだ友人というほどは親しくない話者同士」という観点で, 第 2 週目の収録結果 30 分を対象としてタイミング分析を行なった. なお, 今回の収録被験者は,22 歳男性 (話者 X) および 42 歳男性 (話者 Y) の二人である.

4.1.2 分析結果概要

収録結果を概観したところ, 以下のような傾向が見られたため, 主として, 話者 X から話者 Y への話者交代に着目して分析を行なった.

- 話者 X から, 新しい話題が始まることが多い.
- 一方, 話者 Y は, あいづちや応答が多い.

サンプルデータ中に含まれる話者 Y のターン数は,「話者交代規則 (1a)」によるものが 457 ターン,「話者交代規則 (1b)」によるものが 5 ターンであった. なお,「話者交代規則 (1a)」によるターン取得は, 今回のサンプルデータにおいては, 話者 X からの話者 Y へ向けられた質問に対する応答などであり,「話者交代規則 (1b)」によるターン取得は, 話者 Y が独自に新しい話題を始める場合であった.

4.1.3 発話完了と次発話開始の間隔

話者 X から話者 Y への話者交代において, 3 回以上見られた発話内容について, 話者 X の発話完了と話者 Y の発話開始の時間間隔の, 平均と標準偏差を測定した (表 2).

表 2: 話者 X の発話完了と話者 Y の発話開始の間隔

overlap の有無	内容	件数	平均 (sec.)	標準偏差 (sec.)
overlap なし	overlap なし全体	189	0.519	1.532
	あいづち	75	0.329	0.504
	自発話の継続	89	0.672	2.160
	評価-同意ペア	6	0.431	0.311
	質問-応答ペア	16	0.588	0.633
	返事に困る内容	3	0.521	0.455
overlap あり	overlap あり全体	191	-0.448	1.801
	あいづち	140	-0.406	2.054
	自発話の継続	18	-0.367	0.507
	評価-同意ペア	3	-1.549	0.899
	質問-応答ペア	8	-0.235	0.193
	末尾での重複	79	-0.469	0.643

4.1.4 overlap の有無に応じた発話タイミング

サンプルデータ中の発話を, 話者交代時の overlap の有無により分類し, 発話開始位置と発話タイミングについて分析した結果を以下に述べる.

overlap がない場合

発話開始位置 話し手の発話を最後まで聞き, 意図を理解した上で, 内容や状況に最適なタイミングまで待ってから発話を開始する (図 4).

((映画の話を開始するにあたって、相手に確認))

→ X: 映画はすきですか-?
(0.786)
→ Y: やー、僕?

図 4: overlap がない場合の書き起こし例

発話タイミング 話し手の発話と, 聞き手の発話との間のポーズ長により, タイミングが定義される (図 5).この場合の次発話開始タイミングは, 図 3 の「1. 完了している」に相当しており, 実際の対話においてもよく観察される.

overlap がある場合

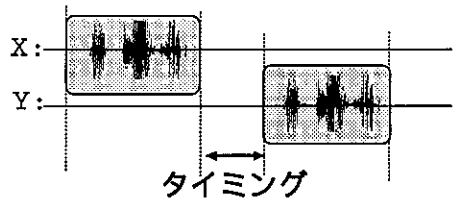

図 5: overlap がない場合の発話タイミング

発話開始位置　話し手の発話完了を待たずに聞き手が発話を開始する (図 6).

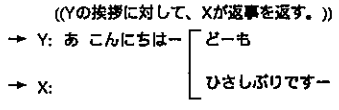

図 6: overlap がある場合の書き起こし例

発話タイミング　話し手の発話の途中で, 聞き手の発話が始まるため, 単なる発話間の間隔ではなく, 「話し手の発話開始点」を起点として, 「話し手発話の, どの発話内容に対応する時刻に, 聞き手の発話が開始されるか」により聞き手の発話タイミングが定義される (図 7).

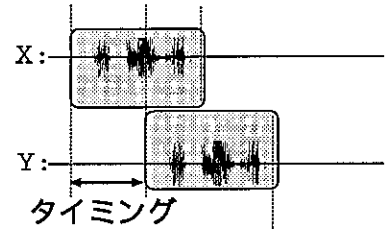

図 7: overlap がある場合の発話タイミング

今回のサンプルデータにおいては, overlap がある場合でも, ほとんど全ての次発話開始タイミングが, 図 3 の「2.1. 聞き手は構文的につながっていることを言う」もしくは「2.2.1. 聞き手は相手の言おうとしたことを踏まえている」に相当しており, 対話の流れに問題なく割り込みを行なえていた. ただし, 録音終了直前に, 発話の overlap が頻繁に発生した際, 話者 Y の発話が, 話者 X の質問に対する応答なのか, 話者 Y 自身の独白なのかを判定しかねるケースが一件だけ見られた.

なお, 一人の話者が長い間しゃべり続ける場合, 次の話者交代を円滑に行なうために, 相手に「割り込むすき間」を提供する必要がある. 例えば, 今回のサンプルデータにおいては, 以下のような現象がみられた.

- 文章を, フレーズなどの短い単位に区切る.
- フレーズ末尾の母音や/s/などを長く伸ばす.

4.2 リアルタイム OS の利用

4.2.1 ART-Linux

発話タイミングの制御モデルにもとづいて, 適切なタイミングで音声情報を生成するためには, 入力音声および出力音声の開始時刻を, 高精度かつリアルタイムに制御する必要があるが, このような処理は通常のマルチタスク OS では不可能なため, Linux の RealTime 拡張である ART-Linux を利用した. ART-Linux は以下のような特長を持つ [4].

- 既存のドライバやアプリケーションが流用可能.
- リアルタイムタスクをユーザモードで利用可能.
- 3 つの API でリアルタイム処理を制御可能.

4.2.2 時間制御の精度改善

クロック周波数 645MHz の PC 上において, C 言語により 1ms 周期のループ処理を実装し, 1 周期に要する時間をシステムクロックにもとづいて計測することにより, Linux および ART-Linux の時間制御の精度を確認した.

動作確認の結果, Linux では, 通常の負荷時でも 20ms 単位程度の精度しか得られなかった. 特に, カーネルコンパイルなど CPU 負荷の高い場合は非常に不安定であった. 一方,ART-Linux では, 常に安定して 1ms の周期が保たれていた (図 8, 図 9).

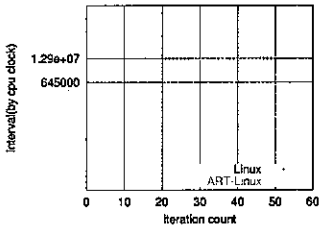

図 8: 通常の負荷の場合

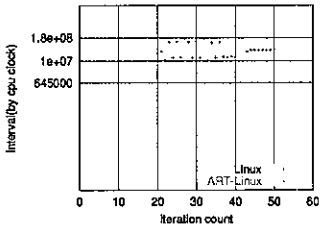

図 9: 大きな負荷をかけた場合

5 発話タイミング制御モデル

聞き手が話し手の発話完了を待つ場合, 聞き手は, 話し手が開始したコミュニケーションをテンポよく継続していくために, ある特定の間だけ待っているように思われる.

一方, 今回のサンプルデータでは, 対話終了直前において発話の overlap が頻繁に発生し, 2 話者が同時に別のことを発話するようなケースも見られた.

このような場合をも含めて, 発話タイミング制御を包括的に扱うためには, 基本的には独立した 2 つの主体を想定するとともに, ある場合には全く overlap なく話者交代が成立しあいながら, またある場合にはタイミングのずれにより完全に発話が重複することも可能なモデルが必要である.

そのようなモデルとして, 本稿では, 自動車の 2 気筒エンジンにおいて, 2 つのピストンがタイミングよく往復運動を繰り返すことによりクランクの回転が継続していくモデルを提案する (図 10).

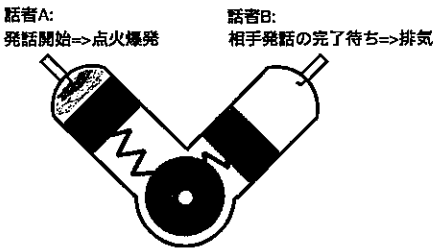

図 10: 対話の 2 気筒エンジンモデル

聞き手が話し手の発話に割り込む場合, 話し手の発話の途中で, すでにその発話内容の全体を予測しているように思われる. さらに, 割り込みは, 「話し手が生成中の話」に対して, 発話内容の補完の役割を果たしている場合がある (図 11).

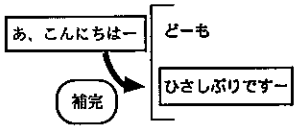

図 11: 聞き手の発話に対する補完

また, 今回のサンプルデータの中では, 話し手の次発話が, 音節や句の境界でなく/h/と/a/の境界などの音素境界, あるいは音素内で発生する場合が 15 件見られた.

overlap のある場合の, このような微妙な発話開始位置を再現するためには, 図 10 の「対話の 2 気筒エンジンモデル」においては, ピストンとクランクを結合するコネクション・ロッドは, 通常のエンジンのように固いものではなく, 発話タイミングのゆらぎを吸収できるような「オイルにより動作をダンプされたバネ」であると考えられる.

今後, さらに大量の音声対話データを用いて, 音響情報, 時間情報, 言語情報などの既知のパラメータと, 今回提案したモデルのダンピング要素との対応関係について検討を進めていく.

6 おわりに

本稿では, 電話対話収録から抽出したデータの分析結果を示した上で, その分析結果から得た知見を「会話分析」の理論と関係づけた. また, これを通して, 日常会話における話者交代を分析するためのフレームワークを提案した. さらに, このフレームワークにより, 発話開始の遅延や, 発話の overlap など発話タイミングのずれについて, 各話者の発話単位ごとに定義されるダンピング要素を用いて説明するためのメカニズムを提案した.

今後は, 現在収録中の大規模対話音声データを用いて, 発話の overlap に対応するダンピング要素について, 分析・検討を深めていきたい. なお, 発話のタイミングは, 話し手の発話と聞き手の発話との相対的な関係であるため, 話者同士の発話速度を考慮しながら分析を進めていきたい.

参考文献

[1] Zue, V.: Conversational Interfaces: Advances and Challenges, Proc. Eurospeech '97, KN-9-KN-18, 1997

[2] 石崎雅人, 伝康晴: 談話と対話, 東京大学出版会, 2001

[3] 木田敦子, 乾裕子, 神崎享子, 高梨克也, 井佐原均: 構文論から見た対話–円滑な話者交替を可能にする構文構造–, 人工知能学会研究会資料, SIG-SLUD-A102-6(11/06), pp.33–38, 2001

[4] 石綿洋一: CQ 出版, Interface 1999 年 11 月号, pp.109–137, 1999