ヒューマンインタフェース 97−5
音声言語情報処理 40−5
（２００２．２．１）

自動応答音声における利用者適応の試み

中川 郷士，ナイジェル ワード
東京大学大学院 工学系研究科
機械情報工学専攻
{nakagawa,nigel}@sanpo.t.u-tokyo.ac.jp

あらまし： 電話を通して予め録音された音声で情報を提供するサービスが便利であるが，必ずしも利用者にとって提供される音声が最適であるとは言えない．本研究では，利用者に合わせた音声スピードの調節によりこのような自動音声応答システムの提供音声を聞きやすくするのが目的である．擬似電話番号案内コーパスを収集，分析し，利用者の反応の速さと発話速度から適応させるスピードパラメータを求める予測式を重回帰分析によって得た．得られた予測式とコーパスにおける実測値の相関係数は 0.46 となった．得られた予測式をもとに疑似電話番号案内システムを構築し，被験者 18 人による評価実験を行い，提供音声が一様なシステムに比べ，利用者適応を行うシステムの方が利用者に好まれることがわかった．
キーワード： ユーザインタフェース， 音声対話， 韻律情報， コーパス分析， 発話速度， 自動音声応答システム

Adaptive Number-giving for Directory Assistance

Satoshi Nakagawa, Nigel Ward
Department of Mechano-Informatics,
School of Engineering,
University of Tokyo

Abstract: Today many services exist which provide information over the phone using a prerecorded or synthesized voice. These voices are invariant in speed. However people giving information over the telephone tend to adapt the speed of their presentation to suit the needs of the listener. This paper presents a preliminary model of this adaptation. Analysis of a corpus of simulated directory assistance dialogs revealed correlations between the operator's speed in number-giving and the speed of the user's response and the user's speaking rate. A predictive formula gives speeds that predictions correlate well (.46) with the speeds observed in the corpus. Experiments with 18 subjects suggest that users prefer an adaptive system.
Keywords: User Interface, Voice Dialogue, Prosody, Corpus Analysis, Speaking Rate, IVR

1 はじめに

あらかじめ録音された音声を用いて電話上で情報を提供する様々なサービスがあり便利である．例えば航空券の電話予約サービスや，住所から電話番号を調べる電話番号案内サービスから総合ボイス・ポータルまである．

こういったサービスにおける，典型的な会話には 1) かかってきた電話の振り分け，2) 利用者の情報（名前や顧客番号など）の獲得，3) 顧客の要望の詳細についてのやりとり，という三つのフェーズはそれぞれ自動化の研究・実装がたくさんある．

オペレータ (o_1): 104 の鈴木です.
利用者 (u_1): あ，もしもし．東京都文京区の東京大学の代表の電話番号知りたいんですけど.
o_2: 東京都文京区の東京大学の代表の電話番号ですね？
u_2: はい.
o_3: ご案内致します…
o_4: **03 3812 2111** です.

図 1: 疑似電話暗号案内コーパスの対話例

しかし第4フェーズ，すなわち情報伝達フェーズ,はあまり注目されてこなかった．我々が知る限りでは，全ての Interactive Voice Response システム，音声対話システムでは録音された固定音声か，固定パラメータによる合成音声が用いられているのが現状である.

提供音声が一定のスピードであれば，例えばスピードが速すぎたり，逆にゆっくり過ぎると感じる利用者もいるだろう．さらに，時間的な効率を考えると，提供音声が遅すぎると利用者にとっては時間の無駄になる．逆に提供音声が速すぎると，聞きとれない場合があり得る．本研究はこうした利用者への不適応を改善することが目的である.

2 利用者適応

システムの出力を利用者個人個人にいかに適応させるかという問題は人工知能の伝統的な研究課題である．そこではユーザ適応はユーザの考え，要望，知識などをモデル化することにより，利用者個人にもっとも有益な情報を提供することを目指す．自然言語処理においては，自然言語生成という分野が，いかに利用者が理解できる意味構造や単語でメッセージを表現するかという問題に関連がある.

このように出力の内容を利用者に適応させるという研究は行われてきた．一方で出力の方法を利用者に適応させる研究は近年になり行われるようになった。

Schmandt[1] は言語情報を使わずに，韻律情報のみを使った道順案内システムユーザの発話を質問とあいづちの 2 つのカテゴリーに分類して，それに対しシステムは，案内を次へ続けるか，もう一度さっき言った道順を繰り返すか決定する．岩瀬ら [2] もユーザに合わせた自然な対話ペースで道案内をするシステムを作成した．日本語において韻律情報を用いることによってあいづちなどの適切なフィードバック，適切な発言のタイミングを達成した.

Tsukahara ら [3] は利用者の刹那的な感情の推察システムを作成した．利用者の発話タイミング・発話の韻律情報を利用し利用者に適したあいづちを選択し生成した.

これらの関連研究のように韻律情報を用いて有効な内容，適切なタイミングによる会話のやりとりを目指した研究は行われているが，適切な発話ペースというものを考慮したシステムは行われていない.

3 分析

3.1 コーパス

自動応答音声が用いられている対話において応答音声を利用者に適応させる場面はいくつか考えられるが，データ収集を簡便にするため電話番号案内の対話を分析対象とし，疑似[1]電話番号案内コーパスを収集した．電話番号案内の対話は短く，様々な年齢，性，職業の利用者，さまざまな環境における対話を簡単に集めることができた．この電話番号案内コーパスの対話例を図 1 に示す.

ここで，現在利用されている電話番号案内サービス（図 2 参照）では最後にオペレータが番号を案内している部分（o_4）が実際に自動応答音声で案内されていることに言及しておく．つまりこの部分を利用者に適応させることが目

[1] 実際の電話番号案内サービスのコーパスを収集するにはオペレータの声のプライバシーなどの問題が発生する

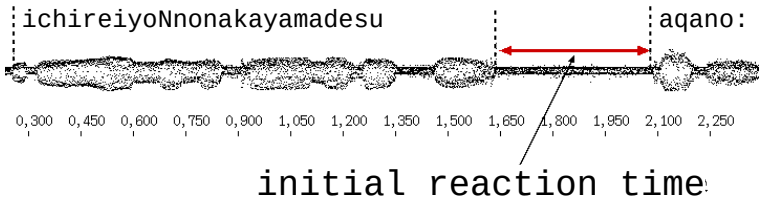


図 3: 第一反応時間:オペレータが始めに発話し，その発話が終了してから利用者が発話を開始するまでの時間

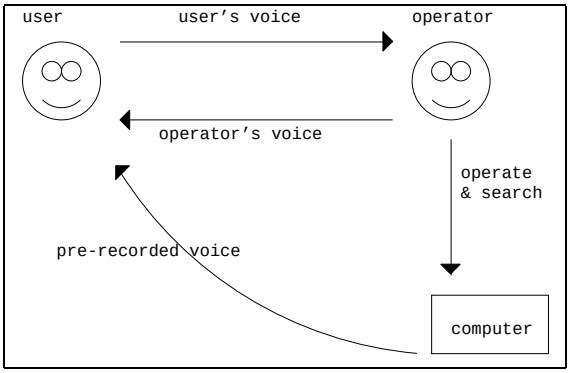


図 2: 現在半自動の電話番号案内

的である.

コーパスは 57 人の利用者と 8 人のオペレータによる会話から収集された．一人の利用者につき 1 対話から 10 対話が収集され，全部で 508 の対話が収集された.

本研究ではコーパスを収集する際に利用者に対してアンケートを行っており，そのアンケートにおいて利用者がオペレータの評価をしている．この評価が良い対話（以下このコーパスを良評価コーパスと呼ぶことにする）について分析することにした．良評価コーパスは全部で 142 対話分である.

3.2 仮説

コーパスを聞くことによって以下の仮説を立てた.

- ゆっくり話している利用者にはゆっくりした案内の方が好まれ，速く話している利用者には速く案内した方が好まれる
- 利用者の反応が遅い（オペレータの発話終了から利用者の発話までの時間が長い）場合はゆっくりした案内の方が好まれ，反応が速い場合は速めに案内した方が好まれる

という二つの仮説である.

この仮説を検証するため，コーパスから

- 「第一反応時間」つまりはじめのオペレータの挨拶 (o_1) から利用者が発話する (u_1) までの時間（図 3 参照）. これは 500 ミリ秒前後は普通だが,50 から 1600 ミリ秒までのバラツキがある.
- 利用者の発話速度. これは 8 モーラ/秒 前後が多いが,6 から 10 のバラツキがある.

を抽出し，オペレータが案内するスピードとの相関を調べることにした.

3.3 純粋伝達時間と総伝達時間

オペレータが案内する電話番号は全て 10 ケタの番号であるので案内するスピードを測るにはオペレータが案内する時間を測定すれば良い.

ここでオペレータが案内する際，利用者の反応が 4 種類あることがわかった．それは: 番号を途中で復唱する, 一部復唱する, あいづちを打つ, と何も発話しないというパターンである.

良評価コーパス 142 対話のうちの内訳は 75 対話はあいづち込の種類で, それを分析対象とすることにした．利用者の「はい」というあいづちを含めたオペレータが番号を案内をしている時間を測定した．これを「総伝達時間」と定義する.

オペレータの発話部分のみを取り出し，それを伝達スピードを示すパラメータ（純粋伝達時

間と定義する）とすることも可能であるが,「間」の重要性を考慮し，総伝達時間をパラメータとして採用した.

3.4 発話速度

利用者の発話速度として聞き取りによって求めたモーラ速度を採用した．本研究ではモーラ速度を，聞き取った文字列（音素表記）から「a」「i」「u」「e」「o」，促音「q」（カナで「っ」），「N」（カナで「ん」），と長母音を伸ばす「:」（カナで「ー」）をとりだし，その個数を発話時間で除算した値と定義する.

4 分析結果

第一反応時間と総伝達時間の関係はやや相関がある（相関係数 0.32）ことがわかった.

発話速度と総伝達時間の関係もやや相関がある（相関係数 -0.25）ことがわかった．発話速度の計算においては，250msec 以上続く言い淀みと沈黙の部分を計算から除外した.

さらに，重回帰分析により総伝達時間を求める予測式を導いた.

聞き取りによる話速を $R[morae/sec]$，第一反応時間を $D[msec]$，総伝達時間の予測値を $L[msec]$ とすると重回帰式は

$$L = m_1 R + m_2 D + b \qquad (1)$$

となる.

重回帰分析により，係数・定数はそれぞれ $m_1 = -355.95[sec{\cdot}msec/morae]$, $m_2 = 1.50[]$, $b = 9048.25[msec]$ となった．また重相関（予測値 L と観測値との相関係数）は 0.46 となり，かなり相関があるという結果が得られた．分析によって得られた式は有意であった（$p < 0.01$）.

5 システム

分析によって得られた式の有効性を検証するために，擬似電話番号案内システムを作成した.

システムは利用者とオペレータの会話をリアルタイムで分析し，分析により得られたパラメータの値から予測式によって導かれたスピードで番号を利用者に案内するというものである．システムはオペレータのキューの後で利用者に番号を案内する.

5.1 ユーザーの発話速度の計算

重回帰分析で得られた予測式（1）には発話速度として聞き取った文章から得られるモーラ速度 R が採用されているが，番号案内のやりとりをしている最中に人間が発話を聞き取ってモーラ速度を求めることはできない．そのため実際のシステムにこの予測式をそのまま導入することはできない.

そこで発話速度を実時間で求めるために，Morgan らの mrate[4] という音響的な指標を用いることにした（良評価コーパスにおいて，音声認識エンジンの出力結果文字列によって得られるモーラ速度と聞き取り文から算出されるモーラ速度との相関係数は 0.04 となり，音声認識エンジンの出力結果によって得られるモーラ速度はシステムに実装できる指標とはならなかった）.

Morgan らによると，英語において，聞き取りによる話速との相関係数は 0.67 という値が得られている.

本システムでは日本語を対象としているので，自発的な連続自然発話コーパス [5] をもとに，日本語のモーラ速度と mrate との関係を調べた．このコーパスは既に聞き取り文が付属しているため利用した．男性 3 人によるコーパス（約 9 分，150 発話分）と聞き取り文を用いて mrate の値とモーラ速度を算出した結果，これらの値の相関係数は 0.68 となり，日本語においても mrate の値と発話速度の間にはかなり関係があることがわかった（図 4 参照）.

回帰分析によって以下の発話速度の予測式が得られた.

$$R = m_r M + b_r. \qquad (2)$$

ここで M は mrate の値であり，係数・定数はそれぞれ $m_r = 2.76[morae/sec]$, $b_r = -5.55[morae/sec]$ となった.

5.2 システムの発話速度の調節

一方で，案内音声には富士通製の音声合成エンジン [6] を利用した．合成音声を利用するこ

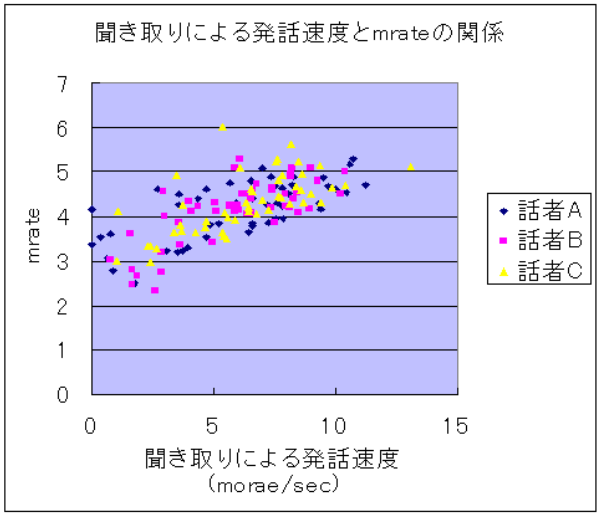


図 4: 聞き取りによる発話速度と mrate の値の関係

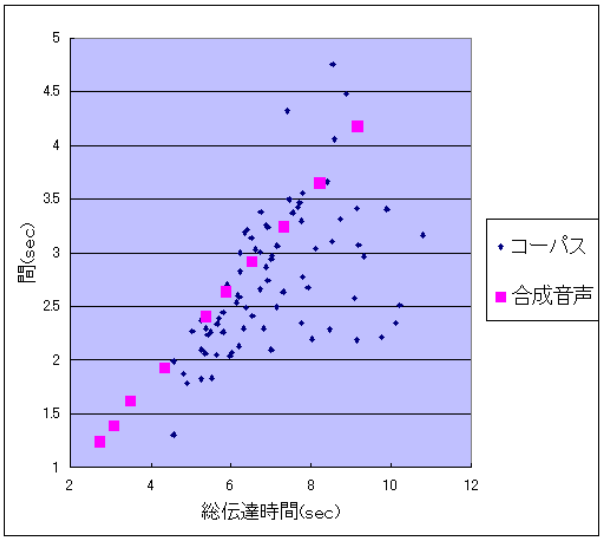


図 5: 合成音声と良評価コーパスにおける総伝達時間と間の関係

とでイントネーションを自動的に付加してくれる，スピード調節が段階的であるが容易である，というメリットがあるが，合成音声の不自然さというデメリットがある．

そこで，間の長さが総伝達時間の約 40%となるように案内部分の間を調節した．この 40%は，良評価コーパスの案内部分を分析して得られた平均値である．

音声合成エンジンは発話速度を 10 段階で操作可能である．10 桁の電話番号を間が全体の 40%になるように調節した合成音声から，合成エンジンの速度パラメータ S（0 から 9 の整数）と音声の全体の長さ L（総伝達時間）との関係式を求めた．

$$S = round(m_L L + b_L). \qquad (3)$$

ここで，$round$ 関数は小数第 1 位を四捨五入し整数値にする関数とする．

回帰分析により，係数･定数はそれぞれ $m_L = -0.001275[1/msec]$（$=-1.275[1/sec]$），$b_L = 12.432[]$ となった．

ここで良評価コーパスの案内部分を改めて分析したところ，速度レベル 0，7，8，9 は実際には現れることが少ない発話速度であることがわかったため（図 5），予測式で求められた S は 1 から 6 の範囲になるように絞った．

式 (1), (2), (3) によって得られる予測値と実測値との相関は相関係数 0.41 となった．

システムはオペレータと利用者の会話をまず分析し，第一反応時間 D と user の発話部分（250msec 以上の沈黙部分は除く）の mrate の値 M を取り出し,S を求める．

6 評価実験

被験者（20 歳代の学生 18 人，うち女性 2 人）を用意し，A，B，C という 3 つの擬似電話番号案内システムを用意した．被験者にはこれらのシステム A，B，C をランダムな順番に利用してもらった．実験は研究室内で，マイク付きヘッドホンを利用して行った．

システム A は提供音声の速度を変えないもの，システム B は本システム，システム C は本システムとは逆の方向[2]に予測値を求めるものである．

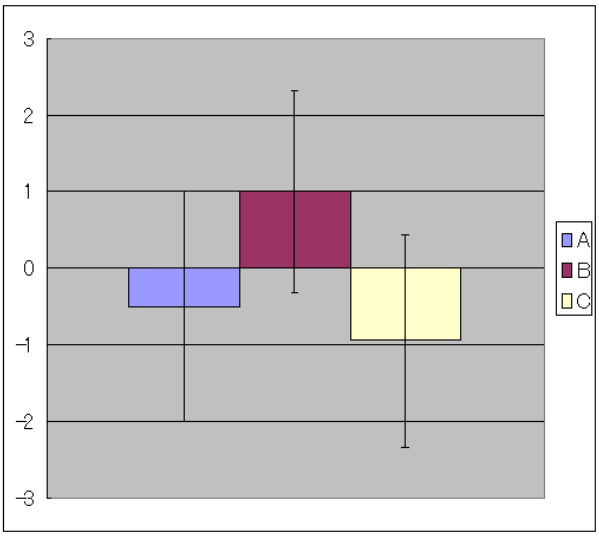
被験者には各システムごとに 3 回利用してもらった．1 回目は普通に問い合わせをし，2 回目は急いでいるつもりで問い合わせをし，3 回目は老人になったつもりで問い合わせをしてもらった．

[2]ゆっくり話している利用者には速く，速く話している利用者にはゆっくりと案内し，反応が悪い利用者には速く，反応の良い利用者にはゆっくりと案内する方向

被験者には 10 個の問い合わせ先のリストが渡され，その中から問い合わせ先をランダムに選んで問い合わせる．システムを利用する際，電話の呼び鈴が 1 回鳴り，まずオペレータ（今回の実験では筆者が担当）が応答する．そこで被験者が選んだ問い合わせ先をオペレータに伝え，オペレータは問い合わせ先の確認をし，「ご案内致します.」と言ってシステムにキューを出す．システムからは電話番号が 1 回だけ案内される（ここでの会話例は図 1 を参照，図の o_4 の部分がシステム出力となる）．

被験者は案内された番号を書き取り，それぞれの案内に対してその案内の速さが適当であったかどうか 9 段階（–4 点から 4 点まで）で評価する．3 回目は老人になったつもりで評価をしてもらった．さらに，3 回を通したシステムの総合的な評価も同様にしてもらった．なお，評価の際には合成音声の不自然さは評価の対象外としてもらった．

適応しないシステム（A）より適応するシステム（B）の間, Wilcoxon の符号付順位和検定を行ったところ適応する方を好むという結果が得られた（$p < 0.01$）．逆システム (C) の評価が一番低かった．



7　おわりに

本研究では擬似電話番号案内コーパスを分析することによって利用者に適応した速度で案内するシステムを作成し，そのシステムの有効性を被験者 18 人による評価実験によって検証した．

利用者に提供音声を適応させるために，パラメータとして利用者の第一反応時間と発話速度を採用したが，その他にも言い淀みや沈黙時間が利用できる可能性も考えられ，今後の課題と言える．韻律情報以外の情報（声から推定される利用者の年代，母国語，出身等) も利用者適応のためのパラメータとなる可能性もある．また，同じ利用者においても環境，電話回線の音質，時間帯などによっても好むスピードが違う可能性もある．また, 今回は便宜的に予測式として基本的に線形のものを採用したが，より最適な他の予測式もありえる．

本研究は，提供音声のスピードにおける利用者適応の試みであり，本研究によって得られた知見は幅広く利用できるだろう．

謝辞　財団法人国際コミュニケーション基金および広瀬啓吉代表の「韻律」科研費プロジェクトを感謝いたします．

参考文献

[1] Christopher Schmandt: *Voice Communication with Computers*, VNR Computer Library, pp.199-204 (1994).

[2] Tatsuya Iwase and Nigel Ward: Pacing Spoken Directions to Suit the Listener, *International Conference on Spoken Language Processing(ICSLP-98)*, pp.1203-1206 (1998).

[3] Wataru Tsukahara and Nigel Ward: Responding to Subtle, Fleeting Changes in the User's Internal State, *CHI 2001:Conference on Human Factors in Computer Systems*, pp.77-84 (2001).

[4] Nelson Morgan and Eric Fosler-Lussier: Combining Multiple Estimators of Speaking Rate, *ICASSP-98*, Seattle, pp.721-724 (1998).

[5] CD-ROM 平成 7 年度　文部省科学研究費補助金重点領域研究, 音声・言語・概念の統合的処理による対話の理解と生成に関する研究, 対話音声コーパス Vol.4 (1995).

[6] Linux 版日本語音声合成ライブラリー, http://www.createsystem.co.jp/linux.html.