雑音環境下におけるヒトの話者識別の誤り傾向*

☆鈴木良平 (上智大)，網野加苗 (科警研)，荒井隆行 (上智大)

1　はじめに

法科学で扱う録音音声資料には，対象話者以外の音声や空調の動作音など，様々な雑音が混入する場合が多い．音声鑑定においては，聴取検査も行うため，雑音環境下におけるヒトの話者識別の仕組みを調べることは，雑音を含む音声も貴重な証拠となり得る犯罪捜査において，音声の活用拡大や鑑定技術向上に貢献できることから有益である．ここで，証拠として用いた音声の鑑定結果に誤りがあれば，誤認逮捕や犯人の取り逃がしを誘発するため，犯罪捜査における音声鑑定には高い精度が求められる．このため，ヒトの話者認識の仕組みの捜査応用を考えるのであれば，ヒトの話者識別にはどういった誤り傾向があるのかも把握する必要がある．

一方，雑音環境に特化して聴取実験を行った例やヒトの話者識別における他者受容率 (FAR) と本人拒否率 (FRR) を分析した研究は少ない．そこで本研究では，ヒトの話者識別の誤り傾向に着目し，30 名の実験参加者から得た回答結果から，識別正答率，他者受容率 (FAR) 及び本人拒否率 (FRR) を集計し，統計分析を行った．この分析結果を基に，ヒトの話者識別の誤り傾向と法科学への応用について考察した．

2　先行研究

先行研究では，スペクトルと基本周波数 (以下，F0) 情報が個人性知覚に重要で，雑音下では特に F0 の影響が大きくなること [1]，SN 比，F0 と雑音の周波数の関係が音韻知覚の精度に関係していること [2]，話者との親密度や話者が既知どうかによって聴取傾向が変わること [1, 3]，刺激音の音節の種類によって正答率が変化すること [4] 等が報告されている．これらの先行研究から，実験を設計する上で，音声資料として用いる話者間の F0 の差，雑音の種類や SN 比，実験参加者と話者の関係，音声のテキストの選定に留意が必要であると考えられる．

また，FAR と FRR に関して，自動話者認識の先行研究では用いる特徴量により FAR と FRR の大小関係が異なる可能性があること，SN 比の低下に伴い FRR が FAR に比べて増加している傾向が指摘されている [5, 6]．ヒトの話者認識の FAR 及び FRR について調べた先行研究は少ないが，自動話者認識の傾向を踏まえると複雑な傾向が存在している可能性があり，分析を行うことは有益と考えられた．なお，FAR と FRR の計算式は先行研究で使用されていた以下 (1)(2) を使用した [5].

Table 1　F0 [Hz] for the words Phone & Mail

話者 No.	「電話」の F0	「メール」の F0
話者 1	111.12	141.36
話者 2	117.77	125.78
話者 3	109.20	142.69
話者 4	122.36	147.22
話者 5	106.57	122.36
平均	113.40	135.88

$$\mathrm{FAR}[\%] = \frac{\text{Number of accepted imposter claims}}{\text{Total number of imposter accesses}} \times 100 \quad (1)$$

$$\mathrm{FRR}[\%] = \frac{\text{Number of rejected genuine claims}}{\text{Total number of genuine accesses}} \times 100 \quad (2)$$

3　実験

3.1　音声資料

今回の実験の音声資料として，科研費「母語識別システムの開発と非母語話者日本語音声コーパスの構築」による録音音声データを使用した (科研費課題番号 JP24810034)．この音声データには無響室において，日本語母語話者も含めた複数の協力者による発話をマイクロフォン (SONY, ECM-23F5) によって録音した音声が収録されていた．

今回はこのマイクロフォン音声のうち，2 種類の単語「電話」「メール」について，平均 F0 が近い 5 名の男性話者による発話を選定し，実験

* Error tendency of human speaker identification in noisy environments. By SUZUKI, Ryohei (Sophia Univ.), AMINO, Kanae (National Research Institute of Police Science), ARAI, Takayuki (Sophia Univ.).

に使用した．各話者，各単語における平均 F0 は Table 1 の通りであった．「電話」「メール」の 2 種類の単語を選定したのは，どちらも 2 音節 3 モーラ語である他，アクセント型が同じで，音声単語親密度も 5 以上と高い単語であったためである．

3.2 雑音の選定と付加

刺激音に付加する雑音として，スピーチノイズとボイラー室の環境雑音を使用した．

音声の長時間スペクトルに似せたノイズは，一般にスピーチノイズと呼ばれている．このため，スピーチノイズを音声資料に付加することで，背景雑音として人間の声が多数存在している状況に近い刺激音を作り出せると考えられる．今回は日本産業規格の定義 [11] に従い，ホワイトノイズに対し，1,000 Hz から 6,000 Hz までオクターブあたり 12 dB 減衰する特性を加えたものを作成し，スピーチノイズ (以下，SpN) とした．

防犯カメラやドライブレコーダを始めとした各機器で録音した音声には，エアコン等の空調装置の動作音が環境雑音として混入することが多数あると考えられる．そこで，空調装置による雑音に似た音と考えられるボイラー室の環境雑音を実験に使用した．今回は ATR 環境音データベースに収録されていたボイラー室の環境音 [12] のうち，人の出入り等，突発的な雑音を含まない冒頭 5 秒を切り出し，雑音として使用した．これを以下，BoilerN と呼称する．

以上の SpN と BoilerN を音声資料に付加する際，[9] に記載があった計算式 (3) を使用し，音声資料と雑音の RMS 比を基準に SN 比の計算を行うようにした．また，いずれの SN 比においても音声資料そのものを同じ音量で提示するため，雑音付加処理後，雑音を重畳していない音声も含め全刺激音で最大レベルとなるサンプルを基準に正規化を行った．

$$x(t) = s(t) + \frac{S_{RMS}}{10^{SNR/20} \times N_{RMS}} n(t) \quad (3)$$

3.3 手続き

聴取実験は，1 名ずつ防音室で行った．実験に参加したのは，18 歳以上の健聴者 30 名 (男性 13 名，女性 17 名) で，全員日本語母語話者であった．実験参加者には PC 上に予め保存しておいた刺激音を，インタフェース (Roland UA25-EX) からヘッドフォン (AKG K72) を通じて diotic 聴で両耳に提示した．音声の提示，聴取者の回答，回答結果の出力には Praat の MFC プログラム [10] を使用し，Replay や回答の修正は許可しなかった．提示する際のレベルは，本実験前に実施した練習試行の際に聴取者にとって快適なレベルに調整するよう指示し，一度レベルを調整した後は，レベルを変更せずに実験を行った．

Table 2　List of pair quantities used in experiment

条件	ペア数
SNR = ∞ 以外，異なる話者	160
SNR = ∞ 以外，同一話者	80
SNR = ∞，異なる話者	40
SNR = ∞，同一話者	20
合計	300

刺激音は AX 法に倣い，2 つの音声を 1 ペアとして続けて提示した．5 名の音声から重複を許す形で 2 名を選択し，同一話者の音声ペアとなる場合は，同じ話者の異なる発話を使用するようにした．

また，ペア内では同じ単語の発話を用い，A の位置で提示する音声には原則雑音を付加したものを使用し (SNR=∞のときは正規化のみ行った原音声を提示)，X の位置で提示する音声には正規化のみ行った原音声を提示した．これは先行研究で，比較資料間で同一のテキストを用いるテキスト依存型の方が精度を高めやすく [7, 8] 現場で活用されやすいと考えられること，未知資料は雑音環境下で録音されることがあるが，これに対する比較資料は後から良好な環境で任意のテキストで録音できる [9] という音声鑑定の特徴を踏まえたためである．

A の位置で提示する音声に付加する雑音として SpN と BoilerN のどちらかを使用し，SN 比は 3 段階 (SNR = ∞, 0 dB, −10 dB) のいずれかとした．聴取者には，Table 2 の通り，合計 300 ペアをランダムに提示し，音声ペアのうち X の位置で聞こえた音声が A の位置で聞こえた音声と同じ話者のものかどうかを回答させた．回答画面には，「2 つ目音声は 1 つ目の音声と同じ話者ですか？」と表示し，画面上に配置したボタン「同一話者」「異なる話者」のいずれかを選択させた．

3.4 結果・分析

SN 比の低下に伴う識別正答率の変化を Fig. 1，FAR と FRR の変化を Fig. 2 に示す．Fig. 1, 2 を参照すると，以下 3 点の傾向を読み取れる．

傾向 1 Fig. 1 より，SN 比の低下に伴い，雑音や単語の種類に関係無く，識別正答率が低下していること．また，SNR = −10 dB となる雑音下においては正答率が約 50%と二肢強制選択法におけるチャンスレベルまで低下していること．

傾向 2 Fig. 2 より，SN 比の低下に伴い，雑音や単語の種類に関係無く，FAR が FRR に比べて上昇していること．

傾向 3 Fig. 2 より，雑音の種類 (SpN, BoilerN) と単語 (電話, メール) によって SN 比の低下に伴う FAR と FRR の上昇度合いに差があること．

以上 3 点の傾向について以下，それぞれ分析を行った．

【傾向 1 SN 比の低下と識別正答率低下の関係】

SN 比の効果に関して分散分析を行った結果，Fig. 1 に記載した雑音と単語の組み合わせ全 4 条件で有意差 ($p < 0.01$) があった．更に，下位検定として Fig. 1 の 4 条件に対し，SNR = 0 dB, −10 dB の間でそれぞれ対応なしの t 検定を行ったところ，有意差 ($p < 0.05$) があった．

【傾向 2 FAR と FRR の大小関係】

SN 比，雑音の種類，単語の 3 条件を揃えた場合の FAR と FRR 間に有意差が見られるかどうか対応なしの t 検定を行った．SN 比，雑音の種類，単語が{−10 dB, SpN, メール}となる条件の場合は $p = 0.05$ と有意傾向が見られ，それ以外の各条件には有意差 ($p < 0.01$) が見られた．

【傾向 3 雑音の種類と単語が FAR と FRR へ及ぼす影響】

SN 比 (SNR = ∞, 0 dB, −10 dB) ×誤りの種類 (FAR or FRR) ×雑音の種類 (SpN or BoilerN) ×単語 (Phone or Mail) の 4 要因分散分析を行った．分散分析の結果，有意差 ($p < 0.01$) が認められた要因を Table 3 に示す．Table 3 を参照すると，SN 比，誤りの種類，雑音の種類，単語に主効果があったことがわかる．主効果があった要因それぞれに対し，下位検定として t 検定を行ったところ，

(1) SNR = −10 dB となる雑音環境下において，雑音種類間で FRR に有意差 ($p < 0.05$) が認められ，SpN 環境下における FRR が BoilerN 環境下の FRR を有意に上回る一方，FAR には有意差が認められないこと

(2) 雑音環境下において単語間で FAR に有意差 ($p < 0.05$) が認められ，電話の FAR がメールの FAR を有意に上回る一方，FRR や非雑音環境下の FAR には有意差が確認できないこと

の二点が判明した．

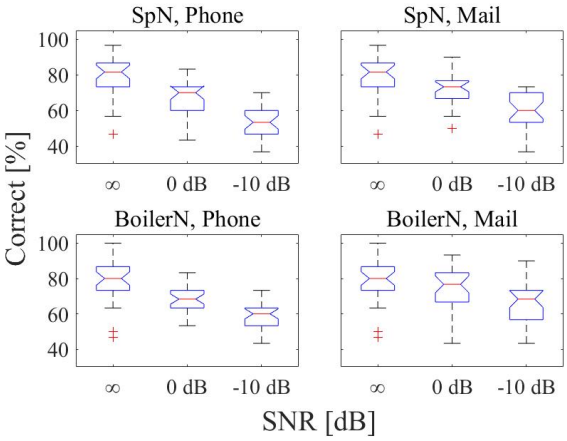

Fig. 1　SNR effect on reduction of accuracy

Table 3　Effect of each factor ($n = 720$)

Factor	F	p
SNR	101	2.00×10^{-16}
FAR or FRR	354	2.00×10^{-16}
Noise	7.30	7.06×10^{-3}
Word	7.11	7.81×10^{-3}
FAR or FRR : Word	10.4	1.29×10^{-3}
SNR : FAR or FRR : Word	5.02	6.79×10^{-3}

注) 表中のコロン : は交互作用を表す．

4　考察

傾向 1 の分析より，SN 比の低下と識別正答率の低下には関連があることを確かめられたと考えられる．先行研究からは，SN 比の低下と母音識別精度の低下に関連があることが指摘されており [2]，話者識別精度についても SN 比の低下と関連があることを確かめられた．

傾向 2 の分析結果を踏まえると，ヒトの話者識別の誤り傾向として，FAR > FRR となる可能性があることが考えられる．また，傾向 3 の分析からは，雑音種類の違いが FRR に影響を与える可能性，単語の違いが FAR に影響を与える可能性があることがわかる．SpN 環境下の FRR が BoilerN 環境下の FRR を上回った理由としては，SpN は語音のマスキングを行うためのノイズ [11] とい

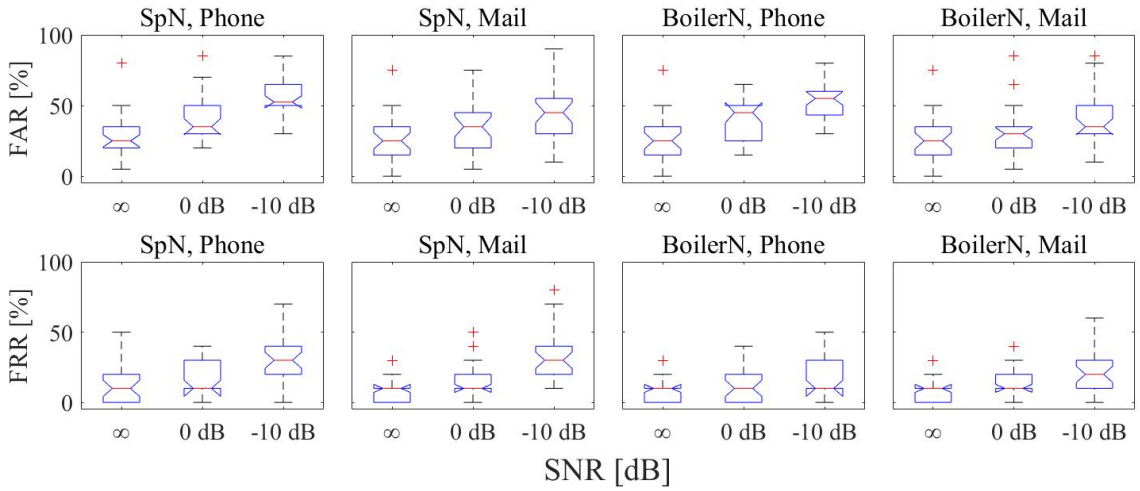

Fig. 2　SNR effect on FAR & FRR

う雑音の特性が，単語の違いによって FRR が変動した理由としては，雑音下で特に識別が困難になるとされる鼻音 [m][n] の存在 [13] や単語中に含まれる鼻音の位置の違いが影響している可能性が考えられた．

FAR と FRR の傾向に関しては，実験で使用したペア数 (Table 2) と数式 (1)(2) の関係から FAR と FRR でサンプル数に差があること (Fig. 2 の箱ひげ図 1 つあたり FAR は $n = 600$，FRR は $n = 300$) や左右に配置していた実験参加者の回答ボタンの位置，指示の仕方によるデフォルト効果 [14] などの影響も受けると言える．引き続き今後の課題としていきたい．

5　結論

今回は SN 比の低下と識別正答率に関係があることがわかった．また，ヒトの話者識別は有意に FAR > FRR となる，雑音種類の違いが FRR に影響を与える，単語の種類が FAR に影響を与える可能性があることなど，FAR と FRR に関して自動話者認識とは異なる複雑な影響がある可能性を確かめられた．

ヒトの話者識別の FAR と FRR の傾向を調べることは，ヒトの話者認識を犯罪捜査の音声鑑定に応用するとどのような誤判定が発生しやすいか把握できるため，今後も本研究を進めていきたい．加えて，身内になりすまして電話を行う特殊詐欺の犯罪抑止策の策定にも FAR と FRR の傾向に関する研究は有益であると考えられるため，更なる応用可能性について検討を進めたい．

謝辞　実験参加者の方々に感謝を申し上げます．なお，本研究は上智大学重点領域研究の一部として助成を得ました．実験に際しては、上智大学倫理委員会の承認を受けています (2021−67) ．

参考文献

[1] 橋本誠 他, 音響学会誌, 54(3), 169–178, 1998.
[2] 石塚, 相川, 情処研報, 2001(16), 153–158, 2001.
[3] Amino and Arai, Acoust. Sci. Tech., 28(2), 128–130, 2007.
[4] 網野, 荒井, 音響学会聴覚研資, 38(6), 579–584, 2008.
[5] Ilyas *et al.*, IEEE SCOReD, 1–5, 2007.
[6] Chen *et al.*, IEEE SP, 694–711, 2021.
[7] 野田, 長内, 信学論 (A), 73(4), 717–724, 1990.
[8] 松井, 古井, 信学論 (D), 79(5), 647–656, 1996.
[9] 鎌田 他, 信学技報, 106(614), 55–60, 2007.
[10] 北原, 田嶋, 音響学会誌, 67(8), 345–350, 2011.
[11] 日本産業規格 JIS T1201-1, 2020.
[12] ATR 環境音データベース I Volume2.
[13] Alwan *et al.*, ICPhS, 167–170, 1999.
[14] Johnson and Goldstein, Science, 302(5649), 1338–1339, 2003.