個人性知覚における音声の類似性について*

◎網野加苗，荒井隆行（上智大）

1　はじめに

法工学・法科学において音声の個人性を扱う場合，聴取による話者識別特性を調べておくことは有益である．人間による音声の個人性知覚には，言語情報の知覚との相互作用が指摘されており[1]，実際，個人性知覚には刺激の言語情報の種類による差があることが分かっている[2]．

母音に多くの個人性が含まれることは多くの研究によって指摘されているが[2-3]，子音別には，鼻音を含む刺激音を提示した際に話者識別正答率が高いということが著者らによる先の研究で示された[4]．また，それに対応する音響特性として，鼻音では話者間のケプストラム距離が大きいことが示された[5, 6]．

しかしそこでは話者の識別正答率のみに着目しており，話者間の混同，つまり聞き手がある話者を別の話者と似ていると感じる際に何を聞いているのかについては言及していない．本研究では，個人性知覚における話者間の混同と音響特徴量の関連について調べた．

2　音響特徴量

2.1 分析方法

本研究で用いたパラメータは 30 次のケプストラム距離である．分析条件を Table 1 に示す．分析フレームの抜粋は波形とスペクトログラムの目視によって行い，遷移区間の判断の際には後続母音 /a/ の F2 遷移が完了する点を基準とした．フレーム抜粋の例を Fig.1 に示す．

フレーム C は子音部，フレーム C-tr. は後続母音への遷移を含む子音区間である．母音部についてはフレーム V-tr. (子音からの遷移区間を含む母音区間) とフレーム V (母音定常部) の 2 種類を設けた．ただし，/ta/，/da/ のフレーム C に関しては，口腔閉鎖による無音区間であるため，分析対象からは除外した．

全話者の総当り方式で，話者内，話者間のケプストラム距離を 6 種類の単音節の種類別に求め，それらを話者内距離に対する話者間距離の比の形でまとめた．比の値が大きい，つまり話者内で安定し，話者間で違いが大きいほど，話者性を反映しているといえる．

2.2 結果

フレーム及び子音別の話者内対話者間距離比を Table 2 に示す．また，そのグラフを Fig.2 に示す．

子音に関して一元配置の分散分析を行った結果，フレーム C 及びフレーム C-tr. では 1%水準で有意な差が見られたが，フレーム V-tr. 及びフレーム V では有意な差はなかった．多重比較検定の結果，子音区間 (フレーム C 及びフレーム C-tr.) では鼻音 /ma/，/na/ がそれ以外のどの口音よりも有意に比の値が大きいことが分かった．つまり，母音区間 (フレーム V-tr. 及びフレーム V) では先行子音の鼻音・口音の区別に関係なく個人性が反映されており，子音区間では鼻音において口音より顕著に個人性が反映されていると言える．

3　聴取実験

本実験は，著者らによる先行研究 [5, 6] で行われたものである．聴取者は話者 10 名全員をよく知っている健聴者 5 名である．刺激音は Table 1 に示した 10 名分の単音節で，録音時のサンプリング周波数 (48 kHz) のものを用いた．実験は録音と同じ防音室で行われ，1 種類の刺激音は 1 話者につき 5 標本用い，1 種類ごとの評価回数は 250 回であった．

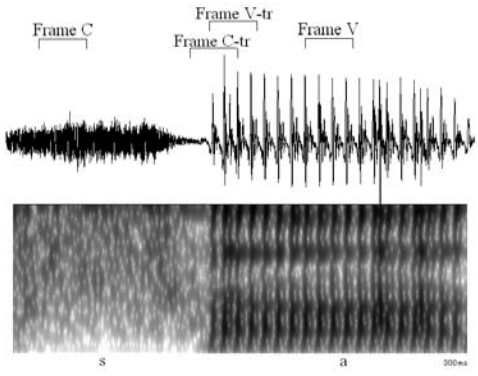

Fig.1 Example of frame excerption

Table 1 Analysis conditions

Speech data	Monosyllables excerpted from carrier sentences: /ma/ /na/ /sa/ /za/ /ta/ and /da/. Five tokens for each speaker and for each consonant were used.
Speakers	Ten male undergraduate students, native speakers of Japanese
Recordings	Recorded onto DAT in a soundproof room, sampled at 48 kHz with 16 bit resolution, then down-sampled at 16 kHz before analysis
Analysis frames	Four types of 30 ms frames were excerpted: 1) C (consonant part), 2) C-tr (consonant incl. transition), 3) V-tr (vowel incl. transition), 4) V (vowel part).

* Speech similarity in perceptual speaker identification, by AMINO, Kanae and ARAI, Takayuki (Sophia Univ.).

Table 2 Rankings of ratios of intra-speaker to inter-speaker distances

Frame C		Frame C-tr.		Frame V-tr.		Frame V	
Syllable	Ratio	Syllable	Ratio	Syllable	Ratio	Syllable	Ratio
/ma/	2.34	/ma/	2.24	/ma/ /na/	2.20	/ma/	2.28
/na/	2.08	/na/	2.06			/sa/	2.25
/za/	1.53	/za/	1.52	/ta/	2.07	/na/	2.21
/sa/	1.44	/sa/	1.39	/sa/	2.06	/ta/	2.11
		/ta/	1.27	/za/	2.05	/za/	1.99
		/da/	1.14	/da/	1.95	/da/	1.93

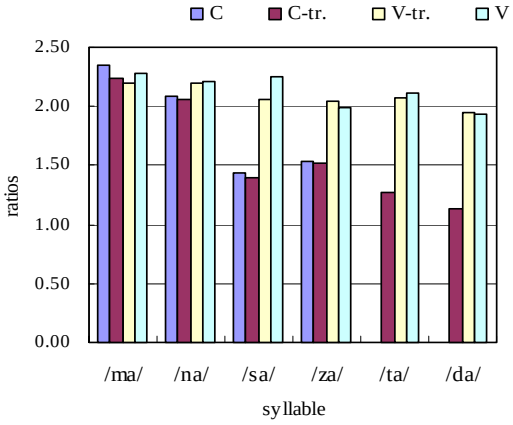

Fig.2 Ratios of intra-speaker to inter-speaker distances

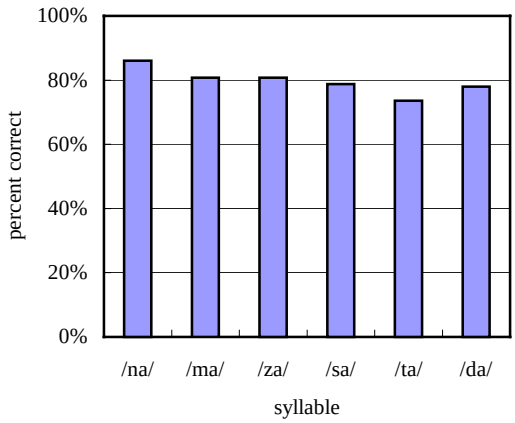

Fig.3 Results of perceptual speaker identification

Table 3 Correlations between perception and distances (N = 100)

	C	C-tr.	V-tr.	V
/ma/	-0.81	-0.79	-0.75	-0.67
/na/	-0.79	-0.77	-0.62	-0.63
/sa/	-0.38	-0.38	-0.66	-0.69
/za/	-0.33	-0.33	-0.63	-0.58
/ta/	N/A	-0.34	-0.60	-0.57
/da/	N/A	-0.31	-0.64	-0.63

先の報告 [5, 6] では Fig.3 に示すような刺激の種類別の正答率のみに着目していた．ここでは，鼻音，摩擦音，口腔閉鎖音の順に高い話者識別率を得たことが分かる．

本研究では話者間の知覚的な距離を，混同行列を用いて表し，前章で求めたケプストラム距離との関連を調べた．混同行列は子音別に作成し，各要素は実際の話者と知覚された話者の混同率を表している．

4　知覚と音響距離の相関

聴取実験での話者間の混同との間の相関を調べるため，話者間・話者内のケプストラム距離も聴取実験の結果と同様，分析フレーム及び子音別に混同行列の形に表した．子音ごとに聴取結果と距離の混同行列間の相関係数を求めた．その結果を Table 3 に示す．

この表から，鼻子音と母音の音節 /ma/, /na/ では全フレームを通じて話者間の音響的距離と知覚における混同の相関が高いのに対し，口子音と母音の音節ではフレーム V-tr. 以降になってようやく高い相関を得ていることが分かる．つまり，/ma/, /na/ では聴取者は話者を識別するのに刺激音全体を用いているのに対し，口子音と母音の音節では母音部のみが話者性判断に利用された可能性がある．

5　まとめ

聴取による話者識別における話者間の混同と刺激音の話者間・話者内距離との関連を調べたところ，聴取実験で正答率が高かった鼻音と母音から成る音節では，子音区間も含め、全体を通じて話者間距離が大きく，聴取者もこの情報を話者性の判断に用いているのに対し，口子音と母音の音節では母音部のみが用いられている可能性が示された．

謝辞

本研究は日本学術振興会特別研究員奨励費 (17・6901) の助成を得た．

引用文献

[1] Nygaard, “The Handbook of Speech Perception,” Blackwell Publishing, 2005.
[2] 西尾，言語生活, 158, 36-42, 1964.
[3] 新美, “音声認識,” 共立出版, 1979.
[4] 網野，信学技報, 104(149), 49-54, 2004.
[5] Amino et al., Proc. Interspeech, 2025-2028, 2005.
[6] Amino et al., Acoust. Sci. Tech., 27 (4), 233-235, 2006.