社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2004-164 (2005-03)

聴取による話者識別における音韻間の格差と音響的対応

網野 加苗† 菅原 勉† 荒井 隆行‡

†上智大学外国語学部 〒102-8554 東京都千代田区紀尾井町 7-1
‡上智大学理工学部 〒102-8554 東京都千代田区紀尾井町 7-1
E-mail: †amino-k@sophia.ac.jp, sugawara@sophia.ac.jp ‡arai@sophia.ac.jp

あらまし 人間が音声のみを聞いて話者を識別する場合，聴取する音の種類によって識別のしやすさは異なる．先行研究によると，有声自鳴音，特に鼻音が含まれる刺激音を用いた場合に話者の識別がしやすいことが分かっている．本研究では 10 名の話者の音声について聴取による話者識別実験を行い，鼻音が有効であることを再度確認した上で刺激音の音響分析によってその理由を説明することを試みた．その結果，各刺激音における話者間のケプストラム距離の大きさと話者の識別のしやすさには関連があることが明らかになった．

キーワード 鼻音，話者識別，個人性，ケプストラム距離

The correspondences between the differences among the phones in human speaker identification and their acoustic properties

Kanae AMINO† Tsutomu SUGAWARA† and Takayuki ARAI‡

†Faculty of Foreign Studies, Sophia University 7-1 Kioi-Cho, Chiyoda-Ku, Tokyo, 102-8554 Japan
‡Faculty of Science and Technology, Sophia University 7-1 Kioi-Cho, Chiyoda-Ku, Tokyo, 102-8554 Japan
E-mail: †amino-k@sophia.ac.jp, sugawara@sophia.ac.jp ‡arai@sophia.ac.jp

Abstract The higher success of human speaker identification depends on the speech contents given to the listeners. In previous studies, the voiced sonorants, especially the nasals, were reported to be available for the identification of the speaker. In this study, perception tests were carried out in order to investigate the effectiveness of some Japanese phones excerpted from the sentences. Again, it was shown that the percent correct was the highest when the stimuli containing the nasal sounds were used. We have then attempted to acoustically explain these differences in stimuli, and we found out that there are correspondences between these differences among the phones and the cepstral distances of the stimuli.

Keyword Nasals, Speaker Identification, Individuality, Cepstral Distance

1. 音声の個人性

人間の音声には発話の意味内容である言語学的情報以外にも，話者個人に関する情報（以下，個人性）も含まれている [1]. 個人性は，聴覚的に特定の話者の音色を与えるものであり，準永久的に全ての種類の音声に含まれる音質であると定義されている [2, 3]. 音声の個人性には，話者の生理学的特徴に起因するものと後天的特徴に起因するものがある．前者は声帯の長さや厚さ，声道の長さや体積によるもので，後者は話し方の癖や話者の所属する言語社会の音体系による特性を意味している．また，広義の個人性には発話のモダリティや話者の感情，異常構音なども含まれると言える．

音声の個人性は，言語学的情報に対して非言語学的情報（extra-linguistic information）と呼ばれ，しばしば言語研究の対象からは除外される [4] が，日常的な言語活動においては円滑なコミュニケーションを行うために活用されている情報であり，その特性を明らかにする必要がある．実用的観点からすると，自動音声認識において個人性は排除すべきものであり，自動話者認識においては抽出すべきものである．これらの実用の場においては，音声の個人性とその音響的対応は大変興味深く，現在までにも様々な研究が試みられてきた [1, 5-7]. また，音声学や社会福祉工学などの学問分野においても，理論の説明や社会福祉の目的のために個人性の研究が行われている [8, 9].

個人性の音響的対応を調べる際のアプローチの一つとして，様々な種類の音に関して人間の聴取による話者識別実験を行い，どのような種類の音を聞いたときに最も話者の識別がしやすいかを調べるという方法がある [10]. この方法を用いた研究によると，聴取による話者識別では母音や鼻音などの有声自鳴音が有効

であるという傾向が報告されている [11-14]. さらに, 統計的手法を用いた研究 [15] や, 日本語以外の言語における聴取実験 [16, 17] からも同様の結果が得られている.

これらの音を用いた際に話者の個人性を知覚しやすいということは, 音響的に考えてもこれらの音に何らかの個人性に関する情報が含まれているということを意味している [18]. 本研究では, 先行する一連の話者識別聴取実験 [12, 13] の結果及び今回新たに行った聴取実験の結果によって, 刺激音間の話者識別のしやすさに見られる格差を再度確認し, その音響的対応について検討する.

2. 聴取による話者識別

2.1. 背景

前述の通り, 聴取による話者識別実験では有声自鳴音を用いた場合に識別正答率が高いという結果が出ている. 日本語の音に関して, 音の種類による識別正答率の傾向を調べるため, 本研究に先行して 2 つの聴取実験（以下, 実験 1, 実験 2）が行われた.

実験 1 [12] では, 3 名の話者グループの様々な発話に関して 14 名の知り合いの聴取者による識別実験が行われた. 刺激音に用いた音声は, 表 1 に示すように, 単独で発話された日本語の 5 母音や子音と後続母音 /a/ から成る単音節などである. 実験 2 [13] では, キャリア文から抜粋された単音節によって, 18 名の聴取者が 3 名の話者グループ 2 つ, 計 6 名の話者の識別実験が行われた. 実験 2 で録音された実験文の一覧を表 2 に示す. 聴取実験で実際に用いられた刺激音は, これらの実験文の前半部である, 子音と後続母音 /a/ の組み合わせ語から抜粋された最後の CV 音節である.

また, 実験 1 及び実験 2 の聴取結果に見られた傾向は, 表 3 に示す通りである. その他の先行研究と同様, 有声自鳴音が話者の識別に有効であるという結果が得られたが, なかでも子音部に鼻音が含まれる CV 音節が特に有効であるという傾向が確認された.

表 1. 実験 1 における実験文

Part 1	単独で発話された 5 母音
Part 2	単独で発話された子音 /s:/ /z:/ /m:/ /n:/（伸ばして発話したものの 2 秒間分）
Part 3	単独で発話された CV 音節（V は /a/, C の種類は /p/ /b/ /t/ /d/ /k/ /g/ /m/ /n/ /nj/ /r/ /ɸ/ /s/ /(d)z/ /ʃ/ /ç/ /h/ /j/ /w/）
Part 4	Part 3 の CV 音節を 5 回繰り返したもの（例: ぱぱぱぱぱ /papapapapa/） アクセントは平板式（低高高高高）

表 2. 実験 2 における実験文

あぱぱぱ /apapapa/	党を支持します /to o siji simasu/
あばばば /abababa/	
あたたた /atatata/	
あだだだ /adadada/	
あかかか /akakaka/	
あがああ /agagaga/	
あちゃちゃちゃ /achachacha/	
あららら /ararara/	
あままま /amamama/	
あなな /ananana/	
あにゃにゃにゃ /anjanjanja/	
あささ /asasasa/	
あざざざ /azazaza/	
あしゃしゃしゃ /aʃaʃaʃa/	
あははは /ahahaha/	

表 3. 実験 1 及び実験 2 の聴取結果の傾向

実験 1	聴取する刺激音の持続時間が長いほど識別正答率が高い
	日本語の 5 母音では, 前舌・中舌母音は識別正答率が比較的高い（/a/, /i/, /e/ の平均 83.33%, /o/, /u/ の平均 77.38%）
	単独で発話された CV 音節では, 子音が鼻音の時に識別正答率が高い
	CV 音節の子音が有声音の場合の方が, 無声音の場合よりも識別正答率が高い
実験 2	話者または話者グループによって, 識別に有効な音の種類は異なる
	一方の話者グループのみにおいて, CV 音節の C が鼻音である時に識別率が高いという傾向が見られた
	鼻音が有効とされた話者グループでは, 調音方法別には鼻音-摩擦音-破裂音の順に正答率が高かった

2.2. 先行研究における問題

前項の 2 つの実験では, いくつかの課題が残されていた. まず, 実験の識別課題が簡単すぎたために聴取結果が正規分布をなさず, データに天井効果が生じていた. また, 聴取による話者認識実験においてはしばしば見られる問題であるが, 聴取者間で話者との知り合いの度合いや日常生活における話者との接触度に差があったため, 均一の聴取結果が得られなかった. これらの問題を解決するためには, 話者数を増やすこと

で実験課題を難しくする，事前に被験者候補に対して面接やアンケートを行うことで，話者・聴取者間の親密度をコントロールするなどの対策が必要である．

本研究では，話者数を10名に増やして発話の録音を行い，10名の話者全員をよく知っている5名の聴取者の協力を得て，再度話者識別の聴取実験を行った．

3. 実験

3.1. 手続き

実験条件は表 4 に示す通りである．被験者は，話者・聴取者ともに日本語母語話者であるが，出生地や海外在住経験は特に考慮しなかった．全員聴力に異常はなかった．話者10名と聴取者5名は全員同じ学生寮で生活している大学生であり，話者全員と各聴取者の間に日常的な接触があることを事前に確認した．

本実験は基本的には実験 2 と同じ要領で行われたが，話者数を10名に増やしたため，聴取実験に要する時間や聴取者の疲労・集中力の低下も考慮に入れて，刺激音の種類を前回の15種類から9種類に制限した．ここでは，先行して行われた 2 つの実験で話者の識別に有効であるという傾向が見られた鼻音 3 種類と，日本語の子音体系のうち最も音素数の多い歯茎及びその付近の位置で調音される 6 種類を選んだ．各話者の各刺激音に対する標本数は 5 個であり，すなわち聴取実験では1種類の刺激音が，250回分（話者 10 名×5 標本×聴取者 5 名）の評価を受けたことになる．

音声資料は全て，防音室で 1 名ずつ，48kHz, 16bit で DAT にディジタル録音された．刺激音は全部で 450 個（話者 10 名×9 種類×5 標本）あったが，恣意的に 150 個ずつ 3 つのパートに分けた上で，各パート内でランダムに並べ替えて編集した．聴取実験も同様に防音室で 1 名ずつ行い，DAT からヘッドフォンを通じて両耳聴取させた．音量は各聴取者の快適なレベルとし，休憩も含めた実験時間は大体 40 分であった．

表 4. 実験条件

話者	男性 10 名，録音時の平均年齢 22.6 歳
聴取者	男性 5 名，実験時の平均年齢 23.0 歳
実験文	実験 2（表 2）に準ずる
刺激音	実験文から抜粋された CV 音節（V は /a/, C は /t/ /d/ /m/ /n/ /nj/ /r/ /s/ /z/ /j/）

表 5. 聴取結果: 刺激音別の識別正答率

刺激音	正答数（/250）	正答率（%）
/na/	215	86.0
/nja/	214	85.6
/ma/, /za/	202	80.8
/sa/	197	78.8
/ja/	196	78.4
/da/	195	78.0
/ra/	186	74.4
/ta/	184	73.6

3.2. 実験結果

聴取実験の結果を表 5 にまとめる．表 5 から，本実験の結果も実験 1 及び実験 2 と同様，鼻音含む刺激音を用いた場合に最も話者の識別がしやすかったことが分かる．鼻音より下位の口音に関しては目立った傾向は見られないが，/ta/-/da/, /sa/-/za/ の対においてそれぞれ有声音が無声音より上位にあることが分かる．全体的に調音方法を基準として見てみると，鼻音，摩擦音，接近音，破裂音（弾き音 /r/ を含む）の順になっていることが分かる．この順序は，実験 2 において鼻音で正答率が高かった話者グループにおける順序とも一致している．

次にこれらの結果を統計的に分析した．まず，刺激音の種類という要因に関して一元配置の分散分析を行ったが，全体的には刺激音間で有意な差は見られなかった（$p = 0.12$）．さらに多重比較検定を行ったところ，/na/-/ra/（$p = 0.015$），/na/-/ta/（$p = 0.009$），/nja/-/ra/（$p = 0.019$），/nja/-/ta/（$p = 0.012$）の 4 対において有意な差が見られた．

また，数種類の弁別素性に関して対をなす 2 群間の差の検定（スチューデントの t 検定）を行ったところ，鼻音-口音の 2 群では有意差が見られた（$p = 0.0044$）が，閉鎖音-摩擦音（$p = 0.25$），有声音-無声音（$p = 0.36$），自鳴音-阻害音（$p = 0.15$）の 2 群間では平均の差は有意ではなかった．

4. 音響分析

4.1. 目的と方法

先行研究における傾向も含め，本研究において再度確認された聴取結果における刺激音間の格差，すなわち，鼻音を含む刺激音を用いた場合の話者識別率が口音を用いた場合よりも（一部の種類に関しては有意に）よいという結果を音響的根拠に基づいて説明することを目的とし，次に音響分析を行った．本実験の刺激音は全て同じ「子音＋後続母音 /a/」という単音節構造から成っており，そのため刺激音間の聴取実験結果に

おける格差は子音部あるいは子音部から母音部への遷移区間にあるものと考えてよい．そこで本研究では，刺激音の子音部のスペクトルに関して，ケプストラム距離を計算することによって子音間及び話者間の比較を試みた．

まず刺激音を 48kHz から 16kHz にダウンサンプルし，母音への遷移前の子音部から 30ms のフレームを抜粋した．子音区間の定義及び抜粋基準の設定があいまいな刺激音 /nja/, /ja/, /ra/ については，今回は分析対象には入れなかった．残りの 6 種類の子音 /t/, /d/, /m/, /n/, /s/, /z/ については，調音方法別に子音部抜粋の様子を一部図 1 に示す．閉鎖音 /t/, /d/ は母音部への遷移が始まる前の解放を含む閉鎖区間，鼻音 /m/, /n/ は同じく母音部への遷移が始まる前の鼻音マーマーの区間，摩擦音 /s/, /z/ は母音への移行が始まる前の摩擦定常部の区間から，それぞれフレームの抜粋を行うことを基準とした．

抜粋後，子音の種類ごとに話者 10 名全員の全 5 標本分の発話について，総当り方式（50×50）でケプストラム距離を求めた．さらに，その距離の話者内・話者間での平均値を求め，話者内距離に対する話者間距離の比の値を計算した．ここでのケプストラム距離の値には，比較するスペクトルに対応するケプストラム係数の差を 2 乗し，さらにその平方根を取ったものを用いている．また，ここでは係数 C_0 も含まれている．

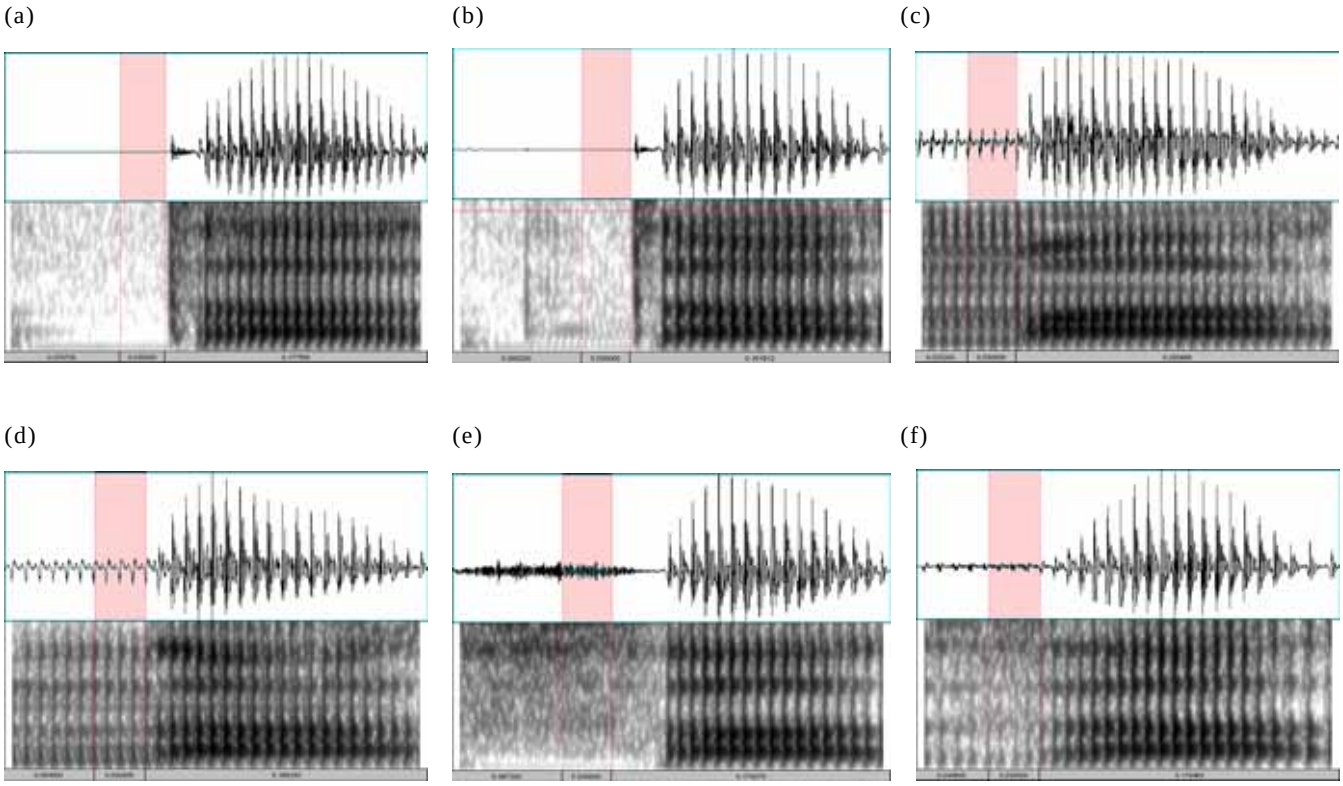

図 1. 聴取された刺激音からの子音抜粋の様子: (a) /t/, (b) /d/, (c) /m/, (d) /n/, (e) /s/, (f) /z/ の抜粋の一例

表 6. 話者内距離に対する話者間距離の比の値

	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6	Sp7	Sp8	Sp9	Sp10	平均
/m/	2.47	1.75	2.51	2.57	3.16	1.58	2.16	1.92	1.55	0.96	2.06
/n/	1.71	1.59	1.83	2.30	2.94	2.07	2.37	1.66	1.01	2.13	1.96
/z/	1.58	1.27	2.37	1.76	1.40	2.35	1.20	2.08	1.64	1.62	1.73
/s/	1.77	1.55	2.45	1.49	1.43	2.02	1.21	1.79	1.64	1.69	1.70
/t/	1.83	1.37	1.60	1.90	1.67	2.21	1.24	1.81	1.51	1.59	1.67
/d/	1.15	1.36	1.33	1.38	2.16	2.38	1.17	1.33	1.10	1.45	1.48

4.2. 分析結果

ケプストラム距離の話者内平均値に対する話者間平均値の比の値を表 6 に，また表 6 をグラフ化したものを図 2 に，子音の種類別に 10 名全員の話者内距離及び話者間距離の平均値をグラフ化したものを図 3 に示す．これら 3 つの図表から話者内距離に対する話者間距離の比の値は，鼻音 /m/, /n/ で大きく，閉鎖音 /t/, /d/ で小さくなっていることが分かる．比の値が大きいほど，一話者内で安定しており，複数話者間で大きく異なる，つまりその音がある話者の特徴をより強く表していることを意味している．

これらの比の値に関して，一元配置の分散分析を行ったが，子音の種類による有意な差は，多重比較検定においても見られなかった．次に，聴取実験の結果の統計分析と同様に，様々な対立する弁別素性に関して 2 群間の差の検定（スチューデントの t 検定）を行ったところ，鼻音-口音の 2 群間においてのみ有意差が見られた（$p = 0.0038$）．

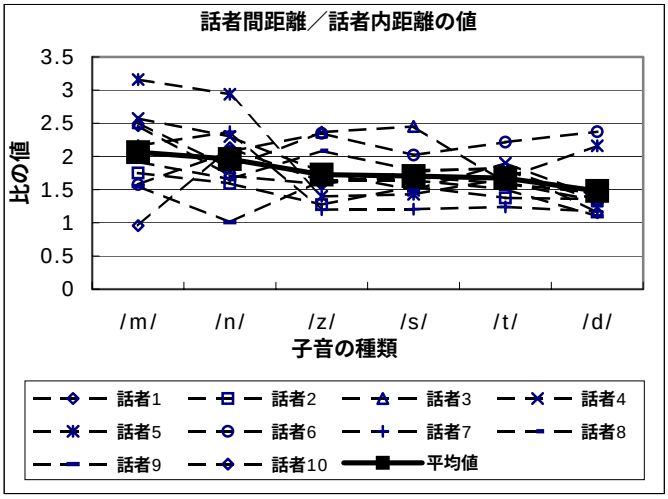

図 2. 子音別 話者内距離に対する話者間距離の比の値

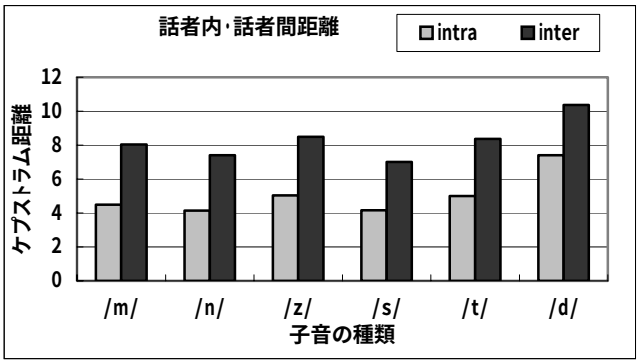

図 3. 子音別 話者内・話者間距離の平均値

5. 考察

聴取による話者識別実験における音の種類間の格差と音響分析結果の対応を見るために，識別正答率とケプストラム距離の話者内･話者間平均値の比の値を，いずれも大きい順に表 7 に示す．

両者を比較してみると，多少の順位の入れ替わりはあるものの，調音方法別に見て鼻音-閉鎖音-摩擦音という順になっているという点において共通している．統計的には両者に有意な相関は見られなかったが，傾向としては，聴取による話者識別と音声の音響的距離の間には関連があると言える．

表 7. 話者識別正答率とケプストラム距離の話者内・話者間比の対応

順位	識別正答率	ケプストラム距離
1	/n/	/m/
2	/m/	/n/
3	/z/	/z/
4	/s/	/s/
5	/d/	/t/
6	/t/	/d/

6. おわりに

音声の聴取によって話者を識別する際，聴取する音の種類によってその有効性に差があるが，なぜそのような差が生じるのかを音響的に説明することを目的として，聴取実験と音響分析を行った．聴取実験の結果，先行研究と同様，鼻音が含まれる刺激音を聞いた場合に最も正確に話者の識別ができることが明らかになった．また，一部の先行研究と同様，調音方法別には鼻音-摩擦音-閉鎖音の順に話者の識別がしやすいという傾向が見られた．音響分析では，聴取実験の刺激音のうち 6 種類を選び，その子音部に関してケプストラム距離を求めたところ，聴取実験における識別正答率と同様，鼻音のスペクトルで最も話者間の違いが大きく，次いで摩擦音-閉鎖音の順になることが判明した．つまり，音声に含まれる個人性に関して，聴覚印象と音響特性の間に関連が見られた．

鼻音が個人性を反映しやすい理由としては，鼻音の調音には口音よりも多くの生理学的特徴が関与することが考えられる．鼻腔・鼻咽腔・副鼻腔などは個人によって形状が異なるため [19]，その共鳴特性は個人性を反映しやすいものと考えられる．また，これらの共鳴腔は話者が意図的に形状を変化させることが困難であるため，その特性は比較的外的要因の影響を受けに

くく，話者個人の特徴の現れとしての信頼性も高いと考えられる．今後はより一層，生理学的観点からの鼻音の研究が必要となるだろう．

また，本研究の音響分析はスペクトルの詳細には関与しておらず，例えばどの周波数帯域が特に話者間で大きな距離を示すのかなどの疑問点には答えられていない．今後はより詳細な観察を行う必要がある．聴取実験の結果に関しても，詳しく考察すると，話者の識別に有効であるとされた鼻音の種類間にも格差が認められる．つまり，本研究における聴取実験及び先行して行った 2 つの聴取実験において，共通して刺激音となっていた /ma/ と /na/ の 2 種類を比較すると，常に /na/ が /ma/ よりも話者の識別に有効であった．今回の音響分析では，ケプストラム距離は /n/ よりも /m/ で話者の個人性を表すという結果が出ており，音響的な説明に至っていない．この格差についてもさらなる調査が必要である．さらに，本実験の刺激音では後続母音を /a/ に限定していたが，その他の母音が後続した場合やその他の音韻構造，例えば撥音 /N/ を含む構造などに関しても検証を行っていきたい．

文　献

[1] 新美康永，音声認識，坂井利之（編），共立出版，東京，1979.

[2] J. Laver, The Phonetic Description of Voice Quality, Cambridge University Press, Cambridge, 1980.

[3] D. Abercrombie, Elements of General Phonetics, Edinburgh University Press, Edinburgh, 1967.

[4] 城生佰太郎，佐藤和之，斎藤純男，福盛貴弘，松崎寛，コンピュータ音声学，城生佰太郎（編）日本語教育学シリーズ第 3 巻，おうふう，東京，2001.

[5] P. D. Bricker, and S. Pruzansky, “Speaker Recognition,” in Contemporary Issues in Experimental Phonetics, ed. N. J. Lass, pp. 295-325, Academic Press, New York, 1976.

[6] 古井貞熙，ディジタル音声処理，東海大学出版会，東京，1985.

[7] T. Kitamura, and M. Akagi, “Speaker Individualities in Speech Spectral Envelopes,” J. Acoust. Soc. Jpn. (E), Vol.16, no.5, pp.283-289, 1995.

[8] P. Ladefoged, A Course in Phonetics, 4th ed., Heinle and Heinle, Boston, 2001.

[9] 飯田朱美，“感情表現と個人性を反映した音声合成手法とコミュニケーション支援への応用,” Sophia Linguistica, Vol.50, pp.179-195, 2003.

[10] D. O’Shaughnessy, Speech Communications ― Human and Machine ―, 2nd ed., Addison-Wesley Publishing Company, New York, 2000.

[11] 西尾寅弥，“声で人を当てられるか,” 言語生活，Vol.158, pp.36-42, November, 1964.

[12] 網野加苗，“話者識別における日本語音韻の特性 ― 聴取による話者識別実験に基づいて ―,” 2003 年度上智大学言語学会予稿集，18 号，pp.32-43, July, 2003.

[13] 網野加苗，“聴覚による話者識別における日本語音韻の特性,” 信学技報，Vol.104, no.149, SP2004-37, pp.49-54, June, 2004.

[14] 松井知子，古井貞熙，I. Pollack，“連続音声中の音節による個人性知覚,” 日本音響学会講演論文集平成 5 年秋，no.2-9-9, pp.379-380, 1993.

[15] 中川聖一，坂井利之，“日本語スペクトルの特徴分析および音声認識・話者認識への考察,” 日本音響学会誌，35-3, pp.111-117, 1979.

[16] M. R. Sambur, “Selection of Acoustic Features for Speaker Identification,” IEEE Trans. ASSP, Vol.23, no.2, pp.176-182, 1975.

[17] G. S. Ramishvili, “Automatic Voice Recognition,” Engineering Cybernetics, Vol.5, pp.84-90, 1966.

[18] G. Fant, “Acoustic Theory of Speech Production,” The Hague, Mouton, 1960.

[19] J. Dang, and K. Honda, “Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation,” J. Acoust. Soc. Am., Vol.100, no.5, pp.3374-3383, Nov., 1996.