

論文 / 著書情報
Article / Book Information

論題	音声認識・合成を用いた計算機利用案内システム
著者	畝 徳治, 笹島 義弘, 管村 昇, 古井 貞熙
掲載誌/書名	研究実用化報告, Vol. 32, No. 11, pp. 2317-2326
発行日	1983, 7

681.3.07.004.14(079.5):534.78

音声認識・合成を用いた計算機利用案内システム*

畠	徳	治**
笹	島	弘***
管	村	昇**
古	井	熙****

あらまし 研究所におけるオフィスオートメーションサービスの一環として、音声情報案内サービスの研究を進めており、新しい音声認識・合成技術(SPLIT 法および LSP 法)を適用した計算機利用案内システムを実現した。本システムの実現では、単語辞書のオーバーレイ方式、単語辞書のカテゴリ化、音声のオンライン登録機能、大量音声メッセージの合成方式、利用者との会話法など、案内サービスに有効な方式を実現した。また、サービスに必要な音声認識装置および音声合成装置は、それぞれ、SPLIT 法および LSP 法にもとづいて試作した。

本論文では、システム構成、音声認識方式、音声合成方式、利用者との会話方式等の検討経過および試行サービス結果の考察を述べる。

1 ま え が き

音声情報処理は、膨大な量の演算処理を要するものが多く、従来は、この点が実用化の大きな障害の1つになっていた。しかし、最近の LSI 技術の進歩により、演算量の問題が次第に解消されつつある。その結果、音声情報処理技術も、次第に実用化の段階に入ってきた。たとえば、電電公社では既に、銀行業務を対象に音声認識・合成技術を適用した ANSER[†]を実用化

している。

一方、研究所では、音声認識については SPLIT 法^{††}を、音声合成については LSP 法^{†††}を開発し、それぞれ研究を進めてきた。新しい音声認識・合成技術を用いたサービスの開発を目的とした研究の一環として、研究所における計算機利用案内サービスにこれらの技術を適用し、案内システムを実現した。

音声認識応答システムを電話機利用による案内サービスに適用するには、多くの課題を解決しなければならないが、本システムの実現に当たり、次のような課題を重点に検討した。

- (1) 電話回線を経由する不特定話者単語音声の高性能な認識の実現。
- (2) 大量音声メッセージの高品質な合成の実現。
- (3) 利用者との柔軟な会話法および案内法の実現。

上記の課題(1)を達成するため、研究所で開発した

* Computer Information Service Using Speech Recognition and Synthesis Techniques. By Tokuji UNE, Yoshihiro SASAJIMA, Noboru SUGAMURA and Sadaoki FURUI.

この研究は情報通信サービス研究室および第四研究室で行われたものである。

** 武蔵野電気通信研究所研究専門調査員

*** 武蔵野電気通信研究所研究専門調査役

**** 研究開発本部調査役

† ANSER: Automatic answer Network System for Electrical Request

†† SPLIT: Strings of Phoneme-Like Templates

††† LSP: Line Spectrum Pair

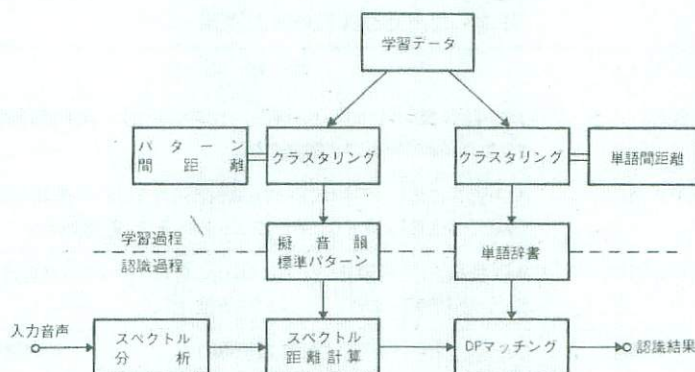


図 2 SPLIT 法による単語音声認識

れ、研究所で開発した SPLIT 法および LSP 法を適用して試作した。

(3) 音声認識方式は、単語辞書の構成法により、特定話者方式と不特定話者方式とに分けられるが、本システムは、利用者の追加に対応できる、不特定話者方式とした。

(4) 案内情報の収集と利用者に対する案内とが、ともに高い即応性を要求されるため、2 台の汎用ミニコンにより機能分担する構成とした。すなわち、案内情報収集装置は、各計算機センタからジョブ処理状況等の情報を常時収集し、案内情報処理装置に転送する。案内情報処理装置は、音声認識・合成装置を介して利用者と音声で会話し、各種の情報案内を行う。

4 不特定話者単語音声認識

不特定話者が発声した単語音声を認識する方法として、種々の方法が提案されている^{(4)~(6)}。本システムでは、多数氏名音声を認識するため、SPLIT 法にもとづく単語マルチテンプレート方式を用いて、音声認識装置を試作した。

4.1 SPLIT 法による単語音声認識

4.1.1 SPLIT 法の概要

SPLIT 法による音声認識方法を図 2 に示す。SPLIT 法では 16 ms ごとに切り出した音声フレーム単位にスペクトル分析し、認識の基本パラメータを抽出している。

(1) 学習過程

学習データを用いて、下記の手順であらかじめ、擬音韻標準パターンと単語辞書を作成する。

(i) 擬音韻標準パターン：学習データに出現する多数の音声フレームのスペクトルパターンを、パターン相互間のスペクトル距離にもとづくクラスタリング手法によりグループ分けし、グループごとの平均パターンを擬音韻標準パターンとする。擬音韻標準パターンは、256 個選択する。

(ii) 単語辞書：多数の学習データを単語別に、単語間のマッチング距離にもとづくクラスタリング手法によりグループ分けし、グループの中心に位置する単語を選択する。選択した単語のスペクトルパターン系列を、フレームごとに最も類似する擬音韻標準パターンのシンボルで表現して単語辞書とする。各シンボル系列を単語テンプレートと称する。

(2) 認識過程

入力音声は、下記の手順で認識される。

(i) 入力音声をスペクトル分析し、フレームごとに擬音韻標準パターンとのスペクトル距離を計算する。

(ii) この距離の値と単語辞書を用いて DP* による時間軸伸縮マッチングを、すべての単語テンプレートについて実行する。

(iii) そしてマッチング距離の最も小さい単語名を認識結果とする。

4.1.2 特徴

SPLIT 法は、下記のような特徴をもつ。

* DP : Dynamic Programming (動的計画法)

表 2 音声認識装置の機能概要

	モジュール	機 能 概 要
1	制御部	ホスト計算機からのコマンドの解析, 音声分析部・音声認識部の制御, オフライン時の各種モニタ機能をもつ.
2	音声入力部	入力音声に対し, アナログ的に高域強調と 4 kHz の帯域制限を行ったのち, 8 kHz サンプリングで 12 ビットの AD 変換を行う.
3	音声分析部	AD 変換された音声データから, 16 ms 幅のフレームごとに音声パワー, 自己相関係数, LPC ケプストラムを抽出する.
4	音声認識部	これらのパラメータを用いて擬音韻標準パターンとのスペクトル距離を計算し, このスペクトル距離を用いて, 単語テンプレートとの DP によるマッチングを行う.
5	インタフェース部	GPIO を用いて, ホスト計算機との間で, プログラム, 擬音韻標準パターン, 単語辞書, 認識結果等の双方向の転送を行う.

(1) スペクトル距離の計算量が単語テンプレートの数に依存しないため, 多数の単語テンプレートを用いる方式に有利であり, 多数氏名の認識に適合する.

(2) 単語辞書を擬音韻標準パターンのシンボル系列で表現するため, 単語辞書に要するメモリが小さく, 単語辞書の入れ替え時間も少なくて済む. 単語辞書として, フレームごとにパラメータそのものを用いる場合に比較し, 約 1/30 である.

(3) 擬音韻標準パターンおよび単語辞書の作成時のクラスタリングは, プログラムにより自動的に実行できる. そのため, 学習データさえ収集すれば, 作成者の主観によらない単語辞書ができる.

4.2 音声認識装置⁽⁷⁾

装置を実現する上で, 下記の点を考慮した.

(1) 多数の利用者の氏名等を認識するため, 発声語長最大 3.2 秒までの単語について, 最大 1000 単語テンプレートまで同時に認識できる能力とする.

(2) 認識率を高めるため, 会話の場面に応じ単語辞書の範囲を限定して認識結果を出力できるようにする.

(3) 利用者の発声回数を抑えるため, 認識結果は, マッチング距離の小さい順に 3 個の単語名と距離の値をホスト計算機に転送する.

(4) 発声終了後, 即時に認識結果を出力させるため, DP マッチングを複数のプロセッサにより並列的に

うとともに, パラメータ抽出, スペクトル距離計算, DP マッチング処理をパイプライン的に実行する.

(5) 認識率とリジェクト能力を高めるため, スペクトル距離尺度として WLR⁽⁸⁾ を, DP として両端点フリー DP⁽⁹⁾ を用いる.

試作した音声認識装置の構成を図 3 に, 機能概要を表 2 に示す.

4.3 単語辞書作成のための音声収集

音声認識性能を高めるうえで, 多数の人の音声を分析し, その特徴を単語辞書に反映させることが重要である. 本サービスのため, 計算機利用者を主な対象に約 300 名の音声を収集した.

これらの音声データは, 伝送経路がサービス時と類似な条件となるよう, 電話回線を経由したものを収集した. 発声者の周囲で生ずるドアの開閉音や他人の話し声等の雑音が, 収集した音声に混入するのを防ぐため, 収集する時に全音声をモニタした.

使用した学習データ収集システムの構成を図 4 に示す.

4.4 単語辞書の構成

4.4.1 認識用語

本サービスで認識の対象とする単語は表 3 のとおりであ

* WLR : Weighted Likelihood Ratio

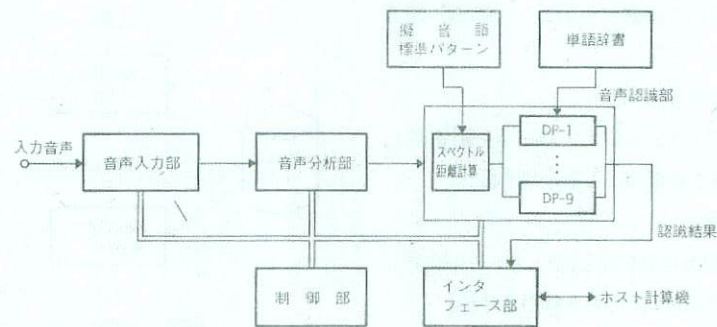


図3 音声認識装置の構成

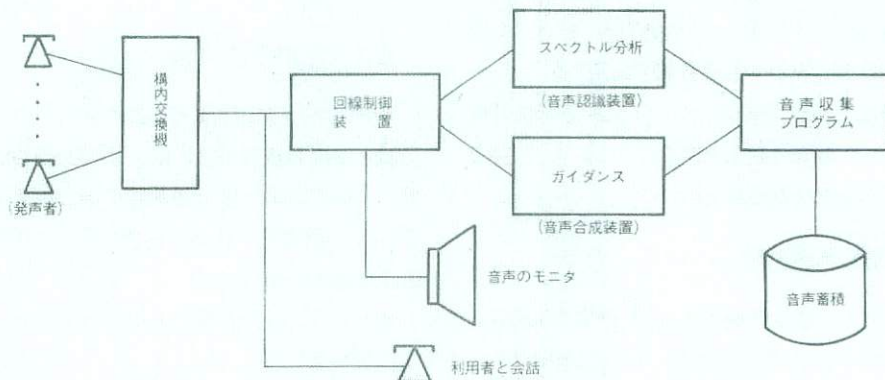


図4 学習データ収集システムの構成

表3 認識用語の種類

種類	認識語
確認語	はい、いいえ
案内名	処理状況、終了時間、運用計画、音声登録
ジョブ種別	センタ・パッチ、Fパッチ
システム名	1号、3号、横須賀
週の別	今週、来週
曜日	月曜、火曜、水曜、木曜、金曜、土曜
所属	(研究室の略称) 約160語
氏名	(計算機利用者の氏名) 約400語

る。

氏名は計算機の利用申請をした者約400名を対象としている。

4.4.2 氏名認識用単語辞書

氏名以外の単語辞書は、サービス中に更新する必要

はなく、これらは項4.1.1で述べた手順により作成した。これに対し、氏名用単語辞書は、その構成上以下のような特徴をもっている。

(1) 本サービスでは、氏数の数は多いが、これは本人が発声したときに正しく認識できれば十分である。この点を利用して氏名の認識性能を向上させるため、氏名認識用単語辞書は、それぞれ1つずつ本人の音声を登録する構成とした。

(2) 一方、計算機利用者の追加は年間を通してあるので、初期の段階で全ての氏名を準備するのは不可能である。また、本人の音声に変化してゆくことも予想される。そのため、氏名は運用開始後も音声登録案内により適宜に追加・変更できる構成とした。

4.4.3 一般語辞書のオーバレイ化

試作した音声認識装置は、1000単語テンプレートまでの実時間認識が可能であるが、さらに、今後のサービスの増加に対応して実効テンプレート数を簡単に増加できるようにするため、SPLIT法の単語辞書がコンパクトで

あることを利用し、下記の機能を実現した。概要を図5に示す。

(1) 単語辞書の分割使用

1000 テンプレートのうち、700 テンプレートを固定部、残り 300 テンプレートを変動部として分割する。

(2) 固定部の用法

固定部には使用頻度の高い確認語、サービス種別、容量の大きい研究室名および氏名を常駐させる。

(3) 変動部の用法

変動部には、サービス種別が決定した時点で、そのサービスに必要な部分を選択して入れ替える。すなわち、変動部をサービス別の複数の単語辞書で共用する。

上記の機能により、実効テンプレート数が音声認識装置のハードウェア容量に制限されないようになった。これを単語辞書のオーバーレイ化と称している。

4.5 認識予備実験

本サービスでは、電話回線を経由した音声の認識が必要であり、マイクロホン等から直接入力される場合より、認識性能が低下することが予測された。そのため、サービスで用いる単語（研究室名および氏名を除く単語）等 31 種を認識対象語として、事前に認識実験を行った。

音声データは、サービスの場合と類似の条件になるよう、電話回線を経由して収集した音声を用いた。31 種の単語について、1 単語当たり平均単語 テンプレート数 18 の条件で、98.0% の認識率が確認でき、所内サービスへ十分適用できる見通しが得られた。

5 大量音声メッセージの合成

本システムでは、音声メッセージの量が多いため、記憶すべき音声パラメータの量が少なく、合成音声の品質が高い LSP 方式を適用して、音声合成装置を試作した。LSP 方式は、現在広く普及している PARCOR 方式*と比較し、約 60% の情報量で同程度の高品質な合成音声を得られる。本システムではフレーム周期 10 ms、1 秒当たりの音声情報量 4800 bit の音声パラメ

* PARCOR : PARTIAL auto CORrelation

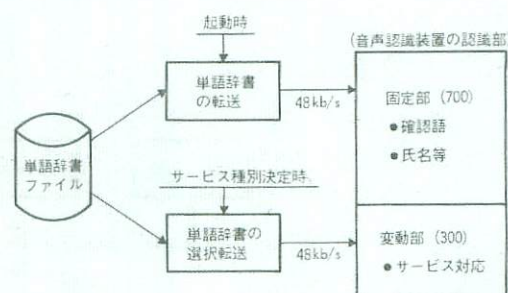


図5 単語辞書のオーバーレイ化

ータを用いた。

5.1 音声合成方式

音声パラメータの作成方法に着目したとき、分析合成方式と法則合成方式とがある。前者は音質が優れているが、あらかじめアナウンサ等が発声したメッセージしか合成できない。反対に、後者は任意のメッセージが合成できるが、音質は前者より劣る。

本システムでは、音質が高いことを重視し、分析合成方式を採用した。

5.2 音声合成装置

(1) 音声パラメータの記憶方式

音声合成装置を実現する場合、①図6のように、音声パラメータをホスト計算機に置く方式と、②音声合成装置に内蔵する方式とが考えられる。前者は一般に、音声合成時のホスト計算機の負荷が大きい。メッセージ量が音声合成装置のメモリ量に制限されない。そのため、大量のメッセージが必要な場合に有利である。また、実験段階で頻繁に必要となるメッセージの修正や変更が容易にできる。これに対し、後者は反対の特徴をもつ。

本システムでは、大量のメッセージが必要なため、ホスト計算機に置く方式①を採用した。

(2) フレーム同期機能

音声の1フレームは、6 バイトのパラメータから成る。パラメータのみを転送した場合、伝送路上でパラメータの1部が消滅すると、それ以降フレーム境界が混乱し、正しい音声合成ができなくなる。フレーム同期用の制御データを音声パラメータに付加することにより、自動的にこの障害を回復する機能をもたせることにした。

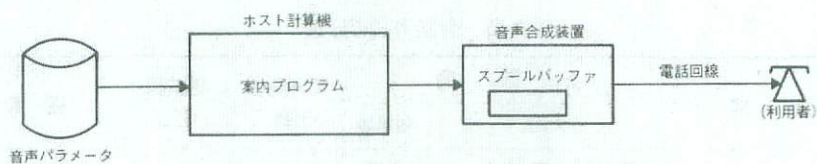


図 6 パラメータ供給形音声合成装置

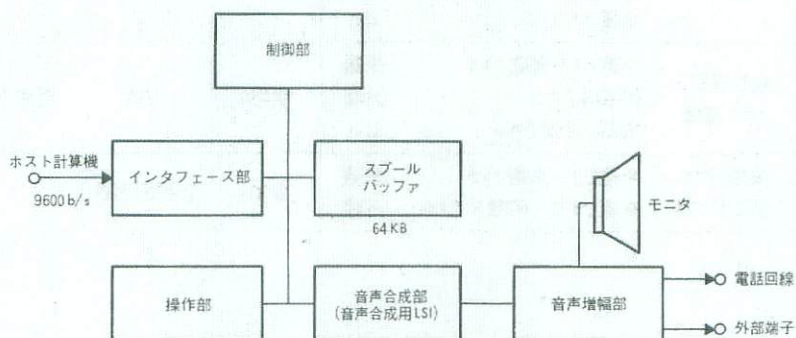


図 7 音声合成装置の構成

表 4 音声合成装置の機能概要

	モジュール	機 能 概 要
1	制御部	ホスト計算機から、制御データおよび音声パラメータを受信し、全体を制御する。制御データは直ちに解釈し、フレーム同期や動作モードの設定を実行する。音声パラメータは、一旦スプールバッファに蓄えた後、フレームごとに音声合成部へ送出する。
2	音声合成部	制御部から受信した LSP パラメータにもとづき、音声を合成する。
3	音声増幅部	音声合成部からの出力を増幅し、電話回線へ送出するとともに、外部端子または内蔵モニターへ出力する。
4	インタフェース部	ホスト計算機との間は、汎用インタフェース RS 232 C で接続され、9600 b/s の速度でデータ転送を行う。

試作した音声合成装置の構成を図7に、機能概要を表4に示す。本装置の音声合成部は、研究所で開発した LSP 音声合成用 LSI (ECL 1565)⁽¹⁰⁾ を用いて実現した。

5.3 音声ファイルの作成

出力回線が1回線であるため、音声セグメントに対し長さ制限の緩い、可変長方式が適用できる。セグメント単位は、音質を維持しながら、セグメントの総量が最小になるよう定めるのが望ましい。本システムでは、音声の自

然さを考慮し、セグメントの境界が無音となる範囲で定めた。

本システムでは、セグメントの管理をホスト計算機のオペレーティングシステムにまかせられるよう、セグメントごとに1つのファイルに格納するようにした。

利用者の音声については、サービス開始後も引続き更新してきた。利用者名を含め、これまで、1100 セグメントの音声ファイルを作成した。

表 5 会話方式の分類

方 式		会 話 例		質問回答 の回数	認識誤り の訂正	備 考
		システム	利用者			
1	1 項目ごと に確認	今週ですか来週ですか 来週ですね 何曜日ですか 月曜ですね	来週 はい 月曜 はい	多い	容易	採用した
2	複数項目ま とめて確認	今週ですか来週ですか 何曜日ですか 来週の月曜ですね	来週 月曜 はい	少ない	複雑	可能である
3	確認と次の 質問を一括	今週ですか来週ですか 来週ですね。何曜日ですか	来週 月曜	少ない	比較的容易	

6 利用者との会話法

6.1 会話の単位

単語音声認識を用いた案内システムでは、利用者はシステムの認識動作に同期し1単語ずつ区切って発声する必要がある。その場合、①システムからの1回の質問に複数項目を連続して発声する方式と、②1回の質問に1項目ずつ発声する方式とが考えられる。

本システムでは、利用者が発声し易い点を重視し、②の方式とした。また、発声のタイミングがはずれた場合、3回まで言い直しができるようにした。

6.2 確認と訂正の方式

音声認識の結果は誤っている可能性が零ではない。そのため、認識結果を確認する手続きと、認識誤りを訂正する手続きとが必要である。確認の手続きに注目して会話方式を分類すると表5のようになる。この表に会話方式の特徴も合わせて示した。

本システムでは、認識誤りの訂正が容易な点を重視し、第1の方式とした。また、利用者の発声回数を抑えるため、認識装置から転送された3個の候補について順次確認し、すべて否定された場合に再度発声させる方式とした。

7 考 察

研究所の計算機利用者約400名を対象に試行サービスを行い、その運用データから音声認識・合成技術の適用法、サービスの改善点等について考察する。

7.1 サービスの利用状況

サービス項目別では、ジョブ処理状況案内(45%)、ジョブ終了時間予測案内(25%)、計算機運用計画案内(22%)、音声登録案内(8%)の順に利用されている。このことから、利用者は、刻々変動する状況が容易に把握できる情報を望んでいることがわかる。ジョブ処理状況案内やジョブ終了時間予測案内は、音声によるリアルタイムサービスに整合するサービス対象であり、運用の定常化により効果が出るものと期待できる。

7.2 音声認識技術の適用法

1回の認識で3個の候補が得られるが、約600回の発声に対し、1位の候補が正しい割合は、96~97%であり、ほぼ100%の割合で3位までに正しい結果が入っている。その結果、利用者はほとんどの場合、1項目の入力に1回の発声と3回以内の確認で済んでいる。このことから、ここで適用した音声認識技術、学習データ収集方式は、所内サービスに十分耐えらると思われる。

7.3 音声合成技術の適用法

音声セグメントの境界が無音となるよう限定して音声ファイルを作成したため、自然性の良い案内メッセージを合成することができた。また、音声パラメータをホスト計算機に置く方式としたため、試行サービス結果に基づくメッセージの修正・追加に容易に対処できた。これらのことから、ここで適用した音声合成技術およびハードウェア構成は、所内サービスに十分耐えうると思われる。

7.4 会話方式

丁寧なメッセージをゆっくりした速度で出力するとともに、利用者が1項目発声するごとに認識結果を確認しているため、会話がまどろっこしいという意見がある。したがって、より使い易くするためには、メッセージの簡略化、発声速度の適正化とともに、認識結果の確認を少なくすることが重要である。

そこで認識結果の確認を少なくするための会話方式として、次のような方式を想定し、予備実験を行った。

(1) 3個の候補が一致しているとき、その時点での確認を省略し、ばらついていいるときのみ確認する。ただし、会話の流れが変わる場面では省略しない。

(2) いつ、どこで、なにを、等の単位で複数項目を一括して確認する。

予備実験では、3個の候補がそろって正しい割合は、約80%である。したがって、この方法により認識結果の確認を少なくすることが可能であり、会話時間(回線保留時間)の短縮等に効果があることがわかった。

8 む す び

本システムの実現およびサービス実験により、新しい音声認識・合成技術を適用したサービスの有効性が明らかにできた。また、電話回線を経由した不特定話者の音声を対象とする認識用単語辞書の作成法、大量音声メッセージの高品質な合成法、利用者との会話方式の改善点などが明らかにできた。

本システムは、下記のような特徴をもっている。

(1) 研究者は自分に必要な刻々変動する情報を、

手近な電話機から即座に得ることができる。

(2) サービスの新規利用者は、電話機から自分の氏名音声を音声認識用単語辞書に登録して、利用者になることができる。

今後より使い易いシステムを実現するため、利用者との自然な会話法、等を検討するとともに、システムの維持・管理の効率化を図るため、音声認識に関する学習データの自動収集、法則合成による音声ファイルの作成、等を検討する必要がある。また、システムの適用対象を拡大して行くためには、公衆回線を経由する音声の認識へ適用した場合の問題点等を検討する必要がある。

今後は本システムの作成経験を生かし、各種予約案内サービス、計算機利用に関する通知案内サービス等を検討する予定である。

謝 辞

本検討に当たり御指導頂いた工務部青柳部長、基礎研究部畔柳部長、寺島前情報通信サービス研究室長、第四研究室板倉室長に深謝します。また、音声合成装置のハードウェアおよびソフトウェアの設計に御協力頂いた第四研究室嵯峨山補佐、音声収集に御協力頂いた第四研究室匂坂研究専門調査員およびソフトウェアの設計・デバッグを担当して頂いた情報通信サービス研究室植島研究主任に深謝します。

文 献

- (1) 管村・古井：擬音韻標準パターンによる大語い単語音声認識，信学論，J 65-D，No. 8，pp. 1041～1048，1982.
- (2) 管村・板倉：線スペクトル対(LSP)音声分析合成方式による音声情報圧縮，信学論，J 64-A，No. 8，pp. 599～605，1981.
- (3) 亘理・渡辺・千葉：不特定話者用単語音声認識装置の試作，音響学会講演論文集(春季)，3-1-23，1981.
- (4) L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpen: Speaker-Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Trans. ASSP-27, No. 4, pp. 336～349, 1979.
- (5) 長島・中津・小島・石井：不特定話者電話音声認

- 識法, 信学会部門別全大, p. 1-107, 1981.
- (6) 麻田他: ハイブリッド構造マッチング法による電話音声の認識, 音響学会講演文集(春季), 1-4-16, 1982.
- (7) 古井・管村・笹島: SPLIT 法による大語い単語音声認識装置, 音響学会講演文集(秋季), 2-1-17, 1981.
- (8) 杉山・鹿野: ピークに重みをおいた LPC スペクトルマッチング尺度の評価, 音声研究会資料, S 30-03, 1980.
- (9) 鹿野: 大語い単語音声認識における LPC スペクトルマッチング尺度の評価, 音声研究会資料, S 30-60, 1980.
- (10) 嵯峨山・管村・板倉・小池: LSP 音声合成用 LSI を用いた音声合成, 音響学会講演文集(秋季), 2-1-1, 1980.

(1983, 7, 14 受付)