雑音によるヒトの話者識別への影響 –反応時間に関する検討– *

☆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

ヒトの話者認識の過程や仕組みを調べ, 傾向を把握することは, 聴取検査も行われる音声鑑定の精度向上や新たな話者認識技術の提案に繋げられることから有益であると考えられる. しかし, 音声鑑定に用いられるような法科学的な音声資料には, 雑音が含まれることが多いと考えられるが, 雑音下のヒトの話者認識傾向に着目した研究は少ない.

先行研究からは, ヒトの音韻知覚過程の研究において識別を行う音節の有声無声や調音位置等が反応時間の増加に影響を与えること [1] や雑音と listening effort の関係を評価する研究において, 与えた課題の内容や雑音の特性によって反応時間が変化したこと [2] 等が指摘されている.

以上の先行研究から, 個人性知覚における反応時間も課題や雑音に依存する可能性が考えられた. もし, この仮説が正しければ, 雑音が含まれた音声に対するヒトの話者認識の反応時間を調べることで, 呈示した音声に対する話者認識の難易度の評価や雑音による話者認識への影響を分析できると言える.

そこで, 本研究では, 信号対雑音比 (signal-to-noise ratio; 以下, SN 比) や雑音の周波数特性が聴取に影響する [3], 音節の種類によって話者識別正答率が変化する [4] という先行研究を踏まえ, SN 比を 3 段階 (SNR = ∞, 0 dB, −10 dB) に設定した 2 種類の雑音を用い, 5 名の未知話者の 2 種類の単語音声によって行われた話者識別聴取実験 [5] の結果の再分析を行った. 再分析にあたっては, 平均識別正答率, SN 比や雑音による反応時間の変化, 回答の正誤による反応時間を検討し, 考察を行った.

2 手法

2.1 実験に使用された刺激音

2 種類の単語「電話」「メール」のマイクロフォン (SONY, ECM-23F5) による録音音声のうち, 平均 F0 が近い 5 名の男性話者による発話を収録したものが既存のコーパスから選定された. 発話の前後の空白時間は各 200 ms であった. これらの音声資料に雑音を付加することで実験で呈示する刺激音が作成された [5, 6]. 選定された各音声資料には, ホワイトノイズに特性を加えて作成したスピーチノイズ (以下, SpN) [7] かボイラー室の環境雑音 (以下, BoilerN) [8] のどちらかが SNR = ∞, 0 dB, −10 dB のいずれかで付加された. この際, 空白時間も含めた音声資料全体にわたって雑音を重畳し, 音声資料の長さそのものは変化しないように配慮された. 雑音付加後, 呈示する音声そのものの音量が統一できるよう, 作成した全刺激音の中で最大振幅となるサンプルを基準に正規化が行われた. なお, 音声資料のサンプリング周波数は 44.1 kHz, 量子化ビット数は 16 bit でエンコード形式は PCM であった.

2.2 手続き

聴取実験は, 18 歳以上の健聴者 30 名 (男性 13 名, 女性 17 名) に対し, 1 名ずつ防音室で行われた. 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) を通じて両耳受聴 (diotic) させた. 音声の呈示, 実験参加者の回答, 回答結果の保存, 反応時間の計測には Praat の MFC プログラム [9,10] が使用された. 音声の呈示レベルは, 本実験前に練習試行を実施し, 実験参加者にとって快適なレベルに調整させ, 本実験ではそのレベルを変更せずに実験が進められた.

刺激音は AX 法に倣い, 2 つの音声が 1 ペアとして続けて呈示された. 音声の呈示間隔は 0.5 s であった. 音声ペアの組み合わせにあたっては, 5 名の音声から重複を許す形で 2 名が選択され, 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配慮された. 実験参加者は, Fig. 1 の通り, 合計 300 ペアがランダムに呈示され, 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画面上に配置されたボタンを通じて回答するよう教示された. なお, この際 Replay や回答の修正は許可せず, 回答ボタンを

* Effects of noise on human speaker identification: –Analysis of reaction time, by SUZUKI, Ryohei (Sophia Univ.), AMINO, Kanae (National Research Institute of Police Science) and ARAI, Takayuki (Sophia Univ.).

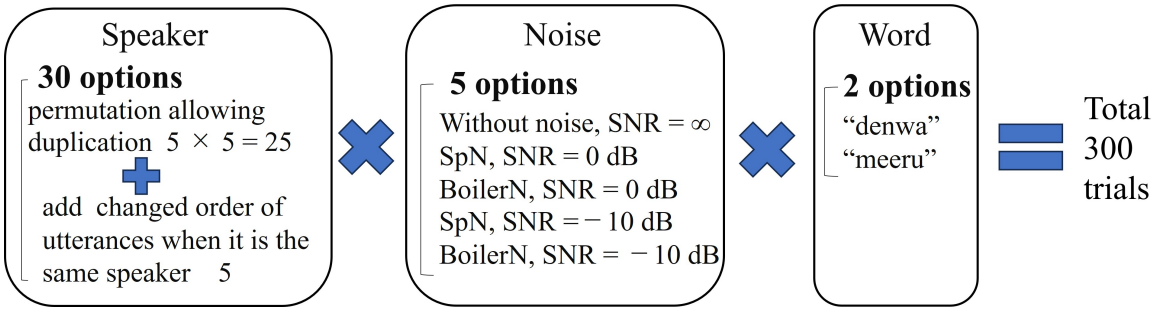

Fig. 1　Number of trials used in the experiment

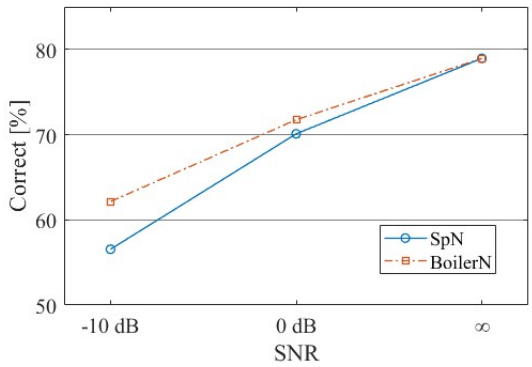

Fig. 2　Noise type and SNR effects on correct rate

押してから 0.5 s 後に自動的に次の音声ペアが呈示される仕組みとされた．

A の位置で呈示する音声に付加する雑音には，SpN か BoilerN のどちらかが使用され，SN 比は 3 段階 (SNR = ∞, 0 dB, −10 dB) のいずれかとされた．また，ペア内では同じ単語の発話が用いられ，A の位置で呈示する音声には原則雑音を付加したものが使用され (SNR = ∞のときは正規化のみ行った原音声を呈示)，X の位置で呈示する音声には正規化のみ行った原音声が呈示された．これらは，比較資料間で同一の発話内容を用いるテキスト依存型の方が精度を高めやすく [11,12]，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発話内容で録音できる [13] という先行研究で指摘されていた音声鑑定の特徴を踏まえたためであった．

3　結果

実験終了後，呈示された刺激音の種類と参加者の回答結果，及び反応時間を集計した．今回は，Praat の仕様に準じ，音声の呈示開始から参加者

Table 1　Average reaction time [s]

Noise	SNR	RT [s]*1	Correct [%]
N/A	∞	3.16	78.9
SpN	0 dB	3.20	70.1
SpN	−10 dB	3.37	56.5
BoilerN	0 dB	3.19	71.7
BoilerN	−10 dB	3.21	62.1

*1: RT stands for reaction time.

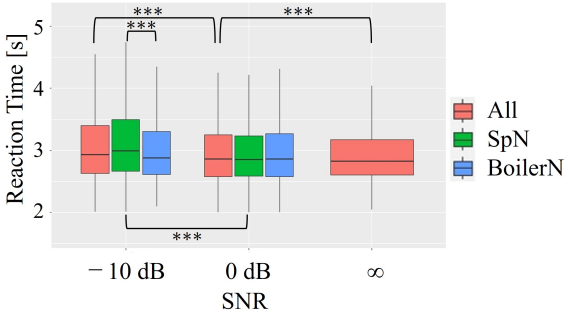

Fig. 3　Noise type and SNR effects on reaction time: *** stands for significant difference ($p < 0.001$)

が回答を行うまでの経過時間を反応時間として分析を行った．なお，当初は話者別の反応時間や単語別の反応時間の集計予定が無かったことや，音声の呈示中も回答が可能であったことから，話者や単語による発話の持続時間の差は考慮に入れなかった．

3.1　平均識別正答率

まず，雑音によるヒトの話者識別傾向の変化を分析するため，平均識別正答率に着目した．雑音の種類と SN 比による平均識別正答率の変化を Fig. 2 に示す．

Fig. 2 を参照すると，雑音の種類を問わず SN 比の低下に伴って平均識別正答率が低下してい

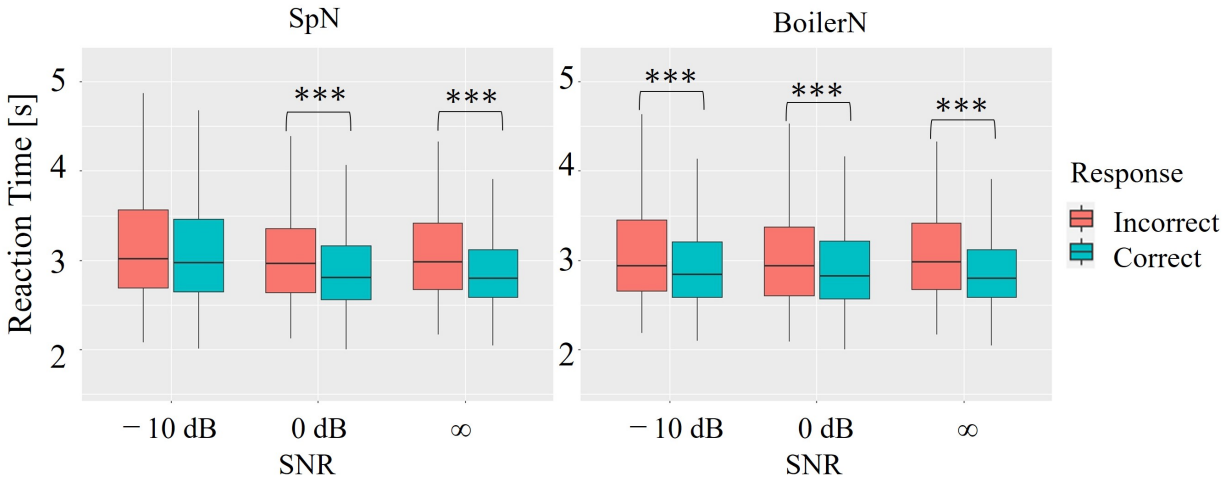

Fig. 4　The effects of correctness on reaction time: *** stands for significant difference ($p < 0.001$)

ること，SpN を付加した刺激音を使用した場合，BoilerN を付加した刺激音を用いた場合に比べ SN 比の低下に伴う平均識別正答率の低下度合いが大きいことが読み取れる．更に，雑音種類と各 SNR に対する反応時間と平均識別正答率を Table 1 に示す．Table 1 の反応時間と平均識別正答率に関して相関分析を行ったところ．相関係数として $r = -0.86$，$p = 0.062$ が得られた．

3.2　雑音の種類，SN 比と反応時間の関係

次に，雑音の種類，SN 比と反応時間の関係について統計分析を行った．雑音の種類，SN 比による反応時間の変化を Fig. 3 に示す．なお，Fig. 3 中 All とは SpN と BoilerN を合算して集計した条件である．Table 1 及び Fig. 3 からは，SN 比が低下するにつれ，反応時間が増加する傾向を読み取れる．また，SNR = −10 dB では，BoilerN よりも SpN を付加した刺激の方が反応時間が長い傾向を確認できる．

雑音の種類，SN 比が反応時間に及ぼす影響について，分散分析を行ったところ，雑音の種類，SN 比，雑音の種類と SN 比の交互作用に主効果 ($p < 0.001$) が見られた．更に下位検定として同一 SN 比の雑音の種類間 (SNR = −10 dB のときの SpN と BoilerN の間，SNR = 0 dB のときの SpN と BoilerN の間)，同一雑音条件の異なる SN 比間 (SpN の SNR = −10 dB と 0 dB の間，BoilerN の SNR = −10 dB と 0 dB の間，All の SNR = −10 dB と 0 dB の間，All の SNR = 0 dB と ∞ の間) で t 検定を行ったところ，以下の 4 条件で有意差 ($p < 0.001$) が見られた．

- All における SNR = −10 dB の反応時間と SNR = 0 dB の反応時間
- All における SNR = 0 dB の反応時間と SNR = ∞ の反応時間
- SNR = −10 dB で SpN の反応時間と BoilerN の反応時間
- 雑音が SpN で SN 比が −10 dB であったときの反応時間と SpN で 0 dB のときの反応時間

3.3　反応時間と正誤の関係

反応時間と正誤の関係について分析するため，参加者の回答結果の正誤を分けた箱ひげ図のプロット，及び回答結果が正解であったときと誤りであったときの反応時間の差の t 検定を行った．この結果を Fig. 4 に示す．Fig. 4 からは，回答結果が誤りであったとき，回答結果が正解であった場合に比べ反応時間が長い傾向があること，そして多くの場合で有意差が見られることがわかる．一方で，SpN で SNR = −10 dB であったときは，有意差が見られず，他の条件と異なり，反応時間に差が無い可能性が高いと言える．

4　考察

まず，SN 比が低下すると識別正答率は低下し，反応時間は増加する傾向があると考えられる．先行研究では，SN 比と識別正答率には関連があること [4–6] や SN 比と音声認識の反応時間の関連 [2] が指摘されている．今回，3.1 節の相関分析から平均識別正答率が低いほど反応時間が長くなるという相関や 3.2 節から SN 比が低下した条件では

反応時間が増加する傾向が確認されており，SN 比の低下が識別正答率の低下，反応時間の増加に関連していると考えられる．

次に，雑音の種類によって識別正答率は変化し，反応時間は増減する可能性があると言える．先行研究 [2] では雑音の特性によって音韻知覚の反応時間が変化したことが指摘されている一方で，音声認識の反応時間に関して雑音の種類による影響は少ないこと [14] を示唆する研究もあり，反応時間に関する傾向は複雑である可能性がある．話者認識の反応時間に関して検討を行った本研究では，SNR = −10 dB，0 dB のいずれでも SpN を付加した音声を呈示したときの平均識別正答率は BoilerN を使用した際より低い点や SNR = −10 dB のときの雑音種類間の反応時間に関して，SpN を付加した刺激音の反応時間が BoilerN 条件と比べて有意に長いことが確認された．これらから，話者認識の反応時間には雑音種類による影響がある可能性が考えられる．反応時間に差が出た要因としては，SpN 条件と BoilerN 条件の間で識別難易度や listening effort に差があったことが挙げられる．

また，SNR = 0 dB のときと比べ SNR = −10 dB のときの雑音種類間の平均識別正答率により大きな差が生じている点や，SNR = −10 dB のときの雑音種類間の反応時間に関して，SpN を付加した刺激音の反応時間が BoilerN 条件と比べて有意に長いことが確認された．これらから，SN 比や付加する雑音の種類によって話者識別の難易度や listening effort が異なり，それらによって雑音下における話者識別の反応時間が変化した可能性があると言える．

最後に，回答の正誤と反応時間の関係について，Fig. 4 より，回答結果が誤りであるとき，反応時間は長くなる傾向にあると言える．これは，即座に判定がしにくい識別難易度の高い刺激であった場合に，同一話者か異なる話者かを判定する判断に迷いが生じた結果，反応時間が増加し，判定精度も低下したためと考えられる．一方で，SpN で SNR = −10 dB の際には，他で見られた有意差が見らなかった．Fig. 2 で示したように，識別難易度が高い SpN のような雑音を付加した雑音条件の際は正誤による反応時間の差が生じにくいとも考えられる．

5 結論

今回は，話者識別実験から得られた反応時間を分析し，ヒトの話者認識の傾向について考察した．この結果，① SN 比が低下すると識別正答率は低下し，反応時間が増加する傾向にあること，②雑音の種類によって識別正答率は変化し，反応時間も変化すること，③回答結果が誤りである際は反応時間が長くなる傾向にあることが判明した．ヒトの話者認識についても，上記の通り，反応時間と関連していることが明らかになったため，今後も研究を継続したい．

謝辞 本研究の音声資料として使用した音声データは，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 (2010)」によるものです (科研費課題番号 JP24810034)．また，本研究の一部は，2023 年 9 月の日本音響学会音声コミュニケーション研究会及び日本音響学会秋季研究発表会での発表内容に基づくものです．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「人を対象とする研究」に関する倫理委員会の承認を受けています (2021 − 67) .

参考文献

[1] P. Stevenson, J. Phon., 1(4), 347-367, 1973.
[2] R. Houben *et al.*, Int. J. Audiol., 52(11), 753-761, 2013.
[3] K. Ishiduka and K. Aikawa, IPSJ SIG technical reports, 2001(16), 153-158, 2001.
[4] K. Amino and T. Arai, Tech. Rep. Physiol. Psychol. Acoust. Acoust. Soc. Jpn., 38(6), 579-584, 2008.
[5] 鈴木 他, 音講論 (秋), 1017-1020, 2023.
[6] 鈴木 他, 音響学会音声コミュニケーション研資, 3(4), 37-42, 2023.
[7] 日本産業規格 JIS, T1201-1, 2020.
[8] ATR-Promotions, ATR ambient noise sound database Ⅱ, 2005.
[9] P. Boersma, Glot. Int., 5(9), 341-345, 2001.
[10] 北原, 田嶋, 音響学会誌, 8(67), 345-350, 2011.
[11] 野田, 長内, 信学論 (A), 73(4), 717-724, 1990.
[12] 松井, 古井, 信学論 (D), 79(5), 647-656, 1996.
[13] T. Kamada *et al.*, IEICE technical report, 106(614), 55-60, 2007.
[14] C. Pals *et al.*, J. Acoust. Soc. Am., 138(3), EL187-EL192, 2015.