話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定

峯松信明[†] 広瀬啓吉[††] 関口真理子[††,☆]

対話システムの高度化にともない，入力音声から単に言語情報（文字情報）を抽出するだけでなく，話者性や感情など，話者の静的および動的特徴を的確に把握しながら効率的に対話を遂行することを目的とした研究が行われるようになってきた．本論文では種々の話者特性の中でも「年齢」に焦点を当てる．特に高齢化社会を考慮し，音声の音響情報より高齢話者を特定する手法を提案する．本論文ではまず，高齢話者音声データベースに対して聴取実験を行い，「高齢者であることを意識した対応が必要である」と考えられる話者を特定した（主観的高齢者）．先行研究より高齢者としての特徴がスペクトル情報に反映されるとの知見があるので，主観的高齢者の同定を話者認識技術を利用して行った．その結果，約91%の正答率が得られた．さらに，聴取実験の結果得られた「高齢者としての対応が必要である」と判断した理由について分析し，スペクトル情報以外の音響情報である韻律的特徴を実験的に検討した．その結果，話速とパワーの局所変動を考慮することで，同定率を約95%まで向上することができた．また，提案手法に基づいて，発話者に対する主観的年代の自動推定に関する分析を行ったのでその結果についても報告する．

Automatic Identification of Subjectively-defined Elderly Speakers and Its Application to Estimating Agedness

Nobuaki Minematsu,[†] Keikichi Hirose[††] and Mariko Sekiguchi[††,☆]

Recent advancement of spoken dialogue systems requires techniques not only to recognize users' utterances, but also to capture their static and dynamic characters, with which more efficient and fruitful dialogue between humans and machines can be realized. In the current paper, research focus is placed upon speakers' agedness as one of the static characters and a method of automatically identifying elderly speakers only with their voices is proposed. Firstly in this paper, a listening test was done for JNAS and S-JNAS databases where subjects were asked to estimate each speaker's agedness subjectively and judge whether the subjects should take special care of their speaking styles when talking to the speakers. Secondly, a series of experiments were carried out to automatically identify the subjectively-defined elderly speakers. In the first experiment, GMM-based speaker recognition techniques were immediately used and 91 % accuracy was obtained. Through experimental examinations of various prosodic features, speech rate and local power perturbation were added to the GMM-based identification in the second experiment. The performance was raised up to 95 %. Finally, a method was also devised to estimate speakers' agedness using the proposed techniques. A rather high correlation between the agedness estimated by the method and that obtained by the subjective listening test indicates the high validity of the method.

1. はじめに

近年の計算機技術の発展に支えられ，実社会で利用可能な音声対話システムを念頭に置いたシステム開発が行われるようになってきた[1),2)]．これらの対話システムは，ユーザ発話における言語内容（すなわち文字情報）を認識，理解し，その結果からユーザの意図を解釈し，その意図に応じた情報提供をするという対話戦略に基づくものが多い．このような枠組みで，ユー

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†† 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo
☆ 現在，NTT-IT
Presently with NTT-IT

ザが（固有に）持つ「特徴」を抽出することを考えた場合，当然発話内容から抽出せざるをえない．人間対人間の対話，特に音声のみによる対話を考えた場合，相手の静的特徴（個人性，性別，年齢，出身地，目上/目下，性格，職業など）や動的特徴（精神状態，感情など）に応じて発話スタイルやマナーを適切に変化させて対話を進める様子が自然に観測される[3]．このような観点から対話システム研究/構築を眺めた場合，発話者切替えの検出[4]と話者の同定[5]，システムの誤動作に起因するユーザ発話の韻律的変動の抽出[6),7]，感情認識[8]，母国語の推定[9]など，研究レベルにおいていくつかの試みが行われている．また，人間対人間の音声コミュニケーションにおいて，相手のどのような特徴に応じて，対話戦略を修正/変更しているのかを分析的に検討した研究もある[3]．

これら種々のユーザ特性のうち，本論文では「年齢」に焦点を当てる．音声の音響情報に年齢の情報が反映されることは，種々の研究例で報告されているが[15)〜18)]，逆に，音声の音響情報から年代を推定する研究例は，筆者らが事前調査した限りにおいては前例がない．その原因の 1 つとしてデータベースの整備があげられる．近年種々の音声データベースが構築，配布されるようになったが，年代別の音声データベースは非常に少なく，日本では高齢話者音声データベース S-JNAS[10] や，子供音声データベース CIAIR-VCV[11]があるのみであり，かつ，これらの配布も 2001 年になって開始されている．

年代の推定に対する実用的価値を考えた場合，自ずと高齢化社会というターゲットが考えられる．21 世紀の日本は高齢化がさらに進み，4 人に 1 人が高齢者になるともいわれている．高齢者にとって使いやすいシステムの実現は，より快適な社会の形成に役立つと考えられる．特に音声のように，その使用に特別の訓練を要さないメディアに基づくマン・マシンインタフェースの構築は，高齢者のように新しい技能の取得が困難なユーザには不可欠なものであるといえる[12]．このようなユーザフレンドリなサービスを実現する 1 つの手段として，システムがユーザの年代を推測し，ユーザの年代に応じた GUI の制御，対話戦略や情報提供戦略の変更，さらには，（音響モデル/言語モデルの変更といった）システム構成の制御をすることが考えられる．ユーザを煩わせることなくこのようなシステムを実現するためには，発話内容によらない音響情報のみから年代同定をする必要がある．特に高齢者を対象とした場合は，高齢者か否かの識別が必要になる．以上の考察より本研究ではまず，音声データベース内の話者について，人間の耳による主観的判断によって「高齢者であることを意識した対応が必要である」と考えられる話者（主観的高齢者）を特定した．その後，主観的高齢者を自動同定することを目的として一連の実験を行った．実験ではまずスペクトル情報に注目し，話者認識の技術を直接利用した高齢者同定を試みた．さらに，高齢者の特徴として他の音響的特徴にも注目し，その中でも特に主観的高齢者，非高齢者間において顕著な差が見られた話速とパワーの局所的変動の情報を考慮し，主観的高齢者同定精度の向上を試みた[13]．

なお，あらかじめ話者情報を登録する話者認識の枠組みが導入可能の場合は，高齢/非高齢の判定を行わず，話者認識結果より話者情報を検索し，話者の実年齢を得る方法が最も確実である．しかしこのような事前の登録を仮定すると，その応用範囲が極端に狭まる（たとえば家庭内だけ，など）ことが予想される．本研究で検討する主観的高齢/非高齢者判定は，あらかじめ話者情報を登録する必要のない，未知話者に対応可能な主観的高齢/非高齢者判定を検討する．

2. 聴取実験による主観的年代推定

2.1 実験の目的

上記した機能の実現に必要な技術は，話者の実年齢を推定する技術ではなく，その話者の音声を聞いたときに，聴取者が発声者に対して感じる “主観的” な年齢/年代を推定する技術であると考えられる．よって本論文では，データベースによって提供されている各話者の実年齢データは無視し，聴取実験を通して被験者の耳による主観的な枠組みで年代を推定させた．

心理学の用語では，聴取者が感じる話者の年代を知覚的年齢（perceived age），話者の実年齢を生物学年齢（biological age）という言葉を用いて区別するが，本研究では混乱を招かないと判断される場合において，主観的年齢，実年齢という言葉を使用する．

2.2 実 験 手 順

被験者は本学大学生 12 名である．年代推定に使用した音声データは，JNAS データベース[14]中の男女各 150 名および，S-JNAS データベース（高齢者データベース）[10]中の男女各 200 名である．聴取実験はあらかじめ定めた PC とヘッドホン（両耳提示）を使用し，音量を一定にして Web 上で行った．被験者は各話者につき 1 文の音声を聴取し，主観的な年代に対するいくつかの質問に答える．なお，提示される文は発声者間で異なっている．この場合，主観判断が，発話文の言語的内容に引きずられることが考えられるので，同一発声者の文音声についても，被験者間で異なるよ

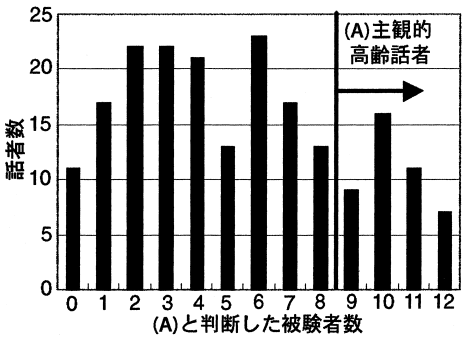

図 1 聴取実験の結果（男性話者）
Fig. 1 Results of the listening test using S-JNAS database (male speakers).

う実験を計画した．また，データベースには方言話者（主に関西）が多く混ざっているため，方言であることを理由に高齢者であるという判断はしないように教示した．

被験者が答える質問は以下のとおりである．まず，電話相談室のオペレータになったつもりで音声を聞いてもらい，その声の人物が電話をかけてきたときの対応として，(A)「相手が高齢者なので自分の言葉遣いや声色に気を遣って対応する必要がある」，(B)「特に対応に気を遣う必要はない」，(C)「どちらともいえない」の 3 段階のうちどれが適切かを選ばせた．と同時に，話者の年代を，(1) 20〜30 代，(2) 40〜50 代，(3) 60 代，(4) 70 代，(5) 80 代以上，の中から選ばせた．なお，極端にノイズが乗っていたり音量が小さすぎたりするため判断が著しく困難な音声（録音状態が極端に悪いと考えられる音声資料）に関しては，そのつど申告してもらい，以下の主観的高齢者同定実験では使用しなかった．さらに，実験後に高齢者と判断した音響的要因について別途アンケート調査（自由記述）した．

2.3 実験結果

S-JNAS の話者に対する聴取実験の結果を**図 1** に示す．横軸を N_x 縦軸を N_y とすると，N_x は当該話者が（A）であると判定した被験者数を示し，N_y は N_x 人の被験者によって（A）であると判定された話者数を示す．S-JNAS データベース内の話者はすべて 60 歳以上であるが，ほとんどの被験者が「特別な対応が必要である」と判断した話者は比較的少なかったことが分かる．実験の結果，12 人中 9 人以上が（A）と判定した話者 43 名を主観的高齢者と定義した．主観的非高齢者については，JNAS データベースからランダムに 43 名分抽出した．3 章以降の実験では，与えられた音声に対して，主観的高齢者/主観的非高齢者のいずれかへ正しく同定することを目的とする．

表 1 アンケートに対する典型的な回答例
Table 1 Typical answers to the questionnaire.

話す早さが遅い
声が震えていたりかすれていたりする
声に力がない
ろれつがまわっていない
つまることが多い

なお，「高齢者判定時に参照した音響的特徴」に関する調査であるが，典型的な回答例を**表 1** に示す．(A) と判断する要因として，話速や声の震えも大きく影響していることが分かる．これらの知見の利用に関しては 4 章で実験的に検討する．

3. 話者認識技術を利用した高齢話者の同定

一般に高齢者の音声認識率は下がるが，これに対し高齢者データでモデルを再学習/適用することで，認識率が改善されることが知られている[17),18)]．また，高齢者は高周波成分が小さくなるという報告[15),16)]もある．これらのことから，スペクトル情報に高齢者らしさ（の一部）が現れていると考えられる．一方，音声認識技術をベースとし，スペクトル情報を GMM でモデル化することで話者を認識する方法が広く使用されている[19),20)]．主観的高齢者の同定というタスクは，話者グループ同定の 1 つと位置付けられることもあり，本論文ではまず，話者認識技術を直接利用した主観的高齢者同定の性能を検討した．

3.1 実験条件

モデルとしては，混合数 32 の GMM を使用した．話者認識では話者 1 名につき 1 つのモデルを作成するが，今回の主観的高齢者同定実験では，各々の話者グループに対してモデルを作成した．また，同定は，各話者グループのモデルに対する対数尤度差をフレーム長で正規化した値の正負に基づいて行った．実験は男女別々に行ったが，どちらも同じ傾向を呈したので，以下では男性話者を用いた実験について述べる．なお，高齢話者（ただし実年齢）音声を用いた音声認識実験[18)]では，「非」高齢者音響モデル利用時の認識性能劣化は男性の方が大きく，高齢者/非高齢者間の音響的差異は女性の方が小さいとの報告がある．本実験でも女性における同定性能劣化が危惧されたが，本実験では主観的に高齢者/非高齢者を定義しており，この枠組みにおいては，男性/女性における傾向差は小さかった．

実験条件を**表 2** に示す．なお，実験に使用した話者は，前節の人間の耳による主観的な分類によって得ら

表 2 実験条件
Table 2 Conditions of the experiments.

学習データ	JNAS（34 名 ×15 文）
	S-JNAS（34 名 ×15 文）
評価データ	JNAS
	（9 名 ×5 音声セグメント（各 5 秒ずつ））
	S-JNAS
	（9 名 ×5 音声セグメント（各 5 秒ずつ））
サンプリング周波数	16 kHz
分析窓	25 ms ハミング窓
フレーム周期	10 ms
プリエンファシス	$1 - 0.97z^{-1}$
特徴パラメータ	12MFCC+12ΔMFCC+ΔPOWER
GMM	対角分散共分散行列，混合数 32

れた 86 名（主観的高齢者，主観的非高齢者各々43 名）である．学習データとしては，各カテゴリ 43 名の話者のうち，34 名の発話を使用し，残りの 9 名を評価話者とした．また，43 名全員が評価話者となるように，学習，評価に使用する話者を移動させ実験を繰り返した（cross-validation）．よって，各話者グループに対して 5 つずつモデルが作成されている．なお，話者認識の分野で広く行われているように，モデル作成に際してはパワー値に基づく無音区間の排除を行い，音声区間のみでモデル化を行った．また，本研究で検討する話者グループのモデリングが，筆者らが意図した要因を反映したモデリングではなく，単に録音環境間の差異のモデリング（すなわちデータベース間の差異のモデリング）となってしまっていることが懸念されたが，種々の異なる音声データベースを用いた予備実験の結果，録音環境間の差異のモデリングとなっている様子は観測されなかった．

3.2 実験結果と考察

本論文で検討する技術を，対話システムの音声インタフェースや音声認識モジュールの切替えにおいて利用する場合，同定に要する発声長が長いのは望ましくない．そこでまず，入力音声長と同定性能との関係を見るために，評価データ長を変え実験を繰り返した．**図 2** に結果を示す．横軸は評価に使用したデータ 1 つあたりの時間長であり，縦軸は区切られた音声データ単位での誤認識率である．図から分かるように，発話長が 5 秒程度あれば結果が収束する．そこで，評価データ長を 5 秒前後に固定し，各話者につき 5 つずつデータを用意し，高齢者同定実験を行った．結果を**図 3** に示す．横軸は 5 データ中いくつのデータが正しく同定されたかを示し，縦軸はその人数を示す．この図より，同定性能を評価する尺度を 2 種類定義した．1 つは 5 データ中，3 データ以上が正解だった場合その話者は正しく判別されたとする話者単位での同定率

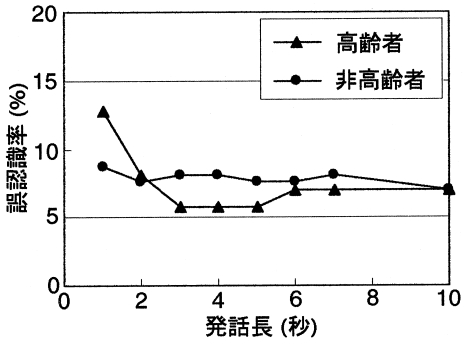

図 2 音声データ単位で集計した誤認識率と提示長との関係
Fig. 2 Misidentification rate as a function of speech length for a set of segmented speech samples.

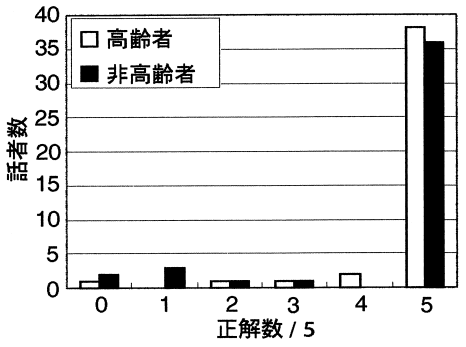

図 3 主観的高齢者同定実験結果
Fig. 3 Results of identifying subjectively-defined elderly speakers.

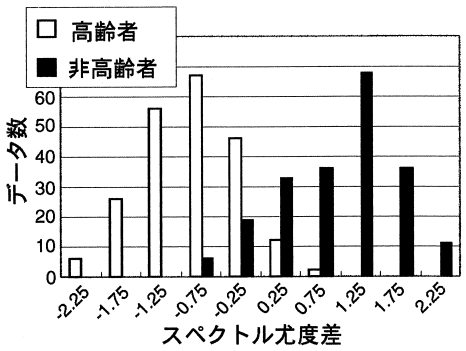

図 4 話者グループ間のスペクトル尤度差の分布
Fig. 4 Distribution of likelihood differences between the two speaker groups.

（話者同定率）であり，他方は約 5 秒単位で区分した音声データ単位での同定率（発話同定率）である．図より話者同定率は 90.7%となり（主観的高齢者 41/43，主観的非高齢者 37/43）比較的高い同定率が得られた．一方，発話同定率は 90.9%であった．

対数尤度差の分布の様子を調べると，2 つの話者グループの違いが**図 4** のように表された．誤認識され

た話者については，話者認識の技術を利用していることから，反対の話者グループの学習データに偶然似た声質の話者がいたためである可能性が考えられる．そこで，複数ある主観的高齢者モデルのうち図 3 の実験で使用したものとは異なるモデルで再実験をしたところ，やはり誤認識が起こることが分かった．また，実際に誤認識された音声を聞いてみたが，誤認識の理由を発見するには至らなかった．このことから，話者認識技術のみを利用した高齢者同定は，ある程度の同定率は得られるものの，その限界が示唆される．そこで以降では，現在の話者認識技術で利用されていない音響的特徴量で，高齢者同定に有効に寄与する特徴がないか，聴取実験時のアンケートを基に調べた．ここで特徴量としては，話者認識技術が声道特性の時間平均パターンのモデル化に基づいていることを鑑み，声道特性とは関連の薄い韻律的特徴に注目した．

4. 主観的年代の違いが反映される音響的特徴量

4.1 人間による年代同定の判断基準

2.3 節で述べたように，聴取実験時のアンケート結果では，発声者をグループ（A）と判定する理由として，声質のほかに，話速や声の震えといった韻律的特徴にも着眼していることが分かった．以下では，話速，パワー，ピッチの各々について主観的高齢者同定に有効に寄与するか否かを実験的に検討した．

4.2 主観的年代による話速の違い

一般に話速を求める場合，認識結果から得られる平均モーラ長が用いられることが多いが，この場合はすでに認識処理が終了していることが前提となる．本論文で検討している主観的高齢話者同定技術の応用として認識モジュールの制御を考えた場合，認識結果の利用は回避すべきである．そこで以下では，認識結果を用いずに推定した話速と，認識結果（連続モーラ認識結果）を参照する形で求まる話速を用いて検討した．

4.2.1 認識処理を必要としない話速とその利用

話速の定義 「話速が上がる」＝「単位時間内に出現する異なる音素間遷移の数が上がる」と考えるならば，話速の上昇はスペクトル遷移数の上昇となり，これは ΔMFCC ベクトルの大きさ（ノルム）のピーク数の上昇として観測される．以下では，ΔMFCC のノルムの単位時間あたりのピークの数を話速として定義する（**図 5** 参照）．ただし，ピークは閾値 θ_r 以上のもののみを数える．また，閾値 θ_r は全データに対し固定し，予備実験で主観的高齢者/主観的非高齢者間の分離度が最も高かった値を使用した．

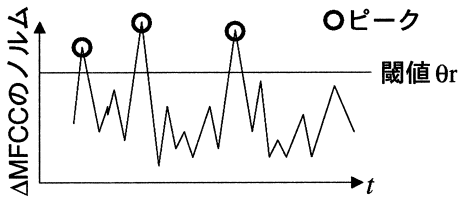

図 5 ΔMFCC ノルムピークを用いた話速の定義
Fig. 5 Definition of speech rate as the number of peaks of the norm of ΔMFCC vectors.

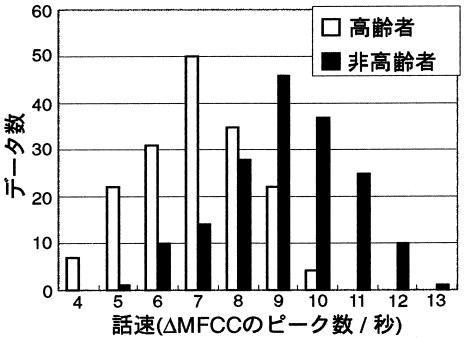

図 6 ΔMFCC を用いた話速の分布
Fig. 6 Distribution of ΔMFCC-based speech rate.

話速の分布 前章の実験と同じ話者を使用した．また予備実験において，話速の収束には少なくとも 5 秒程度は必要であることが分かったので，学習データ，評価データともに，前節で使用した 5 秒の音声データ（5 データ/話者）を用いた．学習，評価データをすべて（43 名 ×2 グループ ×5 音声セグメント）使い，上で定義した話速の分布を求めたところ**図 6** が得られた．話速の分布はおよそ正規分布に従うことが分かる．

話速のみを用いた同定実験 各話者グループ 43 名中 34 名のデータを学習データとして，話速分布に正規分布を仮定してモデル化し，話速尤度差のみに基づいて主観的高齢者同定を行った．評価実験は前節同様，学習に使用していない話者 9 名ずつ（×5 セット）を使用した cross-validation である．結果を**図 7** に示す．話者同定率は 83.7%（主観的高齢者 35/43，主観的非高齢者 37/43），発話同定率は 76.7%となった．

4.2.2 認識結果に基づく話速とその利用

話速の定義 連続モーラ認識結果から得られる単位時間あたりのモーラ数を話速と定義した．

話速の分布 前章と同じ方法で，全データの話速分布を求めた．**図 8** に示す．およそ正規分布に従うことが分かる．

話速のみを用いた同定実験 各話者グループに対して正規分布を仮定して話速をモデル化し，同定実験を行った．結果を**図 9** に示す．話者同定率は 87.2%（主

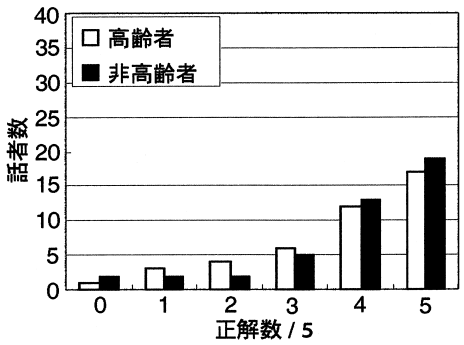

図 7 話速のみを用いた主観的高齢者同定実験結果
Fig. 7 Results of identifying subjectively-defined elderly speaker only with ΔMFCC-based speech rate.

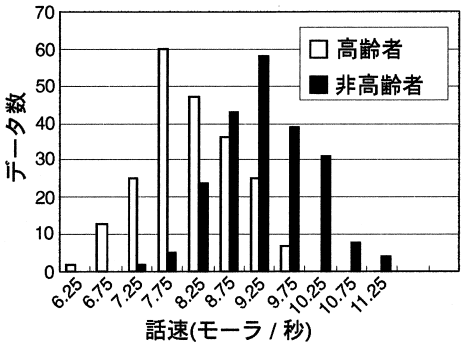

図 8 話速（単位時間当りのモーラ数）の分布
Fig. 8 Distribution of speech rate (#morae/sec).

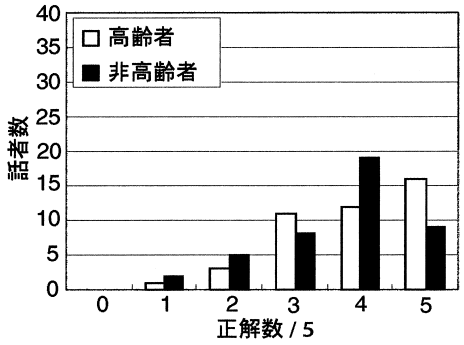

図 9 話速のみを用いた主観的高齢者同定実験結果
Fig. 9 Results of identifying subjectively-defined elderly speaker only with speech rate (#morae/sec).

観的高齢者 39/43，主観的非高齢者 36/43），発話同定率は 75.6%となった．ΔMFCC に基づく話速を用いたときの性能とほぼ同等であることが分かる．

4.2.3 考察と検討

話速だけの情報を用いて，話者単位で約 85%，発話単位で約 76%を正しく同定できたが，その性能は話者認識ベースの手法の方がはるかに高い．しかし，話者認識技術のみを使用した主観的高齢者同定では誤同定された話者が，本実験では正しく同定された例も見られた．このことから話者認識ベースの手法による認識結果と話速による認識結果を組み合わせることで，より高い同定率が得られることが期待できる．2 つの話速定義を行い両者を比較したが，認識結果を参照しない方法でも，認識結果を参照する方法と同等の性能を得ることができた．6 章で種々の音響パラメータの統合を行うが，実用的価値の高い前者の方法のみを使用する．なお，本実験に使用した音声（S-JNAS および JNAS データベース）は読み上げ文であり，実際に対話システムで話速情報を用いる場合は，対話音声の収集，分析をする必要がある．

4.3 主観的年代によるパワー情報の違い

4.3.1 パワーの平均，分散

聴取実験時のアンケートより，高齢者はパワーが全体的に小さいことが考えられた．そこで，パワーの平均，分散について調べたが，予備実験の結果，主観的年代による顕著な差異は検出されなかった．これは，実際に年代による差がないためとも考えられるし，データベースの音声収録時の話者とマイクとの距離や録音環境などの違いが影響している可能性も考えられる．いずれの場合でも，対話システム利用環境に起因する変動が大きい特徴量は，高齢者同定に対して有効なパラメータとは考え難い．

4.3.2 Δ パワーの分散値に対する平均，分散

主観的高齢者は主観的非高齢者と比較して，発声器官の制御機構の衰えから，発声の際のパワー変動の様子が小さく単調，いい換えれば，Δ パワーの分散が小さい，という傾向があると考えられる．Δ パワーの分散に対する平均と分散を求め，その分布の様子を示したのが**図 10** である．この図から，年代による差が観測されるが，年代差と比べて個人差も大きく，予備実験の結果 Δ パワーの分散の大小は，それほど有効なパラメータではないことが示された．なお，Δ パワーそのものは，話者認識技術に基づいた方法でもパラメータの 1 つとして使用されている．

4.3.3 パワーの局所変動

パワーの局所的変動の定義 聴取実験後のアンケート調査では，高齢者音声は「声が震えている」という意見があった．そこで，高齢者音声には，パワーの局所的 “揺れ” が頻繁に観測されるのではないかと考えた．**図 11** に同一文を発声したときの，主観的高齢者/主観的非高齢者のパワーパターンの様子を示す．この図からも，主観的高齢者のパワーパターンにおいて，局所変動が多いことが示唆される．パワーの大きな変

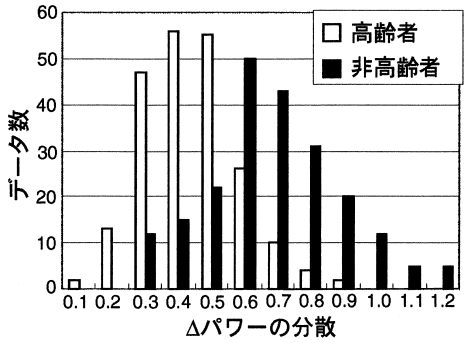

図 10 Δ パワーの分散の分布
Fig. 10 Distribution of Δpower.

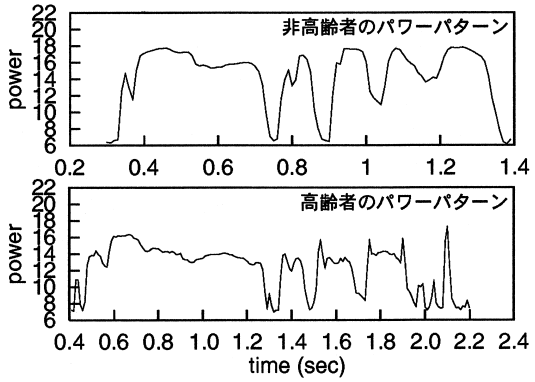

図 11 主観的高齢者（下図）と主観的非高齢者（上図）のパワーパターン
Fig. 11 Examples of power patterns of the two speaker groups (the lower figure is for subjectively-defined elderly speakers and the upper figure is for the others).

動は有声/無声区間，あるいは，音声/無音区間の境界とほぼ一致するが，この場合抽出したいのはパワーの局所変動のみであり，以下のようにパワーの局所変動を定義した．まず，パワーパターンから極値位置（山，谷両方含む）を抽出する．極値の系列は山，谷が交互に現れる形となる．次に，着眼している極値位置のパワー振幅値と，直前の極値位置における振幅値との差分の大きさを計算する．この「振幅差分の大きさ」が閾値 θ_p 以下の極値のみを対象とし，その数を音声時間長で割った値をパワー局所変動と定義した．**図 12** にその様子を示す．すなわち，極値位置（図中，○あるいは×で示された位置）に対して，直前の極値位置の振幅値と現在の極値位置の振幅値の差分が閾値 θ_p より小さいものが局所的変動（○）となり，それ以外を大局的変動（×）としている．ただし，θ_p は全データに対し固定し，予備実験で主観的高齢者/主観的非高齢者間の分離度が最も高かった値を使用した．

パワーの局所的変動の分布 全データのパワー局所変動の分布を**図 13** に示す．およそ正規分布に従うことが分かる．

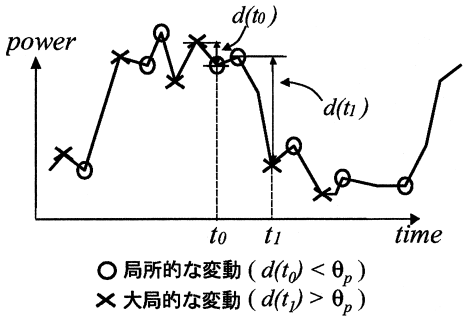

図 12 パワーの局所的変動とその定義
Fig. 12 Local perturbation of power and its definition.

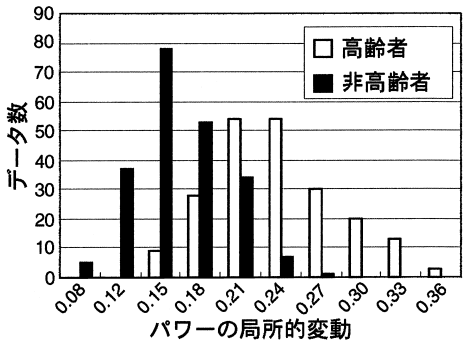

図 13 パワーの局所的変動の分布
Fig. 13 Distribution of local perturbation of power.

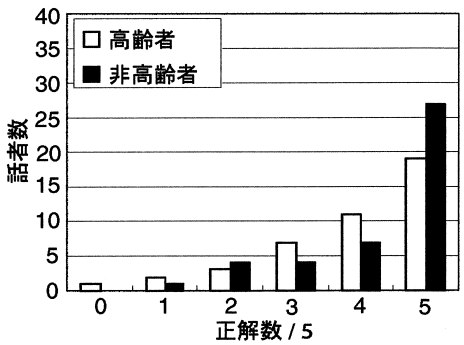

図 14 パワー局所変動のみを用いた主観的高齢者同定実験結果
Fig. 14 Results of identifying subjectively-defined elderly speaker only with local perturbation of power.

パワーの局所的変動の分布のみを用いた同定実験

パワー局所変動を正規分布を仮定してモデル化し，パワー局所変動による尤度差のみを用いた同定実験を行った．実験形態としては，4.2.1 項同様，cross-validation に基づいて評価した．結果を**図 14** に示す．話者同定率は 87.2%（主観的高齢者 37/43，主観的非高齢者 38/43），発話同定率は 81.9%であった．このことか

ら，パワーの局所的変動による主観的高齢者同定は，話者認識技術に基づく方法より性能には劣るものの，話速のみに基づく同定率よりも高く，その有効性がうかがえる．

4.4 主観的年代によるピッチ情報の違い

4.4.1 ピッチ，Δ ピッチの平均と分散

ピッチ，Δ ピッチに関して平均，分散の分布の様子を調べたが，予備実験の結果，年代による顕著な違いは見られなかった．

4.4.2 ピッチの局所変動

前節の実験により，主観的高齢者にはパワーパターンに局所変動が多いことが示されたが，これが声帯の衰えによるものと考えると，ピッチ情報においても同じような現象が観測されると予想される．そこで，パワーと同じようにピッチの局所変動についても分析した．しかしながら予備実験の結果，ピッチの局所変動に関しては，年代による差はあまり顕著ではないことが示された．ただし，一般に高齢者音声のピッチ抽出精度は落ちる傾向があるため，今後，高齢者音声に対しても安定してかつ高精度に動作するピッチ抽出技術が開発されれば，主観的高齢者，主観的非高齢者間にピッチ情報の有意差が検出される可能性も残されている．また，4.3.3 項で扱ったパワーは，25 msec のフレーム（およそ数ピッチ分）に対する二乗平均値として算出している．その結果，ピッチの揺らぎがパワーの揺らぎとして観測される可能性もある．

5. バタチャリヤ距離による特徴パラメータ間の比較

前章までに，話者認識技術を利用して求まるスペクトル尤度差と，いくつかの音響パラメータについて，主観的高齢者/主観的非高齢者間での分布の様子を示した．これらのパラメータはその分布に正規分布を仮定してモデル化しても，各特徴量の単位，レンジが異なるため，異なるパラメータ間の分離の精度（分布間の距離）について単純な比較はできない．また，話者認識技術では特徴パラメータはベクトルとして定義されるが，多次元空間における分布間距離と,（たとえば話速のような）スカラ特徴量の分布間距離は，次元が異なるため定量的比較が困難である．このような比較を可能にするためには，単位/レンジの差，次元数の差を正規化する必要がある．これらの正規化は，次式で定義される多次元正規分布間の距離尺度であるバタチャリヤ距離[21]を用いることで可能になる．

表 3 バタチャリヤ距離による各パラメータの比較
Table 3 Comparison of Bhattacharyya distances between parameters.

パラメータセット	バタチャリヤ距離
(A) 話者認識技術に基づく尤度差	0.995
(B) ΔMFCC のピーク数から求まる話速	0.330
(C) モーラ数から求まる話速	0.222
(D) Δ パワーの分散	0.274
(E) パワーの局所的変動	0.462
(F) ピッチの局所的変動	0.220
(A)+(B)	1.221
(A)+(E)	1.241
(A)+(B)+(E)	1.355

$$
\begin{aligned}
BD(P_a, P_b) &= -\log_e \int_{-\infty}^{\infty} \sqrt{P_a(x)P_b(x)}dx \quad (1)\\
&= \frac{1}{8} u_{ab} \left\{ \frac{\sum_a + \sum_b}{2} \right\}^{-1} u_{ab}^t \\
&\quad + \frac{1}{2} \log_e \left(\frac{|(\sum_a + \sum_b)/2|}{|\sum_a|^{\frac{1}{2}} |\sum_b|^{\frac{1}{2}}} \right)
\end{aligned}
$$

ここで u_a，u_b はカテゴリ a，b の平均ベクトルであり $u_{ab} = u_a - u_b$ である．また $\sum_a$，$\sum_b$ は各々分散共分散行列である．**表 3** に，話者グループ間のバタチャリヤ距離が比較的大きく算出されたパラメータおよびパラメータセットについて，その距離値を示す．表に掲載されていないパラメータ（たとえばピッチの平均，分散など）はいずれも距離値が 0.1 未満である．表より，話者認識ベースの手法に加え，バタチャリヤ距離値が比較的大きく算出されているパラメータである話速とパワーの局所的変動の情報を考慮することで（すなわち (A)+(B)+(E) を用いることで）同定率が上がることが予想される．次章ではこれらのパラメータを統合することを考える．ここでは，線形判別分析とニューラルネットワークを用いた統合を検討する．

6. 話速，局所的パワー変動を考慮した主観的高齢者同定

6.1 線形判別分析

話者認識技術で求められるスペクトル尤度差と，話速，パワーの局所的変動の 3 つをパラメータとし，線形判別分析を行った．なお，話速は ΔMFCC のノルムに基づく話速を採用している．学習，評価話者の割当ては一連の実験に準拠させた．結果を**図 15** に示す．話者同定率は 94.2%（主観的高齢者 40/43，主観的非高齢者 41/43），発話同定率は 92.8%となり，話速，局所的パワー変動といった韻律的特徴を考慮することの有効性が示された．

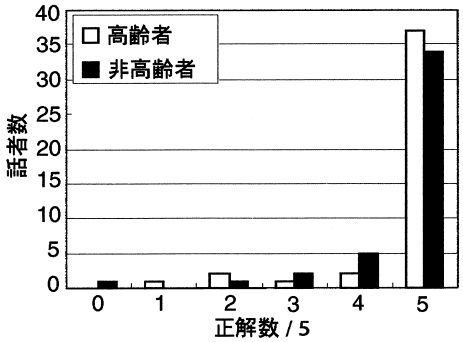

図 15 線形判別分析による主観的高齢者同定実験結果
Fig. 15 Subjectively-defined elderly speaker identification by integrating the parameters with linear discriminant analysis.

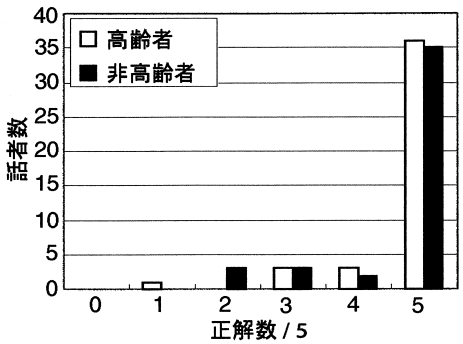

図 16 ニューラルネットワークによる主観的高齢者同定実験結果
Fig. 16 Subjectively-defined speaker identification by integrating the parameters with neural network.

6.2 ニューラルネットワーク

線形判別分析は，各パラメータにかかる重みが明確に示されるという利点がある一方，パラメータ空間を平面で区切るため，モデルとしては単純である．逆にニューラルネットワークでは，基本的には入力パラメータの重み付き線形和により各層（ノード）の値は計算されるが，ネットワークの構造を考慮することでより複雑なモデルが構築可能であり，認識率の向上が期待できる．そこで線形判別分析と同じパラメータを使用し，ニューラルネットワークで高齢者同定を試みた．ここでは 3 層フィードフォワード型ニューラルネットワークを使用した．結果を**図 16** に示す．話者同定率は 95.3%（主観的高齢者 42/43，主観的非高齢者 40/43），発話同定率は 93.0%となり，ここでも話速，局所的パワー変動といった韻律的特徴を考慮することの有効性が示された．しかし今回の実験では，ニューラルネットワークの線形判別分析に対する優位性は観測されているものの，その差異は非常に小さなものとなった．これについては，ニューラルネットワーク構造が最適化されていないことが原因と考えており，今後の課題の 1 つとなっている．なお，一連の実験における同定率の比較を容易にするため，**表 4** に各種手法の話者同定率と発話同定率を示す．話者認識技術のみを用いた場合の性能と比較して，韻律的特徴を考慮することの効果が明確に示されている．

表 4 各種手法による話者同定率と発話同定率
Table 4 Comparison of speaker-based identification rates and utterance-based identification rates among the investigated methods.

パラメータ統合	話者同定率 [%]	発話同定率 [%]
ベースライン	90.7	90.9
線形判別分析	94.2	92.8
ニューラルネット	95.3	93.0

7. 判別得点（discriminant score）を用いた主観的年代推定

前章では判別得点の正負に基づいて，主観的高齢者/主観的非高齢者の同定を行ったが，この得点は各話者グループらしさを表現する定量的尺度と解釈することもできる．そこで本章では，聴取実験で調査した各話者に対する主観的年代と判別得点との対応を分析した．聴取実験における主観的年代推定タスクでは，12 名の被験者に，各発声者を 5 段階の年代のいずれかに分類させた（すなわち各発声者に対して 12 通りの推定結果がある）．この結果を用いて，各発声者の主観的年代を以下のように決定した．2 章の聴取実験では，20〜30 代（20〜39 歳），40〜50 代（40〜59 歳），60 代（60〜69 歳），70 代（70〜79 歳），80 代以上（80〜歳）という 5 つの主観的年代に分類させている．そこで，各年代の「中心」年齢を各年代の中心値として定義し，各発声者に対して各被験者が判定した「中心」年齢（すなわち年代）の平均値をとり，その平均値が属する年代（上記 5 年代のいずれか）を，その発声者の主観的年代と定義した．

図 17 に，上記手順で決定された各発声者の主観的年代と，判別得点の平均，標準偏差との関係を示す．横軸上，5 つの点においてデータが分布しているが，これは各年代の「中心」年齢である．なお,「80 代以上」の年代は便宜的に 85 としてプロットしている．図より，判別得点が主観的年齢と高い相関にあり，主観的年代推定におけるパラメータとしての有効性が示唆される．

8. まとめと今後の課題

本論文ではまず，JNAS および S-JNAS データベー

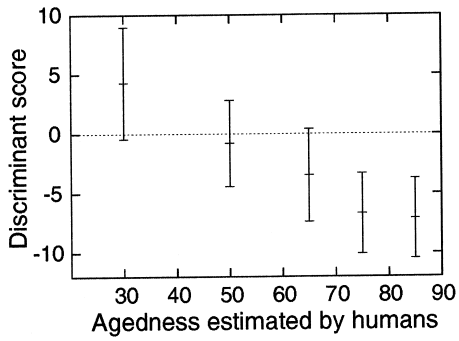

図 17 判別得点と主観的年代推定との関係
Fig. 17 Relation between discriminant scores and subjectively estimated speakers' agedness.

ス中の話者に対して，聴取実験を通して主観的高齢者を特定した．次に，特定された話者の同定を目的として，話者認識技術を用いて同定実験を行い約 91%の精度を得た．精度改善を目的として，アンケートより示された種々の音響的要因を調査することにより，発話速度とパワーの局所変動が主観的高齢者同定に有効に寄与することが示された．これらの特徴量を話者認識技術をベースとする方法論に組み込むことで，約 95%まで精度を向上することができた．また，同定時に得られるスコアと，聴取実験の結果求まった各話者の主観的年代との関係を分析したところ比較的高い相関関係が見られ，本論文で検討したパラメータによる主観的年代の推定可能性を示すことができた．

今後の課題としては以下のことがあげられる．本研究は当初，主観的高齢者の同定を目的として計画されたため，7 章で検討した主観的年代の推定に関しては，聴取実験時の年代設定方法など，実験計画に不備な点があることは否めない．人間が発話者の年代を推定するプロセスを模擬し，その精度を高めることを目的とした場合，年齢という軸を年代というカテゴリに分類する方法，すなわち非線型スカラ量子化の最適解を求めるところから議論する必要があると思われる．そのために使用するデータとしては，今回利用したデータベース以外にも子供音声データを含む幅広い年代のデータを用意する必要がある．また，今回は話者認識技術を直接利用したため，1 状態 HMM（GMM）に基づいて検討したが，特徴量を含め，状態数やトポロジ，韻律的特徴の導入方式に関する検討は十分とはいい難い．話者グループのモデルをマルチテンプレート化する方法論も当然考えられる．さらには，主観的年代のみならず，実年齢の推定というタスクも興味深い．

話者グループのマルチテンプレート化をおし進めた場合，究極的には話者単位でモデルを持つ形態となる．この場合，各モデルが実年齢ラベル，主観的年齢ラベルを持ちうることを考えると，入力話者の実年齢，主観年齢の推定方法として，各モデル尤度を用いた実年齢/主観的年齢ラベル値の期待値を用いる方法が可能となる．現在，このような話者モデルに基づく実年齢，主観年齢推定について実験的な検討をすでに進めており，その推定精度については別の機会に報告する予定である．

主観的高齢者の同定というタスクに対して，提案手法によって約 95%の精度を得ることができた．しかし，誤同定話者の音声を聴取する限りにおいては，その理由が把握できていないのも事実である．本論文で検討したタスクは二者択一タスクであることを考えると，その性能をさらに向上させることが可能であると考えている．モデル形態，パラメータのチューニングなど上記した事項以外にも，本研究では検討できていない有効パラメータのさらなる探求も必要であろう．

本論文で提案した技術の実用的価値を考えた場合，対話システムにどのようにして組み込むかが焦点となる．高齢者と同定できた場合，どのようなサービスが可能となるのか，そもそも高齢者と非高齢者とでサービスを区別することがユーザに望まれるサービスなのか，といった立場から議論する必要性も感じている．いかなる利用者に対しても利用可能なサービスを提供する枠組み（ユニバーサルデザイン[22]）という考えも浸透しつつある．システム側の挙動が利用者の特性によって変化したことを感知されないような制御方式というのも検討課題の 1 つであろう．

参 考 文 献

1) Gustafson, J., Lundeberg, M. and Liljencrants, J.: Experiences from the development of August — A multi-modal spoken dialogue system, *Proc. ESCA workshop interactive dialogue in multi-modal systems*, pp.61–64 (1999).
2) 堂坂浩二, 安田宣仁, 宮崎 昇, 中野幹生, 相川清明：音声対話システム「飛遊夢（ひゅーむ）」, 電子情報通信学会総合大会講演論文集, Vol.1, pp.506–507 (2001).
3) 杉山 聡, 赤植淳一, 小暮 潔：歩行者ナビゲーションにおける情報伝達の利用者適応の分析, 情報処理学会音声言語情報処理研究会資料, SLP36-8, pp.49–54 (2001).
4) Mori, K. and Nakagawa, S.: Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition, *Proc. ICASSP'2001*, Vol.1, pp.413–416 (2001).
5) Martin, A.F. and Przybocki, M.A.: Speaker recognition in a multi-speaker environment,

Proc. EUROSPEECH'2001, Vol.2, pp.787–780 (2001).
6) Hirshberg, J., Litman, D. and Swerts, M.: Prosodic cues to recognition errors, *Proc. ASRU'99*, pp.349–352 (1999).
7) 甲斐敏彦，石丸明子，伊藤敏彦，小西達裕，伊東幸宏：目的地設定対話タスクにおける訂正発話の特徴分析と検出への応用，音響学会秋季講演論文集，2-1-8, pp.63–64 (2001).
8) Li, Y. and Zhao, Y.: Recognizing emotions in speech using short-term and long-term features, *Proc. ICSLP'98*, Vol.6, pp.2255–2258 (1998).
9) 河合　剛，広瀬啓吉：複数言語の音韻論モデルを用いた母語認識手法，日本音響学会秋季講演論文集，1-1-27, pp.53–54 (1999).
10) http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/index.htm
11) http://db.ciair.coe.nagoya-u.ac.jp
12) インタラクティブ音声認識技術，KDD 研究所 (2000).
13) 関口真理子，峯松信明，広瀬啓吉：話者認識技術を利用した高齢話者の同定，電子情報通信学会音声研究会資料，SP2001-77, pp.31–38 (2001).
14) http://www.milab.is.tsukuba.ac.jp/jnas/
15) 粕谷英樹，鈴木久喜，城戸健一：年齢，性別による日本語 5 母音のピッチ周波数とホルマント周波数の変化，日本音響学会誌，Vol.24, pp.355–364 (1968).
16) 粕谷厚生，菊池耕一，増渕伸一：高齢者音声の分析と合成，電子情報通信学会音声研究会資料，SP87-134, pp.33–39 (1987).
17) 小沼知浩，桑野裕康，木村達也，渡辺泰助：高齢者音声の解析と認識評価，日本音響学会秋季講演論文集，2-Q-1, pp.117–118 (1997).
18) Baba, A., Yoshizawa, S., Yamada, M., Lee, A. and Shikano, K.: Elderly acoustic model for large vocabulary continuous speech recognition, *Proc. EUROSPEECH'2001*, Vol.3, pp.1657–1660 (2001).
19) Reynolds, D. and Heck, L.P.: Speaker verification: from reserch to reality, *ICASSP'2001*, tutorial session (2001).
20) Markov, K.P.: Text-independet speaker recognition based on frame level liklihood transformations, Doctor thesis, Toyohashi University of Technology (1999).
21) 中川聖一：パターン情報処理，丸善株式会社 (1999).
22) 高橋　亘：音声的ユニバーサルインターフェイスと日本語解析，電子情報通信学会福祉情報工学研究会資料，WIT2000-10 (2000).

(平成 13 年 11 月 16 日受付)
(平成 14 年 4 月 16 日採録)

峯松　信明（正会員）

昭和 41 年生．平成 7 年東京大学大学院工学系研究科電子工学専攻博士課程修了．博士（工学）．同年豊橋技術科学大学情報工学系助手．平成 12 年東京大学大学院工学系研究科助教授，平成 13 年同大学院情報理工学系研究科助教授．平成 14 年瑞国 KTH 客員研究員．音声認識，音声分析，音声応用，音声知覚，および音声合成の研究に従事．電子情報通信学会，日本音響学会，日本音声学会，人工知能学会各会員．

広瀬　啓吉（正会員）

昭和 24 年生．昭和 52 年東京大学大学院博士課程修了．工学博士．同年東京大学工学部電気工学科講師．昭和 62 年米国 MIT 客員研究員．平成 6 年東京大学工学部電子工学科教授．平成 8 年同大学大学院工学系研究科電子情報工学専攻教授．平成 11 年より同大学院新領域創成科学研究科基盤情報学専攻教授．音声言語情報処理分野一般についての教育研究開発，特に韻律に着目した研究に従事．IEEE，米国音響学会，ISCA，日本音響学会，電子情報通信学会，人工知能学会，言語処理学会等各会員．

関口真理子

昭和 52 年生．平成 12 年東京大学工学部電子情報工学科卒業．平成 14 年東京大学大学院新領域創成科学研究科修了．現在，NTT-IT 勤務．音響情報を用いた話者の年代推定技術の開発，およびその応用に関する研究に従事．日本音響学会会員．