単音節による雑音下の個人性知覚

網野 加苗　荒井 隆行

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: {am ino-k, arai}@sophia.ac.jp

あらまし 聴取による話者識別では, 呈示する音声によって正答率が変化することが分かっており, 著者らによる今までの実験では刺激音に鼻音が含まれる場合に正答率が高くなることが分かった. 法科学においては, 音声資料に雑音が含まれる場合も多く, 雑音付加刺激を用いた識別実験も必要であると考えられる. 本稿では ABX 法を用いた実験により, (1) 雑音を付加する際の SN 比, (2) 刺激音中の音韻, (3) 音声ラインアップ中のターゲット話者の有無の 3 つの要因が聴取による話者識別の正答率に与える影響を調べた. 実験結果から, SN 比の向上に伴い正答率も上昇すること, ターゲット話者の存在が正答率に大きく影響を与えること, 鼻音を含む単音節は雑音下でのターゲット話者不在ラインアップにおいて正答率が特に低下することが示された.

キーワード 話者識別, 雑音, 個人性, 感度, 鼻音

Perceptual Speaker Recognition Using Noise-added Monosyllables

Kanae AMINO and Takayuki ARAI

Faculty of Science and Technology, Sophia University 7-1 Kioi, Chiyoda-ku, Tokyo, 102-8554 Japan

E-mail: {am ino-k, arai}@jsophia.ac.jp

Abstract In perceptual speaker identification, it is known that the identification accuracy depends on the stimuli presented to the listeners. In our previous experiments, the stimuli containing a nasal sound have consistently yielded better performances than the stimuli containing only oral sounds. Taking into account the fact that speaker recognition in forensics often has to do with the speech materials recorded in noise, it is necessary to investigate how the presence of noise affects the speaker identification performances. In this present study, we conducted a speaker identification experiment in an ABX paradigm where noise-added monosyllables were used as the stimuli, and investigated the effects of (1) signal-to-noise ratios, (2) the phonological contents of the stimuli, and (3) presence of the target speakers in the voice line-ups. The results showed that higher SNR and the presence of the target speaker both increased the identification accuracy. Performances on the nasals were more deteriorated than on other syllables with the line-ups where the target speaker is absent.

Keyword Speaker Identification, Noise, Individuality, Sensitivity, SNR, Nasals

1. 背景

聴取による話者識別を調べることは, 法科学や認知科学にとって有益である. 音声コミュニケーションにおいて, ヒトは音声に含まれる情報から発話内容や話者の感情, 話者のアイデンティティを知覚・認識している. 音声信号に含まれるこれらの情報は, 聞き手の知覚において, 複雑に影響し合っている [1].

一例として, 音韻情報と個人性の相互作用が挙げられる. 聞き手が発話内容（音韻情報）を理解する際, 話者との親密度が影響することが報告されている [1, 2]. すなわち, 知り合いの話者による発話は未知話者の発話と比較して明瞭度が高くなると言われている. 逆に, 話者の個人性を知覚する際, どのような発話内容の音声を呈示するかによって, 話者識別正答率が異なることも知られている [3].

著者らによる今までの実験においては, 刺激音中に鼻音が含まれる場合に, 正答率が高いことが示されてきた. 異なる話者セットを用いた場合においても, また話者との親密度や刺激音の音節構造などを変化させた場合においても, 鼻音を含む音声を呈示した場合に話者識別の正答率が高かった [4 -6]. 音響的にも, スペクトルの話者間の違いが口音よりも鼻音で大きく, ケプストラム距離と聴取における話者間混同の間に相関関係があることも示された [4, 7].

しかし, 上記はいずれも雑音のないクリーン環境下での実験結果であり, 日常の音声知覚環境には常に雑音や残響が存在することや[8], 法科学の現場では検証資料が劣悪な録音状況にある可能性が否定できないことを鑑みると [9, 10] , 雑音や残響下の知覚についても調べる必要がある.

橋本ら [1 1] は雑音重畳音声を用いた話者識別実験を行い, 複数の音響パラメータの個人性知覚への寄与

度を調査している．実験の結果，個人性知覚に重要な音響パラメータが，クリーン環境と雑音下では異なることを示した．いずれの条件下でも，基本周波数とスペクトルは個人性知覚に大きく影響するものの，雑音下では，クリーン音声と比較して，特に基本周波数の影響が大きくなることが明らかになった．また，話者との親密度に関して，未知話者の場合は既知話者と比較してスペクトルの影響が少ないことも分かった．

しかし，先行研究では刺激音として文を用いており，文中の出現音素の影響については考察されていない．さらに，A BX 法によって聴取実験を行っているが，X が A と B のどちらの話者でもない試行については調査されていない．実際の法科学の現場では，ごく限られた長さの音声しか利用できない場合も多く，またターゲット話者 (実際の話者) が必ずしも候補として呈示される音声 (音声ラインアップ) の中に存在するとは限らない．

そこで本研究では (1) 雑音付加の際の SN 比，(2) 刺激音に含まれる音素の種類，そして先行研究において影響が指摘されている (3) 音声ラインアップ中のターゲット話者の有無が，雑音下での聴取による話者識別にどのような影響を与えるかを調べるために実験を行った．

2. 実験

2.1. 仮説

上述の検討項目(1)～(3)について，それぞれ以下のように 3 つの仮説を立てた．

【仮説 i 】
SN 比が向上するにつれ，話者識別正答率も上昇する．

雑音下での話者識別の知覚実験で，単音節を刺激とした研究は，管見の限りでは見当たらない．雑音下における音韻知覚の結果からのアナロジーとして，正答率と SN 比は正の相関を成すと考えられる [1 4].

【仮説 ii 】
クリーン音声と同様，雑音を付加した刺激音を用いても，鼻音を含む場合に話者識別正答率が高い．

雑音の種類によっては，スペクトルの形状に起因して，特定の音韻で正答率が低下することも考えられる．今回の実験では，その要因を排除するため，まずは比較的平坦なスペクトルをもつ白色雑音を使用して実験を行う．

【仮説 iii 】
呈示する音声ラインアップにターゲット話者が存在する場合と存在しない場合では，前者において話者識別正答率が高い．

話者の存在の有無による正答率の違いは，先行研究においても示されており [12, 13] , ターゲット話者が存在するラインアップにおいて正答率が高いという見解で一致している．よって本研究においても同様の結果が得られるものと思われる．

2.2. パイロットテストと問題点

網野＆荒井 [15] では, T able 1 に示す条件によって雑音下の話者識別実験を行った．その結果，聴取者は雑音を付加した単音節による識別であっても，十分に話者を判別できることが分かった．結果には天井効果が見られ，刺激音間の十分な比較ができなかった．そこで，課題の難易度およびデザインを再検討し, T able 1 の最右列に示した条件で再度実験を行った．

3. 実験

3.1. 音声資料

電子協日本語共通音声データ [16] から，男性話者 5 名分の単音節発話を使用した．話者選択のおおまかな基準として，関東方言話者であることと雑音の少ない環境下で録音を行った話者であることを考慮した．各話者につき，コーパスに収められた 4 回の単音節発話のうち明確に調音された 2 回分を使用した．

実験では 5 名の話者のうち 3 名をターゲット話者 (target sp eakers, 以下 TS), 2 名を非ターゲット話者 (foil sp eakers, 以下 FS) とした. 5 名のうちどの話者を TS とし，どの話者を FS とするかを決めるにあたって，まず初めに上述の基準に当てはまる 10 名の男性話者をコーパスから選定し，実験者が聴取して特徴的な声質の話者がいないかを確認した．次に各話者の単音節発話における基本周波数 (F0) を全トークンについて測定し，平均 F0 の値が最も近い 5 名を選んだ. Table 2 に各話者の平均 F0 を示す．以上の選定手続きは，未知話者識別では，特徴的な声質の話者や話者間の F0 の違いが大きな手がかりになるとの報告に基づいており [1 1, 17] , それらの大きな要因のみに実験結果が左右されることを避けるためである．

Table 1 Experimental conditions of the pilot test [15] and this present study

Parameters	Pilot test [15]	This study
Speech materials	JEIDA speech corpus [16]	
Stimuli	Noise-added monosyllables	
Experimental task	Naming	Matching
No. of speakers	4 male speakers	3-TS and 2-FS
Familiarity	Previously unknown speakers	
Noise types	White, babble	White
SNR [dB]	0, 5, 10, ∞	-15, -10, -5, ∞

Table 2 Mean F0s of the five speakers' monosyllables

Type	Speaker ID	Mean F0 [Hz]
TS	#1	126.01
	#2	119.57
	#3	130.26
FS	#4	118.87
	#5	130.60

Table 3 Stimulus voice line-ups

Presented positions	A	B	X
SP stimuli	Two out of the three speakers: TS1/TS2/TS3		Either A or B
SA stimuli			FS1 or FS2

最後に平均 F0 が高い順に並べて 2 番目と 4 番目にくる話者を FS, 残りの 3 名を TS とした. これは Holli en [9] による「音声ラインアップでは, 特徴が似通った話者を FS と TS に分散させるとよい」という提案に基づいたものである.

単音節は, 母音部が /a/ である日本語の CV 音節とし, 子音は 7 種類 (/d/ /t/ /m/ /n/ /j/ /z/ /s/) を使用した. これらは著者らによる今までの実験に倣って選択されたもので, /m/ を除いて全て冠音 (coron al consonants) である.

3.2. 刺激音の作成

雑音の種類は, 白色雑音のみとした. 本実験では, SN 比の変化によって刺激音間の正答率の差がどのように変化するかを確認することが第一の目的であり, 周波数上で比較的均一なスペクトルを持つ雑音が望ましいと判断したためである. 信号 (単音節) に対する雑音比 (SN 比) が, -15 dB, -10 dB, -5 dB となるように雑音を付加した. 雑音を付加した刺激以外に原音声 (SN 比 = ∞) も呈示した.

このようにして作成した刺激音を, A BX 方式によって呈示した. A と B には 3 名の TS のうちの 2 名を, X には TS または FS から 1 名を選択した. A, B, および X には, 子音と SN 比の種類が等しい 3 つの単音節を一定の間隔 (600 ms) を空けて並べ, 1 セットの音声ラインアップとして呈示した. X が TS である刺激音を話者存在刺激 (以下 SP 刺激), 含まない刺激音を話者不在刺激 (以下 SA 刺激) とする. 刺激音の呈示順と SP・SA 刺激の内容をまとめたものを Table 3 に示す.

SP 刺激は 3 名中 2 名を選ぶ全組合せについて刺激音を作成し, ABX および BAX の全パタンを呈示した. ここで, A または B で呈示する刺激と X で呈示する同話者の刺激には, それぞれ異なるトークンを用いている. SA 刺激では, 実験時間および聴取者への負担を考慮して, 2 名の FS が同頻度で出現するように各条件間でカウンタバランスをとり, 刺激数を半減させた. このようにして, 合計 504 個の刺激音を作成した.

上記より, 本実験は単音節の種類 7 条件と SN 比の 4 条件 (原音声を含む), さらに話者の存在の有無 2 条件の 3 つを独立変数とするデザインで行い, 話者識別の正答率を従属変数とした.

3.3. 手続き

聴取者として実験に参加したのは, 実験前に 5 名の話者の音声を一度も聞いたことがない健聴者 13 名である. 全員日本語母語話者であった.

聴取実験は一名ずつ防音室で行った. 聴取者はPC上に保存された刺激音をオーディオプロセッサー (ONKYO SE-U55GX) からヘッドフォン (SONY MDR-Z700) を通じて両耳に呈示した. それぞれの刺激セットについて, 3 番目に聞こえた音声 (X) の話者が誰だと思ったかを直感で答えるよう教示し, 「1 番目 (A) の音声と同じ話者」「2 番目の音声 (B) と同じ話者」「どちらでもない」のいずれかのボタンをPC画面上でクリックして回答させた. 話者の総数および単音節の種類については事前に教示していないが, 3 つ目の音声 (X) が 1 番目 (A) と 2 番目 (B) のどちらでもない可能性があることは事前に伝えた. 実験開始前には, T S, FS のいずれも含まない別の話者セットの音声を用いて練習を行った.

実験は, 練習も含めて全て Praat [18] の MFC プログラムを使用して行い, 刺激の再生・回答については, 聞き直しや修正はさせなかった. 各刺激音の繰り返しはなく, 総試行数は 504 問で, 聴取者は途中 3 回の休憩を取った.

4. 結果

実験結果は話者識別正答率によって評価した. 全回答を SP・SA 刺激に分け, 単音節および SN 比ごとに正答数を求めた. 正答数は 13 名中の正解した人数とし, 繰り返し試行数によって平均をとり, 最後に平均正答率を計算した.

4.1. *d'*分析

各要因の影響を分析する前に, 単音節ごとの *d'*値を求めた. *D'* (d-p rime) は, 信号検出システムの精度を測るための指標で, 式 (1) によって与えられる [19].

$$d' = z(H) - z(F) \quad (1)$$

ここで H は HIT（信号が検出されるべき区間で正しく検出すること）の割合, F は FALSE ALARM（信号が存在しない区間で検出の判定を出すこと）の割合, z は *z* 標準化を表している.

Table 4 How to classify the responses into the four categories in *d'* calculations

	Response: SP	Response: SA
Stimuli: SP	HIT correctly identified the speaker	MISS rejected the actual speaker
Stimuli: SA	FALSE ALARM answered a wrong speaker	CORRECT REJECTION correctly rejected the foil speaker

*D'*値は, 聴取実験では聴取者の刺激に対する感度を反映している. 全ての区間で信号を検出するようなシステム (あるいは完全にランダムな回答を行った聴取者) では, *d'* は 0 になる.

本実験では, SP 刺激と SA 刺激の回答を Table 4 のように分け, 式 (1) を用いて単音節ごとの平均 *d'* 値を求めた. その結果を Fig. 1 に示す. この図から, 全ての単音節で正の値をとっていることが分かる. これらをランダムな回答をした場合, すなわち *d'* が全て 0 である場合と比較したところ, どの単音節でも有意に値が大きいことが示された. つまり, 本実験で得られた回答はランダムに与えられたものではないことになる.

グラフから, /ja/ /za/で *d'* 値が高く, /na/ /ta/で低いことが分かる. 子音間の差について分散分析を行ったところ, 有意傾向が見られたが ($p = 0.093$), 差が有意なペアはなかった.

4.2. SN 比の影響

SP 刺激・SA 刺激別の各 SN 比の平均正答率を Fig. 2 に示す. 分散分析の結果, SN 比の主効果は有意だった.

Fig. 2 を見ると, SP・SA 刺激で話者識別正答率が 50%を超えるのは, それぞれ SN 比が-10 dB と-5 dB の時であることが分かる. 全体的には, いずれの刺激においても SN 比が向上するにつれて正答率は高くなっており, 【仮説 3】は支持されたと言える.

4.3. 単音節および TS の有無の影響

SP 刺激と SA 刺激に分けた単音節ごとの話者識別正答率を Fig. 3 (a) および Fig. 3 (b) に示す. 前者は SP 刺激, 後者は SA 刺激についての正答率である.

子音による正答率の違いを分散分析で比較したところ, SP 刺激ではどの SN 比についても有意差は見られなかった. SA 刺激では, どの SN 比でも /ja/ -/na/ 間のみに有意な差が見られた ($p < 0.05$). 鼻音を含む /na/ という音節で正答率が低いという結果は, 著者らによる今までの実験の結果に反するものであり [4-6], またこのことから【仮説 1】は SP 刺激・SA 刺激のどちらにおいても支持されなかったことになる.

続いて SP 刺激・SA 刺激間で正答率を比較したところ, SP 刺激で有意に正答率が高いことが分かった ($p < 0.001$). これは【仮説 2】を支持しており, 他の先行研究で見られた傾向とも一致するものである [12, 13].

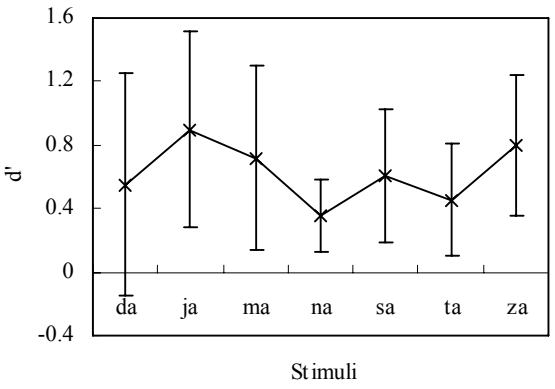

Fig. 1 Listeners' sensitivity: *d'* and its *S.D.* for each syllable

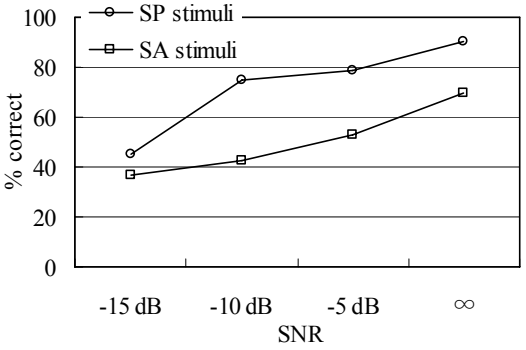

Fig. 2 Identification accuracy as a function of the SNR

5. まとめと考察

本研究では, 雑音を付加した単音節を刺激音として, 聴取による話者識別実験を行い, 3 つの要因が識別正答率にどのような影響を与えるかを調べた.

1 つ目の要因として, 単音節による話者識別における SN 比の影響を調べた. その結果, Fig. 2 に示した通り, ターゲット話者の有無に依らず, SN 比が上昇すると正答率も上昇することが分かった. 雑音下の聴取による話者識別が正答率 50%を超えるためには, SP 刺激では -10 dB, SA 刺激では -5 dB 以上の SN 比が必要であることも明らかになった.

本実験 2 つ目の要因である単音節の影響については, 著者らによる先行研究 [4-6] で聴取による話者識別に有効とされた鼻音を含む音節 (/ma/ /na/) は, 本研究

では同様の傾向を示さなかった. 雑音を含まない原音声の SP 刺激では, 単音節間の差はほとんど見られず, SA 刺激では, 特に /na/ で正答率が低いという結果になった.

考えられる理由の一つとして, 先行研究との実験課題の違いが挙げられる. 今までの実験は, いずれも命名課題であったが, 今回の実験はマッチング課題である. 課題の違いが聴取による話者識別に影響するとの指摘もあり [3], その違いが影響した可能性がある. ただ, Fig. 3 (a) および 3(b) を見ると, SP 刺激の原音声では天井効果が出ていることも考えられる.

ターゲット話者の有無との関連を見てみると, SA 刺激では原音声においても鼻音で正答率が低いことが見受けられるが, SP 刺激ではそのような傾向は出ておらず, ターゲット話者の有無が単音節ごとの正答率に影響している可能性もある.

また, 雑音が単音節間の正答率の差に与える影響についても, 原音声の正答率から雑音を付加した場合の正答率の低下の度合いを比較すると, 特に /na/ で大きく低下していることが分かる. 鼻音は雑音下での弁別が特に困難であることが知られているが [20], 本実験においても雑音を付加した刺激音では, 音韻の弁別ができていなかった可能性がある. SN 比が単音節の知覚に与える影響を調べた研究 [14] でも, 単音節の音韻知覚の正答率が 50%を超えるのは, SN 比が+3 dB 以上の場合だと報告されているため, その可能性は高い. また聴取による話者識別では, 外国語での話者識別が母語による識別と比較して困難であることが知られており [21-22], 発話内容の認識と話者性の判断の間には相互作用があると言われている. それによって, 本実験においても, 特に雑音下での弁別が困難な鼻音で, 話者識別正答率が低下したことが説明できる. 今後, 雑音付加の影響を調べる際は, 各刺激音について単音節 (音韻) を答えさせる課題も行う必要がある.

3 つ目の要因として, 音声ラインアップ中のターゲット話者の存在の影響を調べた結果, 他の研究で得られた結果 [12, 13] と同様, 話者存在刺激において正答率が高いことが分かった. 聴取者は話者を拒絶するよりも受容しようとする傾向にあると言える. この点については, 話者識別時に聴取者がどのようなストラテジーを用いたか, あるいは用いなかったかによっても結果が異なってくる可能性がある [22]. 今後の実験ではフォローアップアンケートの実施も検討したい.

今後の課題としては, 白色雑音以外の雑音や残響が話者識別に与える影響を, 単音節の音韻認識と併せて調べる必要がある. また, 法科学への応用を考慮し, 電話音声や時期差のある発話を用いた実験も行いたいと考えている.

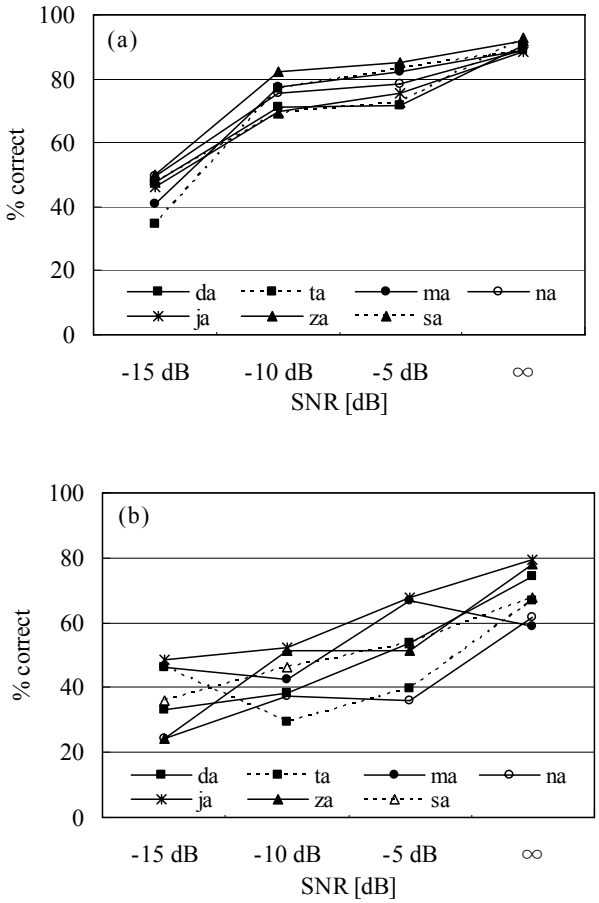

Fig. 3 Identification accuracy as a function of the monosyllables: (a) SP stimuli and (b) SA stimuli.

6. 謝辞

本研究は, 文部科学省私立大学学術研究高度化推進事業 上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の助成を得た.

文 献

[1] L. Nygaard, Perceptual integration of linguistic and nonlinguistic properties of speech, in The Handbook of Speech Perception, eds. D. Pisoni and R. Remez, pp. 390-413, Blackwell Publishing, Oxford, 2005.

[2] S.D. Goldinger, D.B. Pisoni, and J.S. Logan, On the locus of talker variability effects in recall of spoken word lists, J. Exp. Psychol. Learn. Mem. Cogn., vol. 14, pp.152-162, 1991.

[3] P. Bricker and S. Pruzansky, Speaker Recognition, in Experimental Phonetics, ed. N. Lass, pp.295-236, Academic Press, London, 1976.

[4] K. Amino, T. Sugawara, and T. Arai, Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties, Acoust. Sci. Tech., vol. 27, pp.233-235, 2006.

[5] K. Amino and T. Arai, Effects of stimulus contents and speaker familiarity on perceptual speaker

identification, Acoust. Sci. Tech., vol. 28, pp.128-130, 2007.

[6] K. Amino. T. Sugawara, and T. Arai, Effects of the syllable structure on perceptual speaker identification, IEICE Tech. Rep., vol. 105, pp.119-114, 2006.

[7] K. Amino, T. Sugawara, and T. Arai, "Speaker similarities in human perception and their spectral properties," Proc. 9th WESPAC, Seoul, South Korea, June. 2006.

[8] K. Helfer and L. Wilbur, Hearing loss, aging, and speech perception in reverberation and in noise, J. Speech and Hearing Res., vol. 33, pp.149-155, 1990.

[9] H. Hollien, Forensic Voice Identification, Academic Press, San Diego, 2002.

[10] 鎌田敏明, 長内隆, 蒔苗久則, 谷本益巳, "劣悪な雑音環境下における話者照合," 音講論, pp.129-130, Sept., 2005.

[11] 橋本誠, 北川敏, 樋口宜男, "音声の個人性知覚に影響を及ぼす音響的特徴の定量的分析," 音響誌, vol. 54, no. 3, pp.169-178, 1998.

[12] J.H. Kerstholt, N. Jansen, A.G. Van Amelsvoort, and A.P. Broeders, Earwitness: effects of speech duration, retention, internal and acoustic environment, Appl. Cog. Psychol., vol. 18, pp.327-336, 2004.

[13] L.R. van Wallendael, A. Surace, D.H. Parsons, and M. Brown, Earwitness voice recognition: factors affecting accuracy and impact on jurors, Appl. Cog. Psychol., vol. 8, pp.661-677, 1994.

[14] G. Miller, G. Heise, and W. Lichten, The intelligibility of speech as a function of the context of the speech materials, J. Exp. Psychol., vol. 41, pp. 329-335, 1951.

[15] 網野加苗, 荒井隆行, "雑音下における話者識別に発話内容が及ぼす影響について," 音講論, pp. 425-428, Sept., 2007.

[16] JEIDA Japanese common speech data corpus, http://www.sunrisemusic.co.jp/dataBase/fl/voicebase01_fl.html

[17] 北村達也, パーハム・モクタリ, "母音間に共通する個人性情報の知覚的要因," 音講論, pp.493-494, March, 2006.

[18] P. Boersma and D. Weenink, Praat: doing phonetics by computer, ver.4.5.1.4., retrieved from http://www.praat.org/, Computer programme, 2005.

[19] T. Wickens, Elementary Signal Detection Theory, Oxford Univ. Press, Oxford, 2002.

[20] A. Alwan, J. Lo, and Q. Zhu, "Human and machine recognition of nasal consonants in noise," Proc. ICPhS, pp.167-170, San Fransisco, USA, 1999.

[21] J.P. Goggin, C.P. Thompson, C.P. Strube, and L.R. Simental, The role of language familiarity in voice identification, Mem. Cog., vol. 19, pp.448-458, 1991.

[22] A.C. Philippon, J. Cherryman, R. Bull, and A. Vrij, Earwitness identification performance: the effect of language, target, deliberate strategies and in direct measures, Appl. Cog. Psychol., vol. 21, pp.539-550, 2007.