音声対話機能を備えた音色識別学習支援システム

渡辺裕太[†] 関口芳廣[†] 西崎博光[†]

この論文では，数人と対話ができる音声対話機能を備えた音色識別学習支援システムについて述べている．この論文では，特に，音色の例として人間の声を扱っている．まず，最初に音色識別学習支援システムに必要な機能を被験者実験により調査した．その結果，(1) 音声対話機能が必要である，(2) 1 人よりも 2 人で学習すると学習効果が高い，(3) 聴覚情報だけでなく，視覚情報も有効である，という 3 つのことが分かった．そこで，前述の機能を備え，複数人で使用できる音色識別学習支援システムを作成した．被験者実験により，学習支援システムを使って音色識別学習をした場合の方が，システムを使用しないで音色識別学習をした場合より音色識別精度が 50%改善している．

A Tone Color Identification Learning Support System with Speech Dialogue Function

Yuta Watanabe,[†] Yoshihiro Sekiguchi[†] and Hiromitsu Nishizaki[†]

This paper describes a tone color identification learning support system with speech dialogue function that can have a conversation with a few person. In this paper, we especially deal with speech as an example of tone color. First of all, we have investigated functions which are required for the tone color identification learning support system by doing subject experiments. The results showed as follows: (1) a speech dialogue module are needed, (2) training in a twosome gets higher learning effect than training alone, (3) visual information such as spectrograph is useful for training as well as acoustic information. We constructed the support system, therefore, which can be used by a few person and has the functions described above. The experimental result of tone identification by subjects who used the support system showed that the performance of identification achived improvement of 50% accuracy comparing with the case in which the system was not used.

1. まえがき

人間にとって，話し言葉の認識のように，音韻識別を基本にした処理は比較的容易であるが，音韻以外の音情報の識別は大変難しい．たとえば，電話の声だけではすぐに話者を識別できず，困る場合がある．また，経験が浅い母親が乳児の泣き声から，その要求を理解できず，それが原因で育児放棄や育児ノイローゼになる場合もあるともいわれている．このような実用的なことだけでなく，もし野鳥の地鳴きを聞き分けられたり，どんな楽器の音かを判定できたりすれば生活が豊かに楽しくなるとも思われる．この論文では，音，特に音色☆を聞き分ける練習ができる装置の構築を目指している．また，e ラーニングに代表されるように，最近はコンピュータを駆使した学習がさかんである．たとえば，音に関した学習の代表に外国語の練習システムがある．外国語の練習システムでは個人差などの言語外の情報には注目せず，もっぱら音韻の正確さ，韻律の正確さを標準に合わせる訓練が一般的である．また，最近は，犬の鳴き声☆☆や乳児の泣き声☆☆☆を識別する装置が市販されている．これらの装置は，用途によっては便利な場合もあるが，判定を機械に委ねるという面から利用範囲は自ずと限定されてくる．一方，音色の識別を人間が学習する場合は，学習すること自体が楽しく，発展や応用の可能性も広がって，人間自身の能力を向上させることになる．これまで，音色識別の訓練は，1 人で，聴覚を使って学習することが

† 山梨大学大学院医学工学総合教育部
Interdisciplinary Graduate School of Medicine and Engineering, Univercity of Yamanashi

☆ 物理的には音の三要素の 1 つであるが，この論文では音の高さ，大きさなども含む「音が持つ感じ」を音色と表現している．

☆☆ TAKARA, バウリンガル：http://www.takaratomy.co.jp/products/bowlingual/

☆☆☆ WhyCry，赤ちゃんの泣き声識別器：http://www.whycry.com/

基本であった．しかし，筆者らがこれまで従事してきた音色識別学習の一例である「乳児の泣き声識別の研究[1)~4)]」では，乳児の泣き声を学習する場合，振幅や基本周波数などの視覚表示があり，かつ医師や看護師，他の母親などと共同して学習を行う方が効果的であると推測される場面が多々あった．そこで，この研究では，まず，音色の識別を学習する場合に人間を支援するシステムはどのようなものが適当かを検討した．実験の結果，聴覚情報のほか，視覚情報があると学習が効果的にできること，1 人で学習するより 2 人で学習する方が楽しく，効果も上がることが分かった．さらに，ヒューマンマシンインタフェースの 1 つに音声対話機能が要求されていることが分かった．そこで，この研究では，音声対話機能を備えており，2 人で共同して音色の識別を学習できる「学習支援システム」の構築を目指す．

2. 学習支援システムの構成方法

音色識別の一例に，話者識別がある．そこでは，似た話者の声を聞き分けることは，一般に大変難しい．しかし，普段の生活の中で比較的容易に音色を聞き分けられる場合がある．たとえば，太鼓の音や鐘の音などは，すぐにその音色を識別できる．音色を記憶しておくメカニズムの解明は難しいが，言語と対応づけて音色を記憶しているということが考えられる．たとえば，太鼓の音は「ドンドン」という言語情報と対応して記憶されており，鐘の音は「ゴーン」という言語情報と対応して記憶されている．そこで，たとえば，「ゴーン」に対応づけられる音を聞くと鐘の音であると判断されるのである．

人間の声の音色の識別はこのように単純ではないが，一般にいわれる「特徴のある声」については，その特徴（ガラガラ声，低い声など）をすぐにあげることができる．このように，普通の声についても，いろいろな特徴（高い声，低い声，小さい声，かすれた声など）を使って，その音色を表現できるものと考えている．そこで，下記のような仮説を考えた．

1) 人間は音色の特徴を言語で表し，記憶している．
2) 特徴の種類が多いほど，音色を正確に記憶できる．
3) 聴覚のほか，視覚情報も使用するとたくさんの特徴を抽出できる．
4) 複数人で相談すると 1 人で分析するより多くの特徴を抽出できる．

つまり，複数人で相談しながら，視覚情報も使用して音色識別の学習を行えば，効率が良い学習ができると考えている．この考え方を直接証明することは難しいが，視覚情報の利用や複数人の共同学習で学習の効果が上がることを示し，工学的な立場から，間接的に上述の仮説の妥当性を示したい．

声で話者を識別する場合，発話内容が同じなら，「その声が持つ感じ（音色）」で話者を識別すると考えられる．そこで，この論文では音色識別の例として，声による話者識別を対象にする．具体的には，データベースがしっかりしていること，話者識別が比較的難しいといわれていることなどを考慮して，成人女性の音声を使うことにする．実際には，「成人女性の音声の音色を学習して，テスト音声を聞いて話者を識別する」という実験を通して，前述の仮説を工学的な立場から確かめることにする．同時に，学習支援システムのヒューマンマシンインタフェースに関する調査も行う．具体的には，音色を学習する際の「学習支援システムを作成する」という立場から，以下に説明するような調査を行い，この調査結果を基に，学習支援システムを設計，製作することにする．

- 音色識別学習は基本的には聴覚で行われるが，学習に視覚情報が役立つか？
- 音色識別学習は，1 人で行う方が良いか，または 2 人で行う方が良いか？
- ヒューマンマシンインタフェースには何が必要か?
- 学習支援システムは学習者のどのような要求に応えればよいか？

調査の内容を以下に示す．

2.1 視覚情報の有効性

音色識別学習の場合に，視覚情報が役立つか否かを調べる．具体的には話者の声を学習する際，視覚情報がない場合（音声のみ）と視覚情報がある場合（音声＋音声波形の外形・スペクトログラム表示）において，学習効果にどのような影響が出るかを調査する．視覚情報の音声波形の外形とスペクトログラムは，音声の大きさ，音色に対応した情報として採用している．音声の高さは基本周波数で表示することが望ましいが，スペクトログラムにある程度その特徴が出ることから，前記の 2 つの視覚情報にしている．

ここでは，筆者らの経験により，乳児の泣き声の識別や楽器音の識別実験ではユーザは比較的すぐにスペクトログラム表示に慣れ，その情報を利用したため，この研究でも同様の視覚情報を使用することにした．予備的な実験を行った結果，視覚情報を利用する効果が期待できそうであったため，従来の経験で使用してきた視覚情報をそのまま利用しているが，成人の声の場合，その変化が複雑であるため，基本周波数の表示も含め，さらに詳しい検討が必要であるかもしれない．

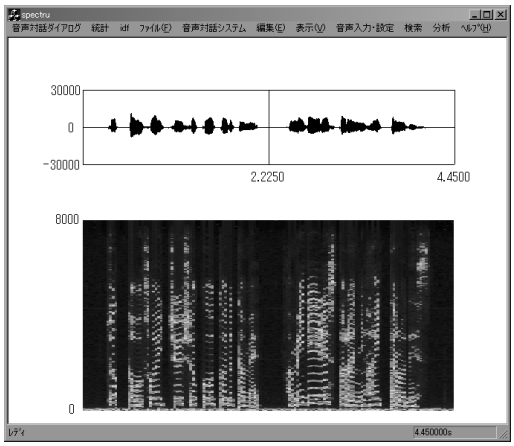

図 1 視覚情報の例
Fig. 1 An example of visual information.

提示する視覚情報の例を**図 1** に示す.

2.1.1 実験の方法

実験開始前に，下記のことを被験者に説明した.

- 話者の音色を学習して，話者識別をすること
- 学習用とテスト用の発話は発話内容が異なること
- テスト発話には同一話者の音声が複数個入っている可能性があること
- テスト発話は，学習した 5 人の音声いずれかに該当すること
- 学習中やテスト中に，メモをとってもよいこと

(1)「視覚情報なし → 視覚情報あり」の順序で実験

まず，成人の被験者 3 人（S1〜S3）に，日本音響学会の音声コーパス[5] に収録されている 5 人の成人女性（A〜E）の同一発話文を各音声間に 5 秒間のポーズを空け，2 回ずつ聞いてもらい，その音声の特徴を覚えてもらった．学習用音声の内容を**表 1** に示す．実験に使用する音声はいずれも東京型のアクセントであるが，若干の個人差がある．以下の実験で使用する試料も同様である.

約 3 分後，テストとして，被験者（S1〜S3）に話者 A〜E の別の 2 種類の発話文（合計 10 発話，表 1 参照）をランダムに聞いてもらい，どの話者かを回答してもらった．つまり，被験者ごとに 10 発話からランダムに選ばれた 10 個の音声を聞くことになる．同じ音声が複数回提示される場合もあり，ある音声は提示されない場合もある．この手法だと被験者ごとに提示音声の差があるが，全体として使用されたテスト発話の差は少ない．なお，被験者に正解を示すことはしていない.

1 週間後，同様の実験を再度行った．ただし今度は，学習用音声を聞くときに，音声波形の外形とスペクトログラムが表示されるようにしている．テスト音声は前述と同様のもので，音声のみ聞かせている.

表 1 2.1 節の実験で用いられた発話内容
Table 1 The utterances used by experiments in Section 2.1.

学習音声	「ちょっと遅い昼食をとるためファミリーレストランに入ったのです」
テスト音声	「あらゆる現実をすべて自分の方へねじ曲げたのだ」 「見上げる藤もいいが，路地植え，鉢植えの花もきれいです」

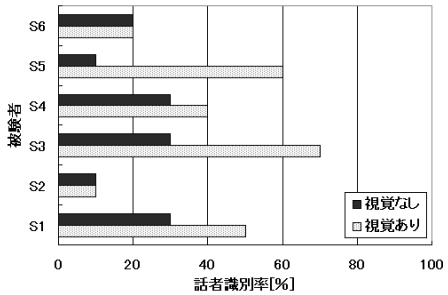

図 2 視覚情報の必要性に関する調査結果
Fig. 2 Investigation results of necessity of visual information.

(2)「視覚情報あり → 視覚情報なし」の順序で実験

前述と同じ方法で，まず視覚情報ありの実験を行い，次に視覚情報なしの実験を行った．被験者は，前述とは別の 3 人（S4〜S6）である.

2.1.2 実験の結果と考察

実験の結果を**図 2** に示す.

実験結果から，視覚情報がある場合の方が，視覚情報がない場合より，話者識別率が約 20%高い．実験後の被験者に行ったアンケートに寄せられた意見によると，ポーズの存在に気づきやすい，発話速度を覚えやすいなどの情報が考えられ，視覚情報がある方がない場合より，学習効果が高いと推測できる．図 2 の結果について,「2 つの母比率の差に関する検定[6]」を用い，有意水準 5%で検定を行った結果，2 つの結果には有意差があることが分かった.

被験者は音声について深い知識はないが，工学部情報系の学生であるため，一般ユーザより，この種の情報に慣れている可能性がある．しかし，この実験は音色識別学習支援システムの構成方法を検討するための実験なので，この実験で得られた知見はある程度役立つと考えている．2 章中の以下の実験も同様である.

2.2 2 人学習の有効性

音色を学習する場合，1 人で学習する場合と 2 人で学習する場合で，学習効果に差が出るか否かを調査す

表 2 2.2 節の実験で用いた発話内容
Table 2 The utterances used by experiments in Section 2.2.

学習音声	「音声研究会はどこで開かれるのですか」
テスト音声	「音声研究会に出席する予定ですが行き方が分かりません」 「機械振興会館です」

る．前節の結果をふまえ，ここでは視覚情報も表示できるシステムで学習を行う．ただし，テスト時には音声のみである．

2.2.1 実験の方法

実験を開始する前に，前節と同様なことを被験者に説明した．

前節の実験とは別の被験者 20 人を 2 つのグループに分け，S10〜S19，S20〜S29 とする．また，前節の実験とは別の 5 人の成人女性（F〜J）の音声を対象とした．対象とした発話の内容を**表 2** に示す．

(1) 1 人学習による話者識別

S10〜S19 の被験者に，1 人で最大 5 分間，表 2 の成人女性（F〜J）の学習音声で，話者識別学習をしてもらった．学習は，制限時間以内なら何回繰り返してもよい．また，被験者の意志で自由に学習を終了できる．実際には，いずれの被験者も，5 分以内で学習を終えた．

次に，前節の実験と同様に，テストとして被験者に話者 F〜J の表 2 のテスト音声（合計 10 発話）からランダムに 10 個を抽出して聞いてもらい，話者 F〜J のどれに該当するかを回答してもらった．記憶に残る程度に差が出るか否かを調べるため，テストは，学習終了直後から 60 分後まで 15 分おきに実施した．被験者に正解を示すことはしていない．

(2) 2 人学習による話者識別

S20〜S29 の被験者にランダムに 2 人 1 組になってもらう．各組，最大 5 分間，成人女性（F〜J）の学習音声に対して，話し合いながら話者識別学習をしてもらった．学習は，制限時間以内なら何回繰り返してもよく，被験者の意志で自由に学習を終了できる．実際には，いずれの組も，5 分以内で学習を終えた．

次に，1 人学習と同様な方法でテストを実施した．学習は 2 人で行うが，テストは 1 人ずつ行っている．

2.2.2 実験の結果と考察

実験の結果を**図 3** に示す．

実験結果から，2 人学習の場合の方が 1 人学習の場合より平均で約 10%話者識別率が高く，また，2 人学習の場合の方が時間による話者識別率の差が小さい．つまり，2 人学習の場合の方が，学習効果がやや高い

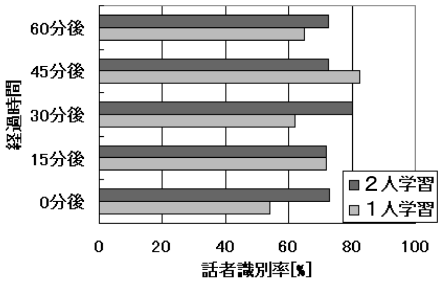

図 3 2 人学習の学習効果に関する調査結果
Fig. 3 Investigation results of the learning effect by two persons.

と推測できる．その差について，前節と同様の統計的な検定を行った結果，有意水準 5%で有意差があった．

時間が経過しても識別率が減少しない理由は，15 分ごとの短い時間に同じ音声を繰り返し聴取した結果，単に音声の比較照合になっている可能性もある．この部分についてはさらに詳しい検討が必要である．

2.3 2 人学習時の対話の分析

(1) 対話音声の収集

2 人学習時の音声を録音収集し，書き起こした．

(2) 被験者同士の会話を特徴づけるキーワード

たとえば，被験者は以下のような発話をしている．

「この声，暗い感じだねえ」

「ちょっと，低いよね」

「おばさんの声に似ているかも」

発話は，もう 1 人の被験者に対するもの，または自分自身に対するものである．全体としては，形容詞（低い，暗いなど）や比較対照（おばさんの声など）などを使って，音声に特徴をつけ，学習を行っていることが推測できる．2 人で音声と視覚情報を使いながら学習することにより，1 人で，音声だけで学習しているときには気づかなかった特徴を抽出でき，それが，話者の識別率の向上につながっていると考えられる．つまり，前述の仮説を間接的ではあるが，ほぼ確認できている．分析した対話に出現した，被験者同士の会話を特徴づける単語の例を大まかに分類して**表 3** に示す．これらを「音色識別のためのキーワード」と呼ぶことにする．

2.4 音色識別学習支援システムに必要なヒューマンマシンインタフェース

音色識別学習を支援するシステムに必要なヒューマンマシンインタフェースの調査，検討を行った．

2.4.1 調査の方法

2.2 節で述べた S10〜S29 の被験者に，テストが終了後，「音色を学習できるシステムでは，スピーカのほか，どのようなインタフェースが必要か，アンケー

表 3 音色識別のためのキーワードの例
Table 3 An example of keywords for tone color identification.

分類	キーワードの例
アクセント	「なまっている」
年齢	「年寄り」-「若い」
元気	「元気」「はきはきしている」-「暗い」「しょんぼりぎみ」「やる気がない」
読み方	「朗読みたい」-「自然」
高低	「高い」-「低い」
速さ	「速い」-「遅い」
明瞭	「こもっている」

表 4 インタフェースに関する調査項目
Table 4 Investigation items concerning interfaces.

選択肢	アイボールセンサ，音声対話，画面表示，キーボード，ジョイスティック，タッチパネル，タッチペン，マウス，画像入力（スキャナ），プリンタ

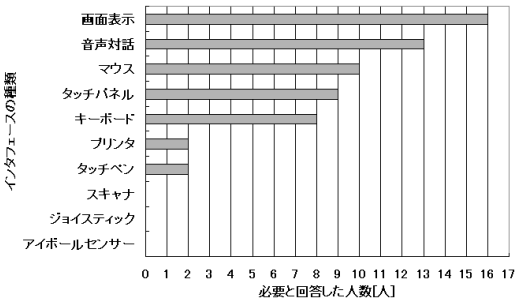

図 4 インタフェースの必要性に関する調査結果
Fig. 4 Investigation results of necessity of interfaces.

ト用紙に示す選択肢の中から 5 個以内で選んでください」というアンケート調査を行った．回答者は 17 人であった．

アンケートに載せたインタフェースの項目を**表 4** に示す．この他自由に必要なものを書き足せる欄がある．

2.4.2 調査の結果と考察

調査の結果を**図 4** に示す．

この調査結果で 17 人中 5 人以上が必要であると回答したインタフェースは，画面表示，音声対話機能，マウス，タッチパネル，キーボードであった．

被験者は 20 代の大学生で，普段から情報機器を使い慣れているので，上記のようなインタフェースが必要であると回答したと予想できる．情報機器の使用に不慣れな人の場合は若干違った結果が出るかもしれないが，ここでは，学習支援システムを構築するという立場から，情報機器の扱いに慣れた人の意見に従ってもよいと判断している．なお，アンケート直前の実験で使用したシステムのインタフェースはスピーカのほか，マウスと画面表示であり，結果に対しこのシステム構成の影響はほとんどないと考えられる．

表 5 インタフェースの用途
Table 5 Usage of interfaces.

インタフェース	デバイス名	主な用途
入力用デバイス	マウス	範囲指定や文字入力の補助
	タッチパネル	音声波形の範囲指定
	キーボード	データベース用文字入力
出力用デバイス	画面表示	音声波形などの表示
	再生用スピーカ	音声の再生
その他のデバイス	音声対話機能	学習の円滑化促進

音色学習実験において被験者の様子を観察した結果から，次のようなことが推測できる．

- 出力デバイスであるスピーカと表示用画面は，音の再生と波形の表示などに必須である．
- タッチパネルは，表示された波形を指しながら「ここからここまで聞きたい」など指示代名詞を含んだ音声入力と併用すると効率の良い入力が可能となる．音声波形の部分再生などにマウスやキーボードの使用も可能だが，学習に少し慣れてくるとタッチパネルを使用する頻度が高くなる可能性がある．
- マウスとキーボードは学習時にはほとんど使用されないと考えているが，データベースに文字データを登録する際などに使用できそうである．

次に音声対話機能について考える．ユーザが学習支援システムの操作に不慣れな場合，システムから「もう一度音声を再生しましょうか？」などの提案があり，「はい，お願いします」などの返事をして学習を進められるとユーザは大変助かる．また，システムの操作に慣れてくると，「2 番目と 3 番目の音声を連続で聞かせてください」などのユーザ主導型の複雑な入力も比較的簡潔に入力可能になる．また，タッチパネルのみの使用では数回のタッチ入力が必要な場合も，音声対話機能と組み合わせて使用することにより素早く学習を進めることも可能になる．つまり，音声対話機能は，学習支援システムを効率良く使用するために，欠かせない機能である．

以上のような観測結果より，**表 5** に示したようなヒューマンマシンインタフェースを音色識別学習支援システムに実装すればよいと考えている（再生用スピーカを含む）．

2.5 音声対話機能で学習者が使う命令について

学習者が音色識別学習支援システムの音声対話機能を使ってシステムに出す命令について調査を行った．

2.5.1 調査の方法

「音色識別学習ができるシステムの音声対話機能で，システムに対する命令項目として必要と思われるものを下記の選択肢の中から 5 個以内で選んでください」というアンケート調査を，前述の 2 人学習の被験者に，テストの終了後に行った（**表 6** 参照）．回答者は 17 人であった．なお，このほかにも必要なものを自由に書き足せる欄がある．

2.5.2 調査結果と考察

調査結果を**図 5** に示す．

まず，このシステムでは，基本周波数やホルマント周波数の表示など，専門的な情報も必要に応じて表示可能にしてあるが，一般の人が音色を学習する場合には，ホルマント周波数などの専門的な単語はほとんど使用されない．そこで，この調査で 17 人中 5 人以上が必要と回答した下記の 3 つの機能を音声対話機能で扱える命令とすることにした．

- 音声の入力命令（録音，編集など）
- データベースの操作命令（登録・削除，検索など）
- 学習時の命令（再生，部分再生，波形表示など）

以上の 3 つの命令のうち，「音声の入力命令」は学習のための音声や分析したい音声をシステムに入力するときに用いるもので，使用頻度は非常に低い．「データベースの操作命令」はすでに登録してある音声を処理する場合に使用される．学習の際には，「○○さんの音声を聞きたい」など，検索に関する命令が比較的多く使用される．学習時に頻繁に使用されるのは 3 番目の「学習時の命令」である．特に，「もう一度聞きたい」など，再生に関する命令が使われる頻度が高い．

表 6 音声対話機能の命令に関する調査項目
Table 6 Investigation items concerning the commands used in the speech dialogue functions.

選択肢	基本周波数（ピッチ），スペクトル（スペクトラム），スペクトログラム（ソナグラム），波形表示，パワー包絡，平均スペクトル，ホルマント（フォルマント），登録・削除，録音，部分再生，再度再生，複数連続再生，編集（カット，ペーストなど），検索，早送り・巻き戻し，一時停止，再生速度変更，選択部分リピート再生，その他（各自記入）

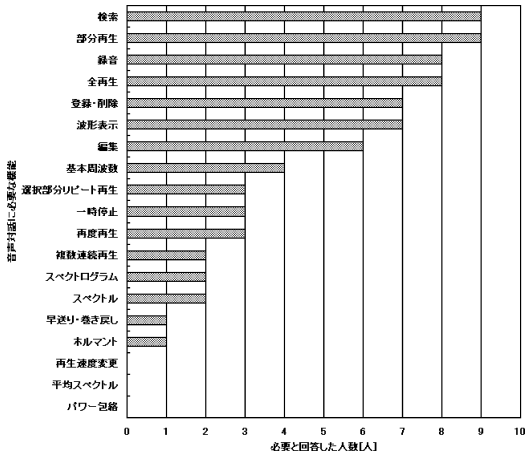

図 5 音声対話機能で使う命令に関する調査結果
Fig. 5 Investigation results concerning the commands used in the speech dialogue functions.

3. 音色識別学習支援システム

音色識別学習支援システムは，**図 6** に示すように「学習音の入出力」，「対話音声の入出力」，「対話処理」，「学習音のデータベース」，「音響分析」の 5 つのモジュールで構成されている．

「学習音の入出力」は学習のための音をデータベースに登録する場合などに使用する．この機能は，一般には学習支援システムの管理者が使用する場合が多いが，必要に応じて学習者が適当な音を登録削除することも可能である．「対話音声の入出力」は 2 人の学習者の音声をシステムに入力したり，システムからの応答音声を出力したりする機能である．システムと学習者の情報の交換は音声対話が基本であるが，学習者は必要に応じて，タッチパネル，キーボード，マウスなどを使用できる．「対話処理」は，学習者の発話を分析理解して，学習者に必要な情報を提供している．「学習音のデータベース」は学習に使用する音を整理して登録しておく場所である．音の登録，削除，編集などのためにインタフェースが用意されている．「音響分析」では，視覚情報表示のためにスペクトル分析などを行う．

図 7，**図 8** に，音色識別学習支援システムを使用し

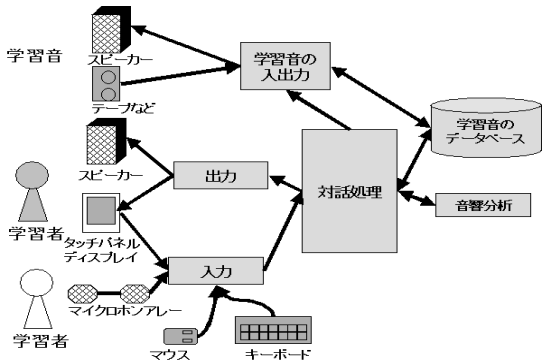

図 6 音色識別学習支援システムのブロック図
Fig. 6 A block diagram of the tone color identification learning support system.

学習者 A：○○さんの声はどうでしたかね？
学習者 B：そーですね．○○さんの声を聞かせてください．
システム：○○さんの声を再生しましょうか．
学習者 B：はい，お願いします．
⇒<○○さんの声再生>

図 7 直接的命令を含む対話例
Fig. 7 An example of a dialogue including a direct command.

学習者 C：この辺は結構高い声ですよね．
学習者 D：この辺を聞いてみましょうか．
システム：再生しますか．
学習者 D：ええ．
⇒<部分再生>

図 8 間接的命令を含む対話例
Fig. 8 An example of a dialogue including an indirect command.

たときの使用者とシステムの対話例を示す．図 7 は，学習者 A が○○さんの声を取り上げ，学習者 B がシステムに直接再生を要求している対話の例である．システムは，直接的な命令を受けなくても，音の再生や波形などの表示処理を行うこともできる．図 8 がその例である．

以下に，音色識別学習支援システムの各部について説明する．

3.1 学習音の入出力

(1) 学習音の入力

学習音は，DAT などの記録メディアに録音したデータを対象に学習を進める場合もあり，データベースに登録してあるデータを使って学習を行う場合もある．必要に応じて，入力データを編集して，データベースへ登録することも可能である．

(2) 学習音の再生

必要な学習音を再生して聞くことができる．タッチパネルなどと音声命令を併用して，指で指しながら「ここからここまで」などと発声し，任意の区間の部分再生をすることも可能である．

3.2 対話音声の入出力など

(1) 対話音声の入出力

基本的には 2 つのマイクロホンアレイを使って重複音声の分離を行うべきだが，まだ機能が不十分なため，学習者の位置の変化などに十分対処できない．そこで今回の実験では，各学習者は別々のマイクロホンの近くに位置し，移動しないものとする．音声応答用スピーカも用意してある．

(2) その他の入出力装置

ディスプレイに音の波形とそれに対応するスペクトログラムを表示して，タッチパネルやマウスなどのポインティングデバイスで指示しながら音声で部分再生命令や表示命令を出すと，部分的な再生，表示や拡大表示などを行うことができる．

3.3 音声対話機能

開発した音色識別学習支援システムは，円滑な学習を可能にするために，学習者と音声で簡単な対話ができる機能を備えている．以下に詳細を説明する．

(1) 音声認識

音声認識エンジンには Julius[7] を使用している．この研究では，音響モデルは汎用のものを使い，認識用辞書と言語モデルは，タスクに合わせて作成している．

言語モデルと認識用辞書を作成するために，Wizard of Oz 法[8] で対話音声を収集した．収集した対話の発声者は学生 20 人，2 人 1 組で 5 分以内で，模擬的に話者識別の学習をしたものである．

収集した対話より，言語モデル作成ツールキット Palmkit[9] を用いて，語彙数 100 のトライグラム言語モデルと認識用辞書を作成した．また，この収集した対話を参考に意味理解で使われる規則なども作成した．

音響モデルは，連続音声認識コンソーシアム 2001 年度版[10] に付随している「ATR 多数話者音声データベース」などにより学習した音響モデルのうち，5000 状態 64 混合の性別非依存のトライフォンモデルを使用した．

音声認識の性能を調べるため，10 人のユーザからこのシステムでよく使われていると考えられる「ここからここまでを聞かせてください」などの 13 種類の発話を収集し（同じ文章を 3 度発話しているため総発話数は 390），音声認識実験を行った．その結果，単語正解精度は 93.5%，文認識率は 86.7%であった．なお，これは実際に音色学習支援システムを使用しているときの発話ではないため，実環境においては若干認識率が低下することが考えられる．

(2) 意味理解

音声認識結果に対して，形態素解析器「茶筌」[11] を使用して，単語の品詞決定を行う．一方，開発しているシステムでは，構文情報の曖昧さなどを考慮して，キーワードベースの意味解析を行っている．キーワードと意味属性の関係の例を**表 7** に示す．表 7 を参照することにより，品詞情報が付与された単語から，キーワードを抽出する．そして，**表 8** を参照することにより，キーワードの組合せと話者の要求の関連から，学習者の意図を推定する．表 7 と表 8 は，対話例から

表 7 キーワードと意味属性の例
Table 7 An example of keywords and semantic attributes.

キーワードの意味属性	キーワードの例
＜検索ワード＞	学習対象者名（○○さん，△△さんなど）性別（男性，女性）
＜処理ワード＞	スペクトログラム，音声波形
＜入力：名詞＞	入力，インプット
＜表示：名詞＞	表示
＜再生：名詞＞	再生
＜位置：名詞＞	始め，終り
＜音声：名詞＞	音声，声
＜入力：動詞＞	入れる
＜表示：動詞＞	見る
＜再生：動詞＞	聞く
＜問い合わせ：動詞＞	教える，知る
＜指示代名詞＞	それ，これ，ここ
＜あいさつ＞	おはよう，こんにちは
＜肯定＞	はい，ええ，そう
＜否定＞	いいえ，ちがう
＜疑問：助詞＞	か

表 8 発話の意図と意味属性の関係例
Table 8 An example of the relation between the intentions of utterances and the semantic attributes.

学習者の発話例	キーワードの組合せ例	学習者の要求
この音声をシステムに入力したい	＜音声：名詞＞＜入力：名詞＞	＜入力命令＞
○○さんの声を聞きたい.	＜検索ワード＞＜音声：名詞＞＜再生：動詞＞	＜再生命令＞
△△さんの声を見たい.	＜検索ワード＞＜音声：名詞＞＜表示：動詞＞	＜表示命令＞
音声波形を見たい.	＜処理ワード＞＜表示：動詞＞	＜表示命令＞
「はい，そうです.」「お願いします」	＜肯定＞	＜肯定＞
「いいえ」「ちがいます」	＜否定＞	＜否定＞

人手で作成している．

表 8 のように，システムが理解できる学習者の音声対話命令は，入力に関するもの，再生に関するもの，表示に関するもの，肯定・否定に関するもの 4 種類に分類できる．

(3) 応答音声合成

フリーの音声合成エンジン☆を使用して応答音声を出力している．

(4) 2 人学習に対する処理

機械と人間の 1 対 1 の対話に関してはすでに研究が進んでおり，その手法も確立されつつあるが，機械と複数の人間との対話の実現にはまだ様々な解決すべき問題が存在する．しかし，この研究では，ユーザを 2 人に限定し，タスクも音色の学習に限定して，実用的な観点から対話システムを構築している．

1）重複音声の分離と認識

ユーザが 2 人になると音声が重複する場合があり，その完全な分離は非常に難しい．また，重複した音声の認識率が下がることはよく知られていることである．このタスクでは,「学習する」という目的があり，学習者はよく考えながら発言する場合が多い．よって，2 人の音声が大きく重なることはあまりない．また，発話の理解のために，キーワードの組合せを利用しているが，キーワードが重複して認識率が低下する場合もほとんどない．また，学習者が移動する場合も少ないので，システムとユーザの位置関係はほぼ一定である．よって，2 つのマイクロホンによるマイクロホンアレーで学習者の方向を見つけて，学習者の区別をしてはいるが，重複音声の分離に関する処理はせずに，音声認識処理を行っている．

実際には，各マイクロホンからの入力に対し，つねに音声認識を行っており，キーワードを抽出し，キーワードの組合せで，ユーザの要求などを理解し，処理を進めるようにしている．

2）発話者・発話対象者の識別

この音色学習支援システムでは，発話者，発話対象者は原則として移動しない前提で，下記の方法で，発話対象者の識別を行っている．

- システムへの発話と判断すれば，発話対象者をシステムとする．
- システムへの発話以外なら，発話者以外の人を発話対象者とする．

具体的には,「聞く，見る」などのキーワードがあればシステムへの命令，つまりシステムへの発話と判断している．またこれらのキーワードがあっても，前述の学習者間の対話を特徴づけるキーワード（表 3 参照),「元気な」「はきはきしている」「暗い」「若い」などが出現した場合は学習者同士の会話としている．

3）2 人のユーザとの対話処理

一般的に複数ユーザの対話解析を行うためには，非常に複雑な対話管理機構をシステム内に構築することが必要になる．しかし，ここでは，タスクの性格を考慮して，音声対話の履歴を残して，それを次の発話の理解に利用する程度の簡単な対話管理手法をとっている．

たとえば「もう一度聞かせてください」などの命令に対応し，対話の履歴から対象音声を判断し，システムは，対象となる音声を再生することができる．また，音声認識結果があいまいな場合の確認や，意味理解の

☆ Microsoft「Text-to-Speech Engine，SAPI 4.0a」と西村誠一ブランドソフトウェアシリーズ「読み上げ TOOL Ver 0.5」

表 9 意味解析のためのフレーム構造の例
Table 9 An example of a frame structure for semantic analysis.

項目	要求	対象	開始位置	終了位置	学習対象者名
内容	再生	音声			○○さん

ために不足している情報の補完のための質問を行うこともできる．

音声対話の履歴として，具体的には，学習者の命令ごとに意味解析のための意味解析フレームを作成し，それを時間に沿って残しておく．**表 9** は，「○○さんの声を聞かせてください.」の要求に対する意味解析フレームの例である．

4. 音色識別学習支援システムの使用例と評価

話者識別学習に，開発したシステムを使用する場合の事前の準備と学習過程の例を以下に示す．

4.1 準　　備

1）必要に応じて，学習に使う試料をデータベースへ登録する．この例では，システムの管理者があらかじめ登録しておいた 5 人の学習用音声を，話者識別の学習に使用する．

2）2 人の学習者は，「データベースに入っている 5 人の音声の音色の個人差を 5 分以内で学習してください」と実験者から言われ，システムの使用を開始する．

4.2 学習過程の例

学習支援システムのスタートキーを押すと，システムが使用できるようになる．学習者が何もしないと，システムが「話者 A の声を再生しますか?」と発声するので，必要なら学習者が「はい」と答えて学習が始まる．たとえば，学習中には，下記のような会話が行われる．

システム：話者 A の声を再生しますか？
学習者 X：はい，お願いします.
→ 話者 A の学習音声が出力される．
学習者 Y：結構高い声ですね.
学習者 X：うん．話者 A の声をもう一度聞かせてください.
システム：話者 A の声を再生ですね.
学習者 X：ええ．→ 話者 A の学習音声が出力される．
学習者 Y：○○さんに似てますよね．次，いきますか？
学習者 X：なるほど．次にしますか.
学習者 Y：話者 B の音声を聞かせてください.
システム：話者 B の声を再生ですね.
学習者 Y：はい，そうです.
→ 話者 B の学習音声が出力される．
：
（同様の対話が繰り返され学習が進む）
：
学習者 Y：だいたい分かりました.
学習者 X：では，学習終わります.

以上の例で被験者は，たとえば，「話者 A の声は，高い声で，○○さんの声に似ている」という特徴などを使って，話者 A の声を覚えているものと推測できる．

4.3 評 価 実 験

学習支援システムの効果を確かめるために，以下の 2 つの方法で，成人女性 5 人の学習用音声について音色を学習し，話者識別実験を行った．被験者ごとの学習時間は 5 分以内とした．また，学習音声，テスト音声は表 1 と同じ試料で，学習は 2 人ずつで行っている．

1）学習支援システムを使用しない学習（被験者：5 組）

パソコンのディスプレイに表示されている 5 種類の学習用音声のマークをクリックすれば，指定の音声が再生でき，それを使って各自，自由に音色の学習を行う．

2）学習支援システムを使用した学習（被験者：5 組）

作成した音色学習支援システムを使用して，成人女性 5 人の音声の音色の学習を行う．

2 つの実験の学習者は，それぞれ 5 組ずつ別々の被験者で，合計 20 人である．学習後，20 人の被験者に対して，それぞれ，テスト音声から 10 個をランダムに選んで聞かせ，話者識別の実験を行った．その結果，学習支援システムを使用しない学習者の平均話者識別率は 22%，学習支援システムを使用した場合は 72%であった．データの数がやや少ないが，学習支援システムの効果が確かめらている．また，被験者から得られたアンケート結果から，開発した学習支援システムにおいて，部分的な波形の再生や表示が容易に行えることなどが，学習に大変役に立っていることが分かった．

5. む　す　び

音色識別の学習の支援に関して様々な調査を行い，その結果，音声対話機能の必要性や視覚情報の利用と 2 人学習の効果を確認した．それをふまえ，音声対話機能を備えており，2 人で共同して使える音色識別学習支援システムを構築し，その有効性を確かめている．この論文では，音色学習支援システムの構築に主眼をおいているが，今後は，学習支援システム使用の効果を「学習」の面から詳しく検討する必要がある．

なお，ここで開発したシステムは，対話理解のための辞書と学習音の入れ替えで，話者識別の学習のほか，たとえば，乳児の泣き声の学習，野鳥の地鳴きの識別学習，楽器音の識別学習などいろいろな分野の音色学習支援に役立つと考えられる．

利用可能な情報が多ければ学習効果が高いことを被験者実験により示したが，情報が多い分学習者の負担が増加することも考えられるため，学習者の疲労との関係の調査を今後行う予定である．

謝辞 実験の実施にあたり，山梨大学大学院医学工学総合教育部大学院生の田中宏和氏に大変お世話になりました．ここに記し深謝します．

参 考 文 献

1) 田中宏和，渡辺裕太，関口芳廣，西崎博光，望月初音，西脇美春：乳児の泣き声の収集とその分析ツール，情報処理学会第 66 回全国大会講演論文集，3Y-4 (2004).
2) 渡辺裕太，田中宏和，関口芳廣，西崎博光，望月初音，西脇美春：乳児の泣き声データベースの試作，日本音響学会講演論文集，2-P-9 (2004).
3) 渡辺裕太，田中宏和，関口芳廣，西崎博光：新生児の泣き声学習支援システム，日本音響学会講演論文集，1-P-25 (2004).
4) 望月初音，西脇美春，渡辺裕太，田中宏和，関口芳廣，西崎博光：新生児の泣き声の音声分析による特徴との関係—空腹による泣きの分析，第 4 回日本赤ちゃん学会 (2004).
5) 板橋秀一（日本音響学会/編）：研究用連続音声データベース Vol.1〜3，日本情報処理開発協会 AI ファジィ振興センター (1991).
6) 内田 治：すぐわかる EXCEL によるアンケートの調査・集計・解析，第 2 版，東京書籍 (2002).
7) 鹿野清宏，河原達也，山本幹雄，伊藤克亘，武田一哉：音声認識システム，オーム社 (2001).
8) 岡本昌之，山中信敏：Wizard of Oz 法を用いた対話型 Web エージェントの構築，人工知能学会論文誌，Vol.17，No.3，pp.293–300 (2002).
9) 伊藤彰則，好田正紀：単語およびクラス N-gram 作成のためのツールキット，電子情報通信学会技術研究報告，SP2000-106，pp.67–72 (2000). http://palmkit.sourceforge.net/index.html
10) 河原達也，住吉貴志，李 晃伸，武田一哉，三村正人，伊藤彰則，伊藤克亘，鹿野清宏：連続音声認識コンソーシアム 2001 年度版ソフトウエアの概要，情報処理学会研究報告，2002-SLP-43-3 (2002).
11) 松本裕治：形態素解析システム「茶筌」，情報処理，Vol.41，No.11，pp.1208-1214 (2000).

(平成 17 年 10 月 12 日受付)
(平成 18 年 4 月 4 日採録)

渡辺 裕太

平成 13 年山梨大学工学部電子情報工学科卒業，平成 15 年同大学大学院工学研究科博士前期課程電子情報工学専攻修了，現在，同大学院医学工学総合教育部博士課程在学中．音声対話システムの研究に従事．

関口 芳廣（正会員）

昭和 46 年山梨大学工学部電子工学科卒業，昭和 48 年同大学大学院修了．同年同大学工学部計算機科学科助手，現在，同大学大学院医学工学総合研究部教授．音声情報処理等の研究に従事．工学博士．信学会，音響学会，電気学会等会員．

西崎 博光（正会員）

平成 10 年豊橋技術科学大学工学部情報工学科卒業．平成 15 年同大学大学院修了．現在，山梨大学大学院医学工学総合研究部助手．音声言語情報処理に関する研究に従事．博士（工学）．音響学会会員．