音声グループ

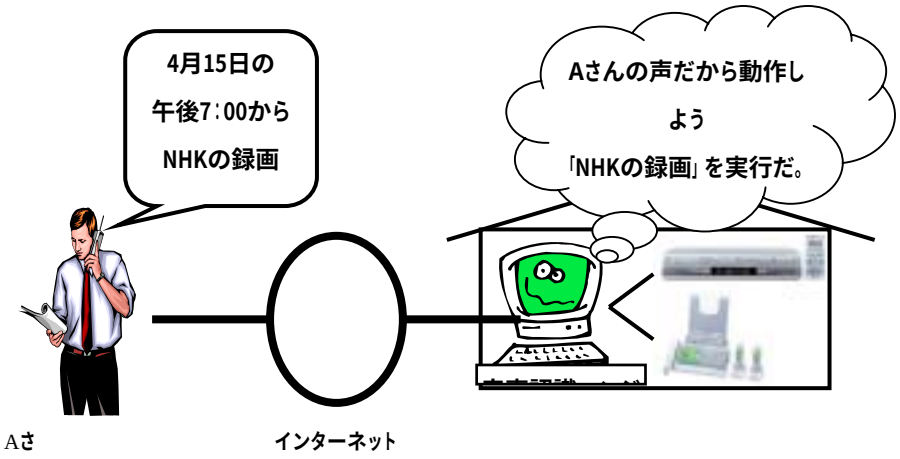
研究概要

音声情報処理は，人間と機械（コンピュータ）のコミュニケーション，あるいは人間同士のコミュニケーションを支援するシステム（例えば音声インタフェース，音声を用いた情報検索，音声翻訳，音声の強調・声質変換，補聴器など），情報システムを構築する上で重要な技術です．

音声グループでは，音声認識・話者認識に関する研究を行っています．音声認識とは「音声信号から何を話しているかを判断する技術」で，話者認識は「音声信号から誰が話しているかを判断する」話者識別と「本人の声かどうかを判断する」話者照合にわけることができます．

私たちの研究室では，雑音があっても，携帯電話でも，正しく音声認識・話者認識ができる方法の研究を進めており，次世代携帯電話用世界標準音声認識エンジンの開発や，特定の人の声だけに反応する自動車やロボット用の音声認識・話者認識システムの開発を進めています．

Keywords：音声認識，音声合成，話者認識，IP 電話(VoIP)，MFT，HMM，GMM，EMD



図：ユビキタス環境化での遠隔操作の例．音声認識エンジンや個人認証サーバでは、登録された人物であるかどうかや何と発話されたかを判定している．

波形合成を用いた音響モデル適応法に関する研究

近年の計算機の性能向上に伴い，大語彙連続音声認識技術が格段に進歩し，音声認識は実用化の段階を迎えている．実際に実環境で収集された音声データで学習した音響モデルを用いることにより数千語彙の電話音声認識が実用化され，音声認識・音声合成による電話応答システムにより道路交通情報や航空券などのチケット予約，株価照会，音声による情報検索ができる音声ポータルなどのサービスが手軽に場所を選ばず利用できるようになってきている．

しかしながら，実環境下では事前に学習したモデルの音響的特性と入力音声の音響的特性間にミスマッチが生じ，音声認識システムの精度が低下する問題が存在する．このミスマッチの原因として，音声通信の際の音声圧縮技術による音声品質の劣化や歪み，背景雑音などが挙げられる．そこで入力音声の歪みに対して頑健で環境に

適した音響モデルの生成法が求められている．認識性能低下の改善策として，最も簡単かつ理想的な音響モデルの学習法は利用環境で音声データを収集し音響モデルを学習することである．しかし，実際には環境毎に音響モデル学習に必要な大量の音声データを入手することは困難な場合が多い．

この問題を解決するために，近年盛んに研究されている音声合成技術を利用し，音響モデルからの波形合成を用いて，環境毎に合成音声を再生・録音する手法を提案した．これは音声認識の分野で広く用いられている，隠れマルコフモデル(Hidden Markov Model, HMM)による音声認識とは逆に，学習された HMM から音声を合成する手法である．提案手法は，HMM から波形合成により合成した疑似音声を当該環境で再生・録音し，その再録音した音声を音響モデル適応に用いる．これにより，提案手法は以下のような利点がある．

- 任意の環境に音響モデルを適応することができ，当該環境における人による発声の手間を削減できる．
- 音声認識に用いる HMM を用い波形合成を行い，適応音声を作成することにより，波形の入出力より直接音響モデルの音響空間の変移(移動)が推測でき，様々な環境に対し柔軟に適応できる．
- 不特定話者で学習したモデルの音響的特徴を用いて適応を行うことから，話者に偏りのない適応が可能である．

評価対象として IP 電話用音声符号化である G.723.1 を用いて，符号化音声の認識実験を行い，提案手法の有効性を確認した．

音声認識・音声合成を用いた音声途切れ補間手法

近年, VoIP(Voice over IP)技術の発展により急速に IP 電話が普及しつつある．IP 電話の本質的な問題として(特に無線区間がある場合)，パケット損失による音声品質の劣化が挙げられる．この問題を解決する手法として，パケット損失直前の波形を繰り返し用いる G.711 Appendix I が利用されている．しかし，この手法は最大 6 フレームまでのパケット損失しか復元できない．さらに，2 音素以上の欠落には原理的に対応していないという問題がある．

そこで，我々はこれらの問題を解決するため，音声認識技術と音声合成技術を用いた音声途切れ補間手法を提案した．提案手法は，(1)Missing Feature Theory に基づく音声認識により，途切れ区間の音素片を前後の言語情報及び音響情報から推定し，(2)推定した音素片列に基づき HMM 音声合成により消失区間の音声波形を生成し補間する．

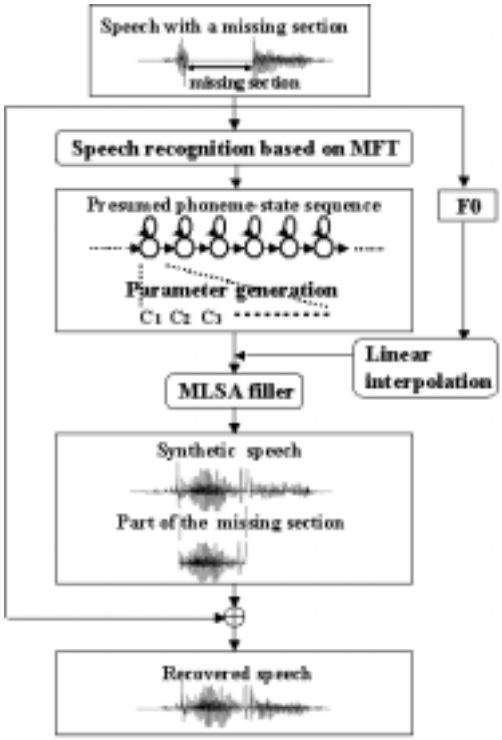


図 1 システムの流れ

時期差に頑健な話者認識に関する研究

音声信号には，一般に音韻性と話者性が含まれていると考えられている．音韻性とは，発話したその内容のことを指し，話者性とは，各個人が独自に持つ個人性情報を指す．話者認識では，音声信号に含まれるその個人性情報を用いて誰の声であるかを識別および照合する技術である．この技術が高い安全性を伴うようになれば，車や建物・部屋の入り口における音声キーなどに応用できると考えられる．また，近年特に発達してきたネットワーク上でのバンキングやショッピングなどにおけるセキュリティのための個人認証として実装できると期待されている．

しかし，今現在の話者認識精度はその実装に見合う程の高い性能には至っていない．その問題点として，以下のようなことなどが挙げられる．

- 音韻性と話者性の分離の困難さ
- コーデック・回線特性の影響や電話帯域の制限
- 学習モデルの作成に用いた音声と入力された音声の発話時期の違い

本研究では，これらの問題点の一つである発話時期の違いに関する研究を行っている．この発話時期の変動は，音声を特徴量で表した空間上でも現れる．その変動に対し，線形的手法である線形判別分析(LDA：Linear Discriminate Analysis)を用いて特徴量の変動を抑えた新しい空間を求め、その上でモデル学習および認識を行っている．実験結果より，(1)音素により同じ時期的影響を受けても異なった変動をする，(2)時期差以外の特性をも押さえ込んでしまっていることなどが考えられる．課題として，音素毎の時期的特徴変動の詳細な調査や変動を含んだ特徴料を学習に用いた実験など行っていく．

音声認識を用いて苗字の漢字表記を認識する手法の研究

近年，ビジネスシーンにおいて音声セルフサービスが広がっている．このサービスは，音声認識を用いて，電話による顧客対応を自動化するというものである．しかし，従来は人間がやっていた作業を機械に任せようとしたら，様々な困難が伴う．何故なら，人間と違って機械は融通が利かないからである．その 1 つとして，日本人の氏名の漢字表記の認識の問題が挙げられる．例えば，「マスコ」という読みの苗字があるとする．そして，この苗字の漢字表記が「増子」だとする．しかし，音声による情報のみでは人間でも「増子」か「益子」か，区別がつかない．それでも，この場合顧客である，「マスコ」さんが「増えるに子供の子で増子です」という説明をしたとする．人間なら，「ああ，増子さんなんだな」と分かるのであるが，従来の機械では，まったく理解できないのである．

そこで，本研究では機械にも，このような説明の文章を理解させようということを目的としている．この研究が成功すれば，音声セルフサービスにおいて，従来は，カタカナなどで認識していた人名を正しい漢字表記で認識できるようになる．

ベクトル量子化と Earth Mover's Distance を用いた分散型話者認識手法に関する研究

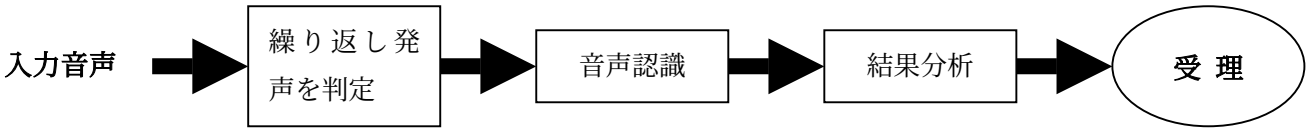
近年、生体情報を用いた個人認証技術であるバイオメトリクス認証の実用化が進んでいる。その中でも、音声を用いた個人認証技術である話者認識技術は、ユーザーの心理的負担が少ないことや、特別なハードウェアの必要性が無い（携帯電話で認証が可能）ことから、今後、需要の増加が期待されている。ただ、携帯電話で話者認識を行う場合、音声を圧縮して伝送するコーデックの影響から認識性能が低下してしまう。そこで携帯端末側で音響分析を行い、サーバー側で認証を行う分散型話者認識方式が提案された。この分散方式を用いることで、コーデックの影響を回避することができる。しかしながら、従来から広く用いられている話者認識手法である GMM を用いた話者認識手法を分散方式に適用した場合、認識精度は大幅に低下してしまう。そこで、我々は分散方式に適した新たな話者認識手法の提案を行った。本手法は GMM のような統計的パラメータの推定を行わないノンパラメトリックな話者認識手法で、話者モデルを特徴パラメータのヒストグラムで構成する。また、認識時におけるテストデータと話者モデルの間の距離尺度には、離散分布間の距離が計算可能な Earth Mover's Distance を用いた。提案手法の有効性を調査するために行った話者識別実験の結果から、従来のパラメトリック手法の一つである GMM に比べ識別誤り率を 31.9%削減することができた。

繰り返し発声を用いた認識精度向上に関する研究

音声認識技術が様々なところで、利用されている。例えば、電話によるライブなどのチケット予約や、コミュニティセンターにおける音声情報案内システム、カーナビの地名入力タスクの音声入力などである。しかし、こういった音声認識を行う際、友達と喋っているのに、入力音声として認識してしまい誤認識を起こしてしまうことがある。こういった誤認識を取り除くための１つの方法として繰り返し発声を用いて、認識精度を上げようというのが、本研究の目的である。なお、今回は、カーオーディオのコマンド入力を想定して実験を行った。

実際にどんな手法なのかを説明すると以下のような流れになる。まず、入力音声に対してそれが繰り返し発声かどうかの判定を行う。普通に会話をしている中で、入力コマンドを繰り返し発声することは、まずないのでそれを検出する。その後、繰り返し発声だと判定されたものに対してのみ、音声認識を行い、それ以外はその時点で棄却する。繰り返し発声だと判定されたものには、音声認識を行って、その結果を分析し、同じ認識結果のみを

入力コマンドとして受理するといった処理を行っている。



・実験とその結果

上記のような流れの中で、繰り返し発声の判定するための実験を行い、繰り返し発声の検出精度は 96.7%といった数字が得られ、またその後の音声認識部において、認識結果から正しく入力コマンドが受理されるのは、100%という数字が得られた。今回はクリーンな音声で、実験を行ったので、今後としては、実際の車内雑音下での音声を用いて実験を行っていきたいと考えている。

中学理科教授学習システムにおける SAPI を用いた音声インタフェースの実装

現在，私たち A1 研究室では，自然言語処理技術を利用した中学理科教授学習システムの研究構築が行われている。しかし，現在開発されている教育システムのインタフェースは，マウス操作とキーボード操作によるものが主である．キーボード操作においては，欧米に比べ日本ではタイプライタ文化が盛んではなかったため，キーボードになじめない学習者が多い．そのため，タイピング速度による学習の遅れや，なじめないキーボード操作による教育システムへの苦手意識が，学習効率の低下につながるといった問題がある．そこでこれらの問題を解決するために，キーボードに代わる入力手段として音声インタフェースを使用する．音声インタフェース構築において，Windows の API 群である SAPI を使用する．SAPI を使用することにより，エンジンに依存しない開発が可能である。以下に音声インタフェースの構成図を示す.

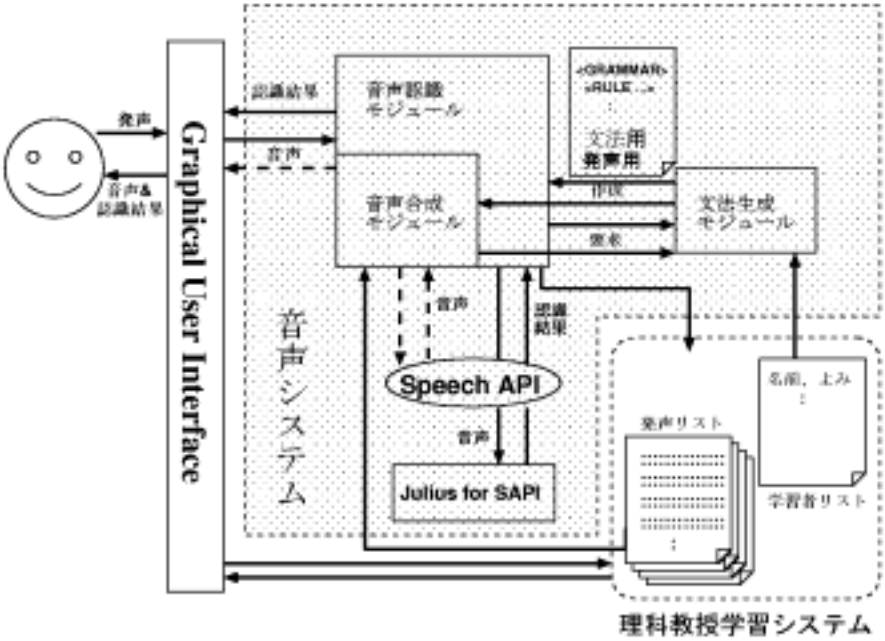


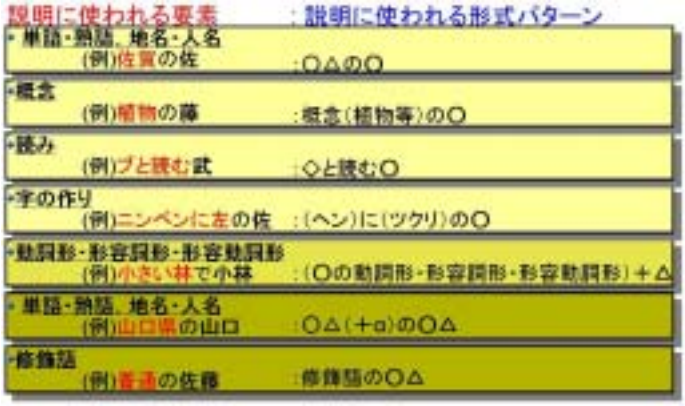
図３．音声システム構成図

音声による姓名漢字入力インタフェース

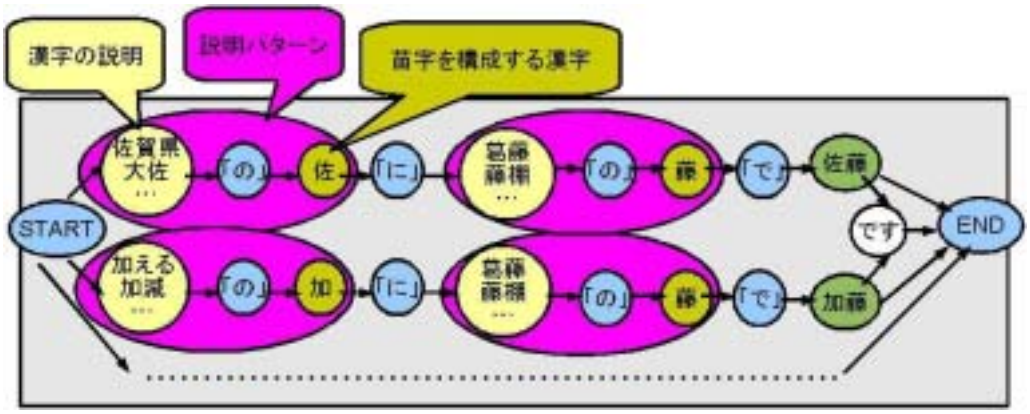
近年，音声認識・音声合成技術の進歩により，通信販売等で，電話対応を自動で行う音声セルフサービスの実用化に期待が高まっている．

しかし，従来の音声認識システムには，日本人の苗字を正しく認識出来ない問題があった．原因の一つは「佐藤」「佐東」等，音声だけでは，どの漢字表記が正しいのかを判別出来ないことにあった．

そこで本研究では，「佐賀県の佐に葛藤の藤で佐藤です」のように，漢字を言葉で説明してもらうことで漢字表記を正しく認識させることを目指した．アンケート結果から，以下の表に示すように，苗字を説明行うにはある程度きまった形式パターンや要素があることがわかった．



現在はこれらの情報を基に，ＦＳＡ(有限状態オートマトン)を用いて作成した言語モデルを構築することで，認識結果道程過程に意味制約を与え，正しく認識させようと試みている．



スペクトル微細構造を考慮した風雑音除去手法にする研究

屋外での音声収録や撮影で風雑音の影響により音質劣化を招くという問題がある．現在，風雑音対策として簡易なスポンジの風防が用いられているが，その雑音抑制能力は十分とは言えず，強風下での音質劣化は避けられない．現在，強風下での雑音調査等ではウインドスクリーン（大型の風雑音防止装置）が用いられている．しかし，ウインドスクリーンを用いた場合，録音機器の大型化、コストの問題により市販の携帯型撮影機器や携帯性の高

いマイクなどに用いることができない.
そこで本研究では市販のビデオカメラやピンマイク向けの雑音除去手法として,
機器の大型化を避けることができ，低コストな雑音除去手法を確立するために信号処理技術に着目した．本研究では風雑音に対して十分な分析を行い，その分析結果を踏まえて，風雑音除去に特化した雑音除去手法を提案した．評価実験の結果，従来手法（信号処理による）に比べ大幅な音質改善を達成した.

音声認識を用いた帯域制限音声の広帯域化に関する研究

電話音声などに挙げられる帯域制限された音声は人間にとって聞き取りにくい場合が多々ある。これは音声が帯域制限されることにより、高周波数成分（高音域）や低周波数成分（低音域）が欠落し、明瞭度や臨場感の低下が起こっているためである。我々は、このような帯域制限された「聞き取りにくい音声」を広帯域の「聞き取りやすい音声」にすることを研究の目的としている。実現方法としては、帯域制限音声の情報をもとに欠落した成分を推定・付与することで広帯域化を行う。これまでにも様々な広帯域化に関する研究が行われてきているが、広帯域化された音声品質の更なる改善が必要とされている。そこで今回、新たに「音声認識を用いた帯域制限音声の広帯域化法」を提案する。具体的には HMM（Hidden Malkov Model）に基づいた音声認識、音声合成、および音声信号処理による技術を用いて実現する。 HMM に基づいた音声認識は帯域制限音声の発話内容を高精度で推定することが可能である。また推定した内容に沿って音声合成の技術により欠落した成分の作成・付与を行なう。従って、欠落した成分をより正確に推定することで高品質な音声を得ることができると考えられる。

繰り返し発声を用いた頑健な音声認識手法に関する研究

現在、音声を用いたアプリケーションは開発されてはいるがあまり利用されていないのが現状である。原因として２点あり、１点目は日常会話の中で動作してほしくないときに音声インターフェースが誤受理してしまうことである。２点目は動作してほしいときに正確に音声認識できないことである。

そこで「スイッチ切って、スイッチ切って」のような繰り返し発声を用いたらどうかを考える。日常生活の中で「スイッチ切って」のようなコマンドを２回以上言うことは考えにくい。これより１点目の問題はまず解決できると考えられる。２点目の誤認識に関する問題だが認識率を下げる大きな原因は雑音である。雑音は大きく定常雑音と突発性雑音の２つに分けられる。定常雑音に関しては過去に様々な除去手法が提案されてきたが突発性雑音に関しては有効な手法は提案されてこなかった。しかし今回の繰り返し発声を用いる事で１発声目と２発声目の同じ時間帯に突発性雑音が入る事は考えにくい。その事を利用して突発性雑音に頑健な音声認識手法を構築し、有効性を確認した。今後は定常雑音に対しても頑健な手法を構築していくと同時に日常会話から繰り返し発声コマンドを検出する手法も構築する。
図１から図３に示しているのは横軸が時間で縦軸が周波数のスペクトルグラムである。図１と図２で確認される突発性雑音のスペクトルグラムが図３では除去されているのが確認できる。

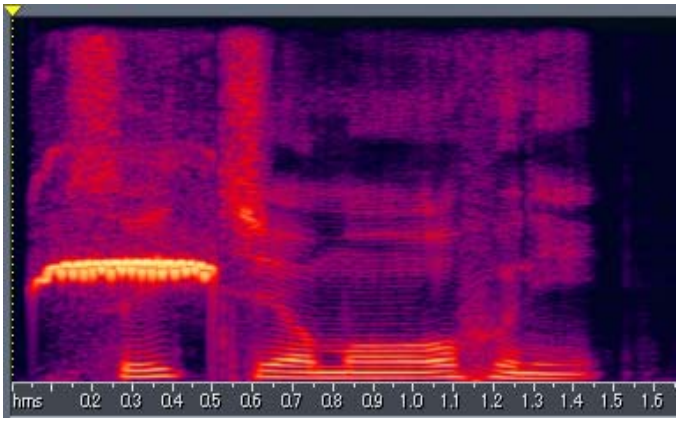


図１：１発生目のスペクトルグラム

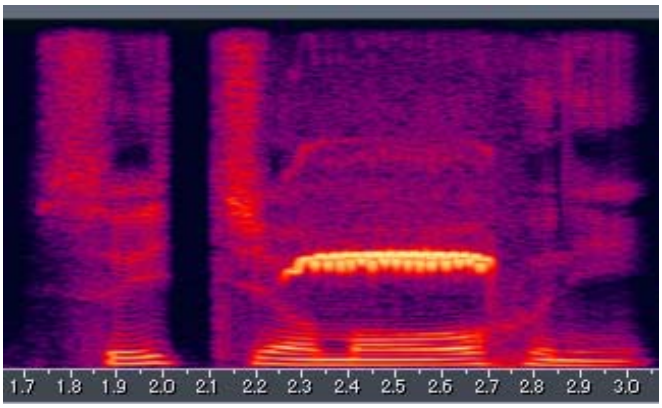
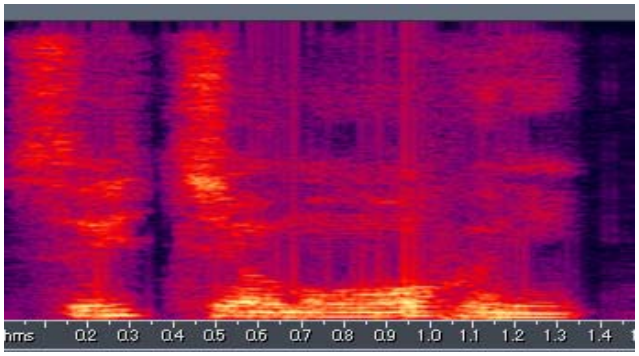


図２：２発生目のスペクトルグラム

図３：提案手法を用いた後のスペクトルグラム



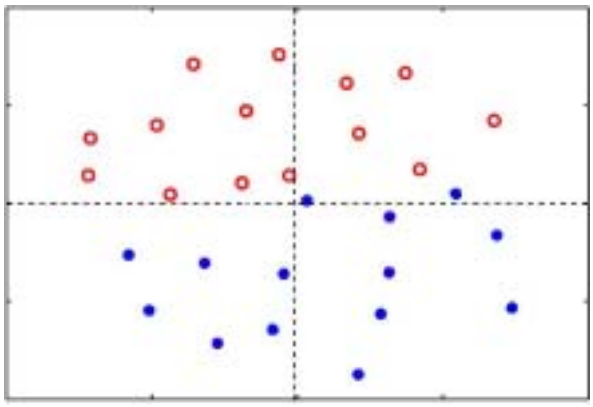
COSMOS に関する研究

近年では、カーナビなど音声インターフェイスを備えた機器が普及して来ている。今後も音響モデルを用いた音声・話者認識機能の需要は増していくだろう。しかし、不特定話者認識技術では、異なる日時や環境等で認識性能にばらつきが生じ、高精度の認識性能を提供しているとは言いがたい。

認識性能を向上させるためには、そういったばらつきの要因を調査する必要がある。そのためには、音響空間上で音響モデルの分布を分析する必要がある。そこで、多次元正規分布である音響モデルを 2 次元平面上に可視化

し、視覚分析能力を利用して音響空間の分析を行なう COSMOS(aCOustic Space Map Of Sound)法が提案された。COSMOS 図を実際に紹介する。これは、男性と女性の音声を COSMOS で描画したものである（○が女性、●が男性）。図によると男性と女性がはっきり別れており、男女間の特徴の違いが COSMOS で確認された。このように、COSMOS 図からモデルの違いを確認することができる。

COSMOS 図



マルチプラットフォームを目指した音声による DB 検索システム

現在，データベースに対して情報を検索したい場合，キーワードとなる文字の入力には，インターフェースとしてキーボードやタッチパネルシステムが利用されている．このようなインターフェースでは，荷物で両手が塞がっている場合や，手に障害を持っている人の場合，手による入力ができない．また，キーボードの入力に不慣れな人や，文字を入力するタイプのタッチパネルシステムを操作する時では，文字の入力に時間を取られてしまう．音声をインターフェースとして利用できれば，誰もが速く検索を行うことができる．

現在の音声認識の精度は，明瞭な発声の場合，概ね正しく書きおこしができる．しかし，音声認識の処理は，かなりの時間を要する．このことから，音声認識は高速な処理が可能なサーバーで行う必要がある．

検索エンジンを利用して情報を得る場合，キーワードとなる単語を入力フォームに入力し，情報を得る．この操作を自然言語に置き換えるなら，「～に関する情報を下さい」や「～について教えて下さい」というのが適当である．このことから，データベースから情報を得る為だけに用いられる自然言語の文法は，ある程度限定されると考えられる．

これらのことにより，データベースから音声によって情報を検索することは可能であると考えられる．

本研究では，データベースからの音声による情報の検索を目的としたシステムを開発することである．上述した通り，音声認識の処理にはかなりの時間を要するので，本システムはサーバー・クライアント型の音声認識システムである．また，クライアントシステムには，マルチプラットフォームでのしようを目指しているので，システムを開発する為の言語に JAVA を使用している．データベースの構築には MySQL を使用した．開発したシステムの動作画面を以下に示す．


IPアドレス 192.168.0.190 ポート番号 7008 切断
フレーム 320 間隔(ミリ秒) 0 辞書ファイル DB選択
マイク マイク+ファイル ファイル(1個) ファイル(複数)
プライマリ サウンド ドライバ, version Unknown Version
認識開始
ジャンレノが出てない映画を見せて下さい
ttle NATURAL JOIN lactor WHERE NOT (lactor.Lactor='ジャンレノ');
セブンイヤーズインチペット
スリーパーズ
リバーランズスルーイット
セブン
デビル
ジョーブラックによろしく
ファイトクラブ