音声言語情報処理　12－14
（１９９６．７．２７）

音声認識合成による応用構築の容易な電話音声対話システム

北井 幹雄[1]，山田 智一[1]，塚田 元[2]，嵯峨山 茂樹[1]
[1] ＮＴＴヒューマンインタフェース研究所 〒238-03 横須賀市武 1-2356
[2]ATR 音声翻訳通信研究所 〒619-02 京都府相楽郡精華町光台 2-2

あらまし　本稿は，次の４つの特徴を持つ音声対話実験システムについて述べている．（１）電話網を介して収集した約１万人の音声データを使って学習した音素環境依存型音素ＨＭＭモデルによる不特定話者大語彙音声認識機能，（２）トライホン環境に依存した音素波形セグメントの接続により高品質な音声を合成するテキスト音声合成機能，（３）音源カードと音声モデムを搭載したＰＣ以外に特別なハードを必要としないソフトウェアベースのシステム構成，および（４）２種類のサービスシナリオを記述するだけでシステムの構築を可能にする容易で且つ迅速なプロトタイピング環境．

キーワード　音声対話，音声処理システム，電話音声認識

Experimental Voice Interactive System for Telephone Applications with Speech Recognition and Synthsis Functions

Mikio KITAI[1], Tomokazu Yamada[1], Hajime Tsukada[2], and Shigeki Sagayama[1]
[1]NTT Human Interface Labolatories
1-2356 Take Yokosuka-Shi Kanagawa 238-03 Japan
[2]ATR Interpreting Telecommunications Research Laboratories,
2-2, Hikaridai, Seikachoo, Souraku-gun, Kyoto, 619-02 Japan

Abstract　This paper describes an experimental interactive system featuring (1) high accurate speaker independent and large vocabulary speech recognition based on context-dependent accurate acoustic phoneme HMM models trained with speech data from more than 10,000 speakers collected over telephone network, (2) high quality text-to-speech synthsis that generates speech by concatenating triphone-context-dependent waveform segments, (3) software-based configuration that requires no special hardware except a PC equipped with a sound board and a voice modem, and (4) easy and rapid prototyping which enables the developer to build a system by writing two types of service scenarios.

key words　speech dialogue, dialogue system, voice activated system

1 はじめに

電話サービスへの音声認識技術の適用は従来から各研究所などで検討され続けて来たが，ANSERサービス以来，一般のユーザが使用するサービスに音声認識が適用されることはここ1年程前までは殆んどなかった．しかし昨年以来，少しずつではあるが実際のサービスに音声認識が使われ始めている．例えば，ＫＤＤの悪戯呼撃退サービス [1] や，ＮＴＴで昨年度から開始されている鳥や蝉の鳴き声を案内するサービスなどである．近い将来に予定されているサービスとしては，三菱電機の電話音声認識ユニットを用いた観光地の天気予報案内サービスなどがある．

近年，音声認識合成機能が実サービスに使われ始めた理由としては，以下が考えられる．

(1) 電話網を介して収集した大規模音声データの学習による，仮名による語彙変更が可能な不特定話者電話音声認識の性能向上，
(2) 波形合成法によるテキスト音声合成音の質の向上，
(3) 計算機の高速化，低価格化，高機能化による，ＰＣ上で実時間で動作する音声認識，合成ソフトウェア構築および利用の容易化，および
(4) ＡＰＩの改良．

電話網を介した音声データの収集で有名なのはＴＩによる “Voice Across Japan” であるが，三菱電機などでも独自で大規模な音声データベースを構築しており，電話回線を経由してHMMにより不特定話者音声認識を行なう場合，大規模なデータベースの構築は必須と言える．

ＡＰＩの重要性やラピッドプロトタイピングの重要性は，東芝の新田ら [2] によって従来から指摘されている．ＮＥＣの磯ら [3] は，ＰＣの環境上で，音声認識合成を利用したアプリケーションを比較的容易に構築できる環境を開発している．

本稿では，我々が開発した，特別なハードウェアを必要とせず，音声認識，合成に関する特別な知識がなくても簡単に且つ迅速に電話応用システムの作成が可能な，対話システムについて述べる．

2 システム概要

2.1 ハードウェア構成

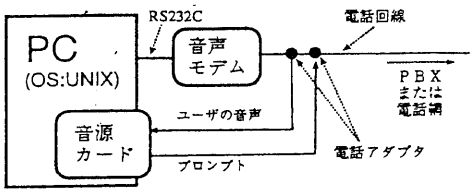

図 1: ハードウェア構成

本システムは，図1に示すようにパソコン (Pentium 90MHz 以上，メモリ 10MB 以上)，音源ボード (SoundBlaster16)，音声モデム，および電話アダプタ (2線／4線変換装置) から構成される．音声モデムは，電話回線制御とＰＢ受信に使用する．認識入力音声と合成出力音声は，電話アダプタを通して電話回線からダイレクトに入出力する．

音声認識は，音素同期型ＨＭＭ－ＬＲによる不特定話者電話音声認識ソフトウェア [4] を使用しており，認識対象語彙の設定は単語リストまたは文法 (ＢＮＦまたはＣＦＧ) で簡単に行なえる．音声合成は，トライホン環境を考慮した波形合成型のテキスト音声合成ソフトウェアＦＬＵＥＴ [5] を用いており，漢字かな混じりの通常の日本語テキストが扱える．

2.2 ソフトウェア構成

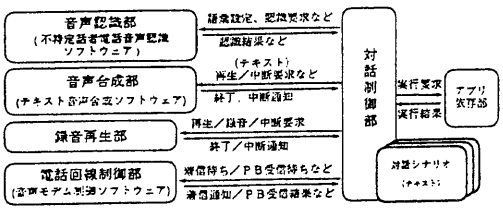

図 2: ソフトウェア構成

システムのソフトウェア構成を図2に示す．システムは，(1) 音声認識部，(2) 音声合成部，(3) 録音再生部，(4) 電話回線制御部，(5) 対話制御部，および (6) アプリ依存部，の6つの機能部分から構成される．

音声認識部は，単語リストまたはＢＮＦやＣＦＧで記述された単語文法で与えられた語彙

を対象として，不特定ユーザの電話音声を認識する．音声合成部は，仮名漢字混じりのテキストからシステムプロンプトを合成する．録音再生部は，録音された音声ファイルを再生したり，ユーザの声を録音する．電話回線制御部は，着信やＤＴＭＦ信号などの検出や，システム側電話機のオンフック，オフフックを行なう．対話制御部は，５章で説明するサービスの処理手順を記述した対話シサリオに従って，上述の４つの部と通信し，システムがユーザと音声で対話できるよう制御する．通信は，サーバ・クライアント方式を採用した．アプリ依存部は，例えばデータベース検索処理などの，アプリケーションに依存した処理を行なう．

3　音声認識部

音声認識部の原形は，電話番号案内タスクへの音声認識の実験的な適用を目的に７万人規模の電話ユーザの住所，氏名の認識に使用したもの [6] であるが，更に認識性能の改善，認識応答時間の短縮，および簡単で迅速なアプリケーションの開発を可能にする，次の５つの新しい特徴を有する．

(1) ４階層共有構造ＨＭＭ，
(2) 音素環境依存ＨＭＭ対応ＬＲパーザ，
(3) 前向きヒューリスティック関数を用いたビーム探索，
(4)client-server model に基づくＡＰＩ．
(5) 全国１万人から収集した電話音声データによるモデルの学習，

(1) ～ (4) に付いては既に他の文献でも紹介されているが，以下で簡単に説明する．

3.1　４階層共有構造ＨＭＭ

認識性能を向上させるためには，音響モデルの精緻化が有効である．そこで従来の単純な音素モデルではなく，前後の音素環境を考慮した三つ組音素 (triphone) モデルを作成した．この際，単純にモデルのバリエーションを増やすと，限られた学習データからよい音素モデルを作成することが難しく，かつ処理時間が増大する．このため (1) モデル，(2) 状態，(3) 基底分布，(4) 特徴量の４階層を共有化することにより，学習効率を高く，計算量を少なくした．新しく提案したのはパラメータレベルの共有化であり，(1) ～ (3) のレベルの共有化は既存の方法を利用した [7]．

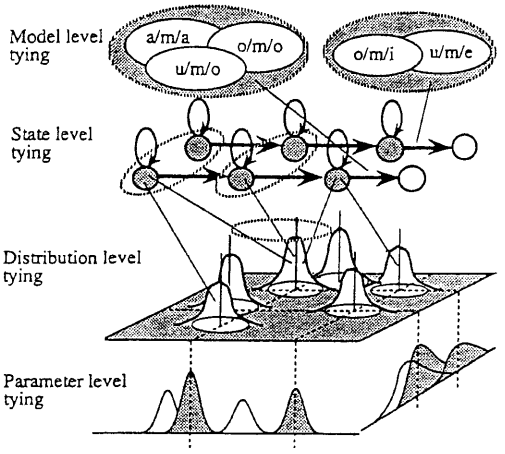

図 3: ４階層共有モデルの概念図

４階層共有モデルでは．特徴パラメータレベルの共有化により，モデル全体で 1600 個ある分布の平均値ベクトルを，各次元 16 点の代表平均値の組み合わせで表現できる．またパラメータ表現の階層化が十分なされることにより，モデルが効率的に表現され，認識性能をほとんど低下せずに，認識時の平均値に関わる計算量が大幅に削減できる．

3.2　音素環境依存 HMM 対応ＬＲパーザ

前後の音素環境を考慮していない環境独立の音素を扱っていた仮説生成用のＬＲパーザは，そのままでは環境依存のＨＭＭ (トライホンモデル) を扱うことができない．ＬＲパーザの使用するＬＲテーブルには，従来と同じ環境独立の音素を終端記号とするＬＲテーブルを利用し，最後に予測された音素ではなく，それより前に予測された音素を中心音素として設定する方法を開発し，音素環境依存の HMM を扱うことを可能にした [8]．

本方法は，ＬＲテーブル自体を音素環境依存モデルに対応するように変換する方法に比べると，テーブル作成に時間を要しないので，ダイナミックな語彙の変更に向いており，また bi-

phone(2つ組音素)の場合でも，またtriphoneより長い環境を考慮に入れたモデルでも，照合のタイミングをそれに応じて変化させるだけなので，容易に扱うことが出来る．

3.3 前向きヒューリスティック関数を用いたビーム探索

認識処理において，全ての仮説のパターン照合を行なうのは効率的ではない．一般的には，探索空間を絞り込む何らかの手法がとられる．本システムの原型[6]では，音声の入力が終了してからでないと処理が始められないという原理的制約があったが，本システムでは前向きヒューリスティック関数を用いたビーム探索により，音声の入力と並列して高精度で効率的な探索が行なえる[9]．

3.4 ＡＰＩ

表１．音声認識の基本仕様

項目	仕様
話者性	不特定話者／話者適応
音響モデル	電話用（4 kHz、話者１万人） マイク用（5.5kHz or 6kHz）
語彙（孤立単語）	仮名／ローマ字による登録
文法（連続音声）	CFG／BNF
入力音声	12 or 8 kHz sampling 16 bit or μ-law 8 bit
OS	Unix／Windows-NT
インタフェース	クライアント・サーバ方式、 ライブラリ
その他の特徴 （オフライン処理）	雑音適応NOVO、 話者適応（MAP/VFS）

表１に認識ソフトウェアの仕様を示す．音声認識の専門家でなくても簡単に扱えるように，音声認識の専門的な知識を必要とするアルゴリズム部分と，具体的なアプリケーション部分とを分離した．具体的には，音声認識処理自体の部分をサーバとして実現し，アプリケーションはクライアントとして音声認識APIを介してサーバを利用する方式を採用した[4]．

4 テキスト音声合成部

図４に，テキスト音声合成処理フローを示す．処理はテキスト解析と音声合成の２つからなる．以下では，不自然な読みを減少させ，高品質なテキスト合成音を得るために採用した，新たなテキスト解析の仕組みと音声合成法，更には本ソフトウェアの仕様と構成上の特徴について簡単に説明する．

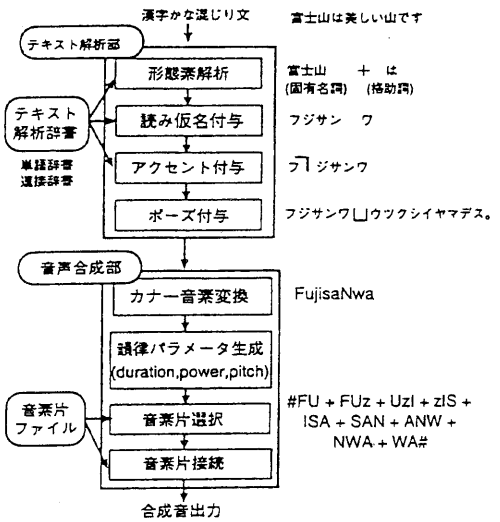

図 4: テキスト音声合成処理フロー

4.1 テキスト解析処理

形態素解析部は，入力テキストを解析し，単語の境界や品詞情報を抽出した後，漢字に対する読みの付与，アクセントおよびポーズの設定を行なう．従来は，漢字を中心とした単語の最長一致に基づいた簡易な形態素解析法を採用していたため，特に複合語やひらがなの部分での解析誤りによる，不自然なアクセントの付与や，読み誤りが比較的多かった[10]．

これらの誤りを減じるため，新たに仮名自立語を加えた約１０万語の単語辞書を用意し，更に単語間の品詞情報をの接続関係を連接コストで表現した連接辞書を用い，コスト最大となる単語列を抽出する探索法により解析精度を向上させた．アクセント付与処理では，テキスト解析の精度向上と新たに付属語アクセントを追加することで，平仮名部分を中心に，アクセン

ト付与精度を大幅に向上することが出来た[11].

4.2 音声合成処理

音声合成部では，テキスト解析部から送られてくるアクセント付仮名文を音素記号列に変換し，音素片ファイルから隣接する音素環境が一致する音素片データを選択，結合した後，規則で設定されたピッチ，時間長データに基づいて，波形データを加工する．音素片データは，日本語に存在する15,000個の3音素連鎖(トライホン)の中から，音素素性を考慮して選択された6,000個のトライホンを含む，無意味単語発声データベースから作成した．

均一に発声されたトライホン発声データベースから音素片を作成することにより，合成音声の安定性を大幅に向上させることが出来た[12]. 音素片の数は，ＣＯＣ法を用いることにより，品質をほとんど劣化させることなく，約3,000個程度までに減らすことが出来た[13].

4.3 仕様および構成上の特徴

仕様を表2に示す．ＦＬＵＥＴへの入力テキストは，仮名漢字混じり文，またはアクセントおよびポーズ情報などが付与されたカナ文である．ＥＵＣコードまたはシフトＪＩＳコードが許容される．出力音声は，１６ビットとμ law ８ビットのＰＣＭ音声である．入力テキストにコマンドを挿入することで，男声と女声の選択，声の速さやボリュームの変更が可能である．この変更は，合成処理中にも可能である．

ＦＬＵＥＴはソフトウェア構成上，テキスト前処理，テキスト解析，音声合成処理の各モジュールで構成される．各モジュールは，機能拡張や今後予想される様々なプラットホームへの移植に対して，最低限の変更量で対処できるようにするため，プログラム設計上，以下の工夫が行なわれている．

(1) ＵｎｉｘやＷｉｎｄｏｗｓなどのＯＳに極力依存しないプログラミング，

(2) サンプリング周波数，システム規模，要求性能に柔軟に対処できるソフトウェア構成，

(3) 音素片数の削減に柔軟に対処できるファイル構成．

フィルタ演算を必要としないため，ｉ４８６ＤＸ２以上のプロセッサであれば，ソフトウェアリアルタイム合成が可能である．

表2．テキスト音声合成の基本的な仕様

項目	仕様
入力テキスト	漢字仮名混じり文(JIS/EUC) アクセント情報付カナ文(JIS/EUC)
辞書	単語辞書：一般語、固有名詞約１０万語 音素片ファイル：約6,０００個
音声出力の制御	入力テキスト中での指定、プログラム側での機能設定が可能 発声速度 (8 段階) 音量 (16 段階) ピッチダイナミックレンジ (5 段階) 声質 (男女各１名)
出力音声	12 or 8 kHz for WS with UNIX 11.025 kHz for PC-windows 16 bit or μ-law 8 bit
OS	Windows／Unix

5 音声対話機能

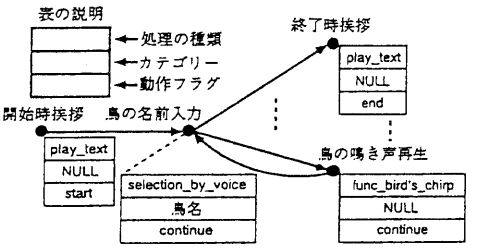

図５：高次レベルシナリオの状態遷移の例

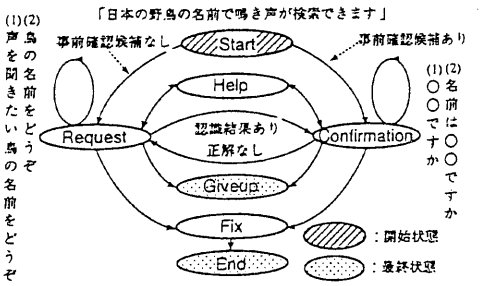

図６：低次レベルシナリオの状態遷移の例（音声によるメニュー選択の場合）

対話は，図５と図６に示す２つのレベルの状態遷移で表現されるシナリオにより駆動される．図５は高次レベルの状態遷移の例であり，各状態はそこで実行される処理を表す．各状態は実行する処理の名前と種類，入力される項目のカテゴリ，および動作フラグなどを情報とし

て持つ．主な処理の種類としては，音声 and/or ＰＢ信号によるメニュー選択，複数カテゴリに対する認識候補の音声による一括正誤確認，ＰＣＭファイルまたは合成音によるプロンプト再生，および音声録音などがある．通常の対話は，動作フラグ start を持つ状態から始まり，動作フラグ end を持つ状態で終了する．

図６は音声によるメニュー選択処理の状態遷移の例であり，図５の一つの状態に対応する低次レベルの処理を表す．認識を実行する状態は，Request と Confirmation である．

各状態での動きを例で説明する．例えば鳥の名前を言って鳥の声を聞くようなサービスの場合，Start 状態で「日本の野鳥の名前で鳴き声が検索できます」と出力し，通常は Request 状態に移行して「声を聞きたい鳥の名前をどうぞ」(内容は Request に来た回数により異なる) と入力を要求し，話者の応答音声を認識する．認識結果があれば，Confirmation 状態に移行し，「(認識候補) ですか」と確認し (最大N位候補まで確認)，応答音声を認識する．確認音声の認識結果が肯定であれば，Fix 状態に移行し，その候補に同定した場合に特別に出力する内容があれば再生し，End 状態に移行して，その単語に応じた次の処理 (図５のレベルに戻る) に移行する．認識時に音声入力なしや結果がなしが続く場合や鳥の名前が同定できない場合は，その繰り返し回数に応じて Help 状態に移行して操作説明を行なったり，Giveup 状態に移行して他の入力手段，例えばDTMF 信号，による代行入力処理に変更するか，認識を諦めて処理を中断する．

認識候補の確認処理の要否は，フラグにより指定できる．確認処理を行なう場合でも，認識結果の確度に応じて省略できる処理を導入している．すなわち，確度が予め決めたしきい値より高い場合は１位候補を正解と見なして確認を省略し，確度が予め決めたしきい値より低い場合は確認せずに再入力を要求する．再入力音声を認識する場合は，初回より大き目のビーム幅を設定する．正誤を確認した候補がすべて否定されて再入力を要求する場合には，認識対象から否定された候補を外し，同じ誤りが連続することを防いでいる．認識確度は，音響ガーベージモデルからの認識尤度で補正した１位および２位候補の認識尤度から決定する．

扱う対話の種類は，システム主導の１問１答穴埋め型に限定した．これは操作方法に関する事前知識のない不特定のユーザに使用されること，電話回線経由では音声認識精度が低下することを考慮したからである．

システム開発者は，図５と６に対応する情報を表現したシナリオをテキストで記述するだけで簡単に対話システムを構築できる．

6 まとめ

本稿では，特別なハードウェアを必要とせず，音声認識，合成に関する特別な知識がなくても簡単に且つ迅速に電話応用システムの作成が可能な，対話システムについて述べた．今後は，本システムが扱える対話の拡張を図る．また対話記述言語についても検討を続ける．

謝辞 日頃御指導を頂き，発表の機会を与えて下さった北脇ＮＴＴヒューマンインタフェース研究所音声情報研究部長に感謝致します．

参考文献

[1] S. Yamamoto et. al. : Proc. JASJ Conf., 1-5-14, pp. 33–36, March 1996
[2] T. Nitta, Proc. ICSLP94, vol. 2, pp. 671-674, Sept. 1994.
[3] 磯ほか：Proc. IEICE Conf., D-654, pp.213–213, March. 1996.
[4] 山田ほか：Proc. JASJ Conf., 2-8-2, pp.41–42, Oct. (1994).
[5] K.Hakoda et. al.: Proc.AVIOS95, pp.65–72, (1995).
[6] Y. Minami et. al. : Proc.IVTTA92, Oct. 1992.
[7] S. Takahashi et. al. : Proc. ICASSP95 (Detroit), pp. 520–523, 1995.
[8] T. Yamada et. al. : Proc. JASJ Conf., 3-8-8, pp. 123–124, Oct. 1994.
[9] Y. Noda et. al. : Proc. Eurospeech95 (Madrid), WEam1A.5, pp. 913–916, Sept. 1995.
[10] 箱田ほか : Proc. IEICE Conf., A-128 Oct. 1990.
[11] K. Hakoda et. al. : Proc. AVIOS95, pp. 65–72, 1995.
[12] 吉田ほか : Proc. JASJ Conf., 2-1-3, Sept. 1995.
[13] Y. Yoshida, et.al. : ICSLP96, Oct. 1996.