



HAL
open science

Evaluating the effects of task design on unfamiliar Francophone listener and automatic speaker identification performance

Benjamin O'Brien, Christine Meunier, Natalia Tomashenko, Alain Ghio, Jean-François Bonastre

► To cite this version:

Benjamin O'Brien, Christine Meunier, Natalia Tomashenko, Alain Ghio, Jean-François Bonastre. Evaluating the effects of task design on unfamiliar Francophone listener and automatic speaker identification performance. *Multimedia Tools and Applications*, 2023, 10.1007/s11042-023-15391-0. hal-04138760

HAL Id: hal-04138760

<https://hal.science/hal-04138760v1>


Submitted on 23 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Evaluating the effects of task design on unfamiliar Francophone listener and automatic speaker identification performance

Benjamin O'Brien^{1,2}  · Christine Meunier² · Natalia Tomashenko¹ · Alain Ghio² · Jean-François Bonastre¹

Received: 16 November 2022 / Revised: 17 March 2023 / Accepted: 18 April 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Many questions remain with regards to how context affects perceptual and automatic speaker identification performance. To examine the effects of task design on perceptual speaker identification performance, three tasks were developed, including lineup and binary tasks, as well as a novel clustering task. Speech recordings of native French speakers were compared similarly across tasks evaluated by unfamiliar Francophone listeners. True positive (sensitivity) and true negative (specificity) response rates across tasks were measured. Our results showed participants had similar sensitivity and specificity for binary (88%) and clustering (84%) tasks, but random selection rates for the lineup task. Pearson correlation procedures were used to evaluate the efficiency of scores produced by a state-of-the-art automatic speaker verification to model perceptual responses (equal error rate = 89%). Automatic scores modelled lineup ($r^2 = 0.6$) and clustering ($r^2 = 0.5$) task accuracy quite well, however, they were less robust when modelling binary task responses ($r^2 = -0.2$). The results underscore the role task design plays in shaping perceptual responses, which, in turn, affects the modelling effectiveness of automatic scores. As evidence points to humans and algorithms modelling speakers differently, our findings suggest automatic speaker identification performance might be improved with a greater understanding on how context shapes perceptual responses.

Keywords Voice similarity · Perceptual speaker identification · Task design · Automatic speaker verification

✉ Benjamin O'Brien
benjamin.o-brien@univ-avignon.fr

¹ Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, Avignon, France

² Aix-Marseille Université, CNRS, LPL UMR 7309, Aix-en-Provence, France

1 Introduction

When hearing a voice, listeners are oftentimes capable of obtaining information associated with the speaker, such as their sex, age, emotional state, and whether they are familiar [59]. Recent literature reviews on voice perception [25, 34, 35, 59, 64] have distinguished listeners who are familiar or unfamiliar with speaker voices, where the former relies more on identity processing [25] and the latter perceives and compares voice qualities [70]. However, the methods used by unfamiliar listeners remain unclear.

While evidence suggests unfamiliar listeners use various acoustic features, such as F0 and formant frequencies [4, 26, 29] and phonetic content [6, 51, 57], when identifying speakers, the situation in which they are tasked is equally influential on performance (see Levi [30] and Perrachione [49] for methodological overviews of speaker processing tasks). For example, contextual parameters, such as the number of speakers and environmental factors, have been shown to affect perceptual speaker identification (SID) performance [25]. Several perceptual SID tasks have been designed, and a key distinction between them is whether listeners are tasked with identifying a certain speaker from a set or discriminating between voices. The artificial constraints of the task can influence how listeners engage with speech materials, which, in turn, can affect their voice perception performance.

The potential influence of task on performance can be contextualized in the domain of automatic speaker verification (ASV) systems. Like unfamiliar listeners, ASV systems rely on various acoustic features to develop speaker models that are used to produce scores describing the likelihood that a pair of speech recordings were produced by the same or different speakers (see Singh et al. [60], Poddar et al. [50], and Naika [41] for recent developmental trends in ASV systems). As a way of modelling the perceptual experiences of listeners with voices, ASV systems typically train models on large datasets with many different speakers. Only a handful of studies have compared human and machine performances, which have focused on the effects of speech types, such as the length of utterances [17, 22, 48]. However, to the authors' knowledge no findings have been reported with regards to the effectiveness of automatic scores to model listener responses across perceptual tasks.

Many questions remain regarding the effect of context on perceptual and automatic speaker identification performance. The primary goal of the current study was to examine the effects of task design on perceptual performance by unfamiliar listeners. Since listeners vary in their detection of inter- and intra-speaker variability, e.g., Lavan, Burston, and Garrido [27] and Clopper [10], a major challenge was to develop a method of standardizing speech recordings across tasks, i.e., avoid acoustic feature bias emerging from the different artificial task constraints. In addition, a common metric was required to evaluate performance per speaker across tasks. After collecting listener responses, our second goal was to evaluate how effective automatic scores were at modelling perceptual responses across tasks. Any observations could offer insight into ways ASV systems could be designed differently to improve speaker identification estimation performance, as growing research has shown there are distinctions between how humans and algorithms model speakers.

2 Related work

2.1 Perceptual speaker identification tasks

Listeners evaluating perceptual speaker identification (SID) tasks are typically asked to either identify a *target* speaker in or make discriminations from a set of speech recordings.

A target refers to a specific speaker in a set. Speech recordings from speakers outside of the set are considered *non-target* speakers. Due to study objectives, task designs may vary with regards to the number of (non-)target speaker speech stimuli, which, in turn, can affect perceptual performance. The following outlines three perceptual SID tasks developed for the current study.

The *lineup* [20] task asks listeners to identify a target from a set of speech recordings. Typically, the *utterance* (speech recording) of the target differs from lineup utterances, which are oftentimes the same to avoid phonetic bias. The lineup method has been employed in a number of voice perception studies, including the effects of telephone recordings on identification performance [24, 36, 43, 74]. Although the general facial-recognition task is transformed from the visual to the auditory domain, it is not obvious that visual and auditory displays are comparable from the point of view of memory or perception. While vision provides a global, static scene of individuals to be processed, audition requires sequential processing, as each voice in a set needs to be compared directly to a target voice. This sequential process suggests listeners require a more complex memory process in comparison to viewers. Moreover, evaluations may be compromised, as listeners characterise and compare target and lineup voice qualities while considering the possibility that the target is absent from the lineup [47]. This observation raises further questions regarding the effects of a lineup task on working memory, as the artificial framework introduces the possibility of true negative response (“correct reject”) into an already complex auditory-memory processing chain (see Smith et al. [61, 62] for evidence that the task is error-prone).

The much simpler *same-to-different* (SD) task presents listeners with a pair of speech recordings separated by a short pause and asks them to judge whether they belong to the same or different speakers. SD tasks have been used to test the impartiality of non-target speakers used in lineups [32, 56], as well as to examine the effects of such things as speaker familiarity [4], language familiarity [13, 31], noise [61], and stimuli selection methods [32, 39]. The selection of stimuli used in SD trials requires careful control, as design biases have been shown to drive listener responses. For example, Sussman showed that performance by unfamiliar child and adult listeners was influenced by manipulating same-to-different ratios [66]. Although effective, the low-level SD task is not optimized for identifying which voice qualities listeners associate with target speakers. To do so would require numerous speech recording repetitions for each target speaker, which would be time-consuming and possibly introduce fatigue (see Mühl [39] for a protocol with an approximate 10 minute duration). This observation raises concerns for memory bias or speech priming, as a “fresh voice” is not equivalent to a voice that was presented in a previous SD trial.

The voice sorting task requires listeners to organise speech recordings into groups or *clusters* that represent perceived speaker identities. An alternative to the restrictive types of responses derived from lineup and same-different tasks, voice sorting provides listeners with the opportunity to (re-)listen and (re-)group speech recordings until they are satisfied with their judgements. A major boon of the task is that it neutralises the concept of a “target” speaker, as listeners merely organize voices in terms of their perceived likeness. One potential drawback is that it requires numerous speech recordings per trial. Thus, in order to increase the speaker group homogeneity, listeners are required to make numerous comparisons between speech recordings, which, in turn, can be time-consuming. Nevertheless the method has been used in a number of studies examining the sorting behaviors of familiar and unfamiliar listeners [27, 28, 65]. By instructing participants to sort 32 stimuli into 1

to 32 different speaker groups, Johnson et al. suppressed the possibility of introducing artificial bias [21]. Recently O'Brien et al. [44–46] developed a perceptual *clustering* method and reported unfamiliar listeners were effective at navigating the intuitive interface.

2.2 Automatic speaker verification

Very few studies have compared human and machine speaker identification (SID) performance [3, 16, 55, 58]. One area where perceptual and automatic speaker verification (ASV) system performance have been compared has focused on grouping speech recordings into similar and dissimilar speaker groups. Kelly et al. [23] developed an i-vector-based ASV system to make similar and dissimilar speaker groups and reported that unfamiliar listeners were able to judge male speakers and their similar comparison speakers, but not their dissimilar comparison speakers. In addition, they reported no significant findings for female speakers and their similar and dissimilar comparison speakers, which suggests humans distinguish similar and different voices differently. To maximise and minimise the similarities between speaker groups, O'Brien et al. [46] used a similar process by compressing acoustic features into i-vectors and producing similarity scores via cosine distance with Within Class Covariance Matrix procedures. Park et al. [48] reported that humans outperformed an i-vector-based ASV system when completing a text-independent speaker discrimination task. Revealing a weak correlation between human and machine performance, the authors suggest the two represent speakers differently.

3 Methods

3.1 Stimuli

Speech recordings from 10 female and 10 male native-French speakers were selected from the PTSVox database [7]. Speaker descriptions are detailed in Table 1. The age range of the speakers was 18 to 24 years (mean age 19.7 ± 1.6 years). All speakers read three traditional French-texts, entitled “Ma soeur est venue chez moi hier”, “Au nord du pays, on trouve une espèce du chat”, and “La bise et le soleil se disputaient”. The texts were selected due to their familiarity with native French speakers and rich phonetic content. Speech was recorded in a double-walled, sound attenuated room with a Zoom H4N stereo microphone (sampling rate: 44.1 kHz; bit-depth: 16-bit).

Female and male speakers were separated. The decision to separate speakers by sex was based on findings that have shown listeners are quite capable of discriminating male from female speakers and vice-versa (see Titze [67], Mendoza et al. [37], Whiteside [72], and Wu and Childers [73]). As the goal of our study was to examine the effects of task design on perceptual performance, it was decided to eliminate the potential confounding factor of speaker sex.

Each speaker was assigned to either a Target (in-set) or Non-Target (out-of-set) group. As it was important to create balance in terms of the acoustic difference between each Target speaker and all Non-Target speakers, the fundamental frequency (F0) and speech tempo of each speaker was extracted (Table 1) and used to calculate the standardized euclidean distances (SED) between speakers. A YIN algorithm [8] written in MATLAB 2016b (MathWorks Inc, USA) was used to calculate F0, while speech tempo (phones per second) was obtained in Praat [5]. A custom script was written to select Non-Target speakers with the

Table 1 Description of PTSVox speakers

Speaker	Group	Sex	Age	Region (FR)	F0 (Hz)	Tempo (pho/s)
LG001	Target	Female	19	Alsace	222	12.8
LG002	Target	Male	24	Lorraine	114	13.7
LG003	Target	Female	19	Rhône	200	11.2
LG004	Target	Male	19	Donzère	111	15.3
LG005	Target	Male	18	Saint-Etienne	147	13.4
LG006	Target	Male	22	Loire	125	14.2
LG007	Target	Female	19	Grenoble	190	13.7
LG008	Target	Male	20	Isère	138	14.8
LG009	Target	Female	19	Picardie	202	12.0
LG010	Target	Female	19	Haut-Rhin	227	12.0
LG011	Foil	Female	21	Chaumont	220	13.9
LG012	Foil	Female	20	Bourgogne	220	14.5
LG013	Foil	Male	22	Loire	117	13.5
LG014	Foil	Male	18	Rhône	101	13.8
LG015	Foil	Female	18	Rhône	190	14.0
LG016	Foil	Female	18	La Tour du Pin	180	14.0
LG017	Foil	Male	19	Aisne	115	14.0
LG018	Foil	Male	20	Loire	111	14.4
LG019	Foil	Female	19	Grenoble	206	12.4
LG020	Foil	Male	20	Auvergne	117	14.1

smallest SED. Figure 1 illustrates SED across female and male Target and Non-Target speakers.

For each speaker 24 utterances were extracted with Praat [5] and evenly distributed across Target and Non-Target speaker groups (see Appendix A for French text and English translations). The duration of the utterances ranged from 1.1 to 3.5 s (mean duration 1.9 ± 0.4 s). Limiting the duration of the utterances reduced the possibility of introducing fatigue to participants. In order to compare the effects across tasks, it was important to avoid a threshold effect. All 480 speech recordings were normalised in MATLAB, such that the maximal amplitude of each recording was adjusted to a target of 100% of the signal dynamic.

3.2 Perceptual task designs

Figure 2 provides an illustrated overview of the three perceptual tasks developed for the study. Table 2 describes the number of trials, stimuli per trial, design, and mean duration across tasks. For each task, the sex of the speakers remained the same.

Participants evaluated 30 *Lineup* task trials (random order, non-repeating) programmed in Lancelot [2]. Each Target was presented six times (1:1 ratio present-to-absent in Lineup). Participants were instructed to first listen to the Target utterance located at the top of the interface (see Appendix B) and then each Lineup utterance (unlimited listens). Their task was to decide whether the Target speaker was present in the Lineup by selecting a circle below the Lineup voice. If they believed the Target was absent from the Lineup, they selected a circle below a red 'X'.

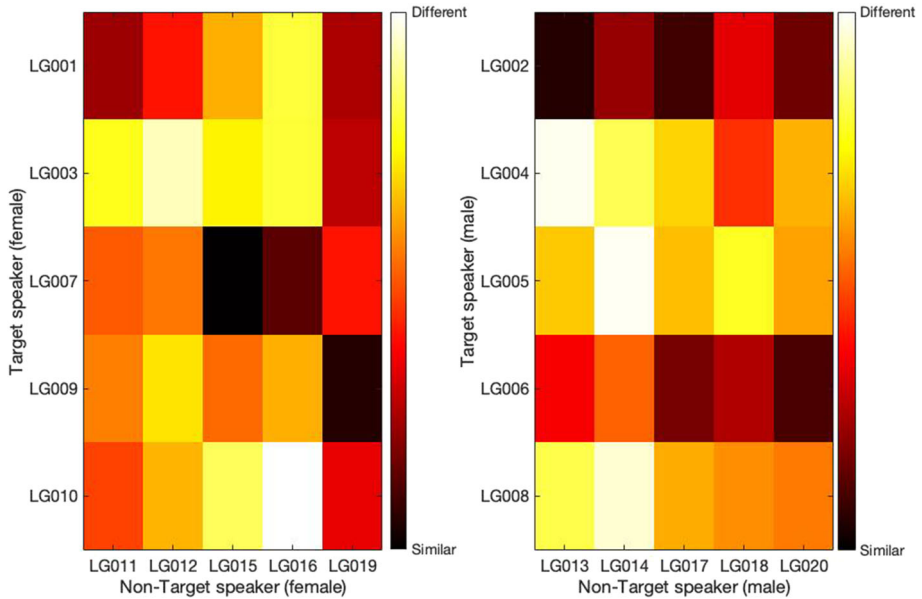


Fig. 1 Heat maps of standardized euclidean distances between each female (Left) and male (Right) Target and Non-Target speakers

Participants evaluated 100 *Same-Different* (SD) task trials (random order, non-repeating) programmed in Perceval [2]. Each SD trial began with a short “beep” generated by a sinusoidal oscillator (frequency: 500 Hz; duration: 0.8 s). Following 2 s of silence, a speech recording was automated with *Voix A* (“Voice A”) text displayed in a yellow rectangle. Following 0.5 s of silence without an image, a different speech recording with *Voix B* (“Voice B”) in a blue rectangle. Participants had 5 s to decide whether the two voices belonged to the same speaker or different speakers by pressing a button on the left or right, respectively (see Appendix B). Each Target (A) and Non-Target (B) speakers were presented 1:1 ratio same-to-different. For each different trial, Target (A) and Non-Target (B) speakers were presented AB and BA.

Participants evaluated 10 *Cluster* task trials programmed in a state-of-the-art interface developed at Laboratoire Informatique d’Avignon, Université du Vaucluse-Avignon (open source and available upon request). Each trial was composed of 12 speech recordings derived from the Lineup task trials: the six utterances of each Target speaker were randomly distributed across two trials (balanced) with the remaining nine speech recordings composed of two to five Non-Target utterances. For each Cluster trial participants were tasked with listening to each recording (unlimited) and classifying it into a cluster representing a unique speaker identity. To classify a speech recording, participants were instructed to right-click on the circle, which revealed a drop-down menu with classification colors (Appendix D).

3.3 Participants

All participants were native-French speakers and reported good hearing. They consented to voluntary participation in the study and were compensated for their time. The study was approved by the Ethics Committee of Aix-Marseille University.

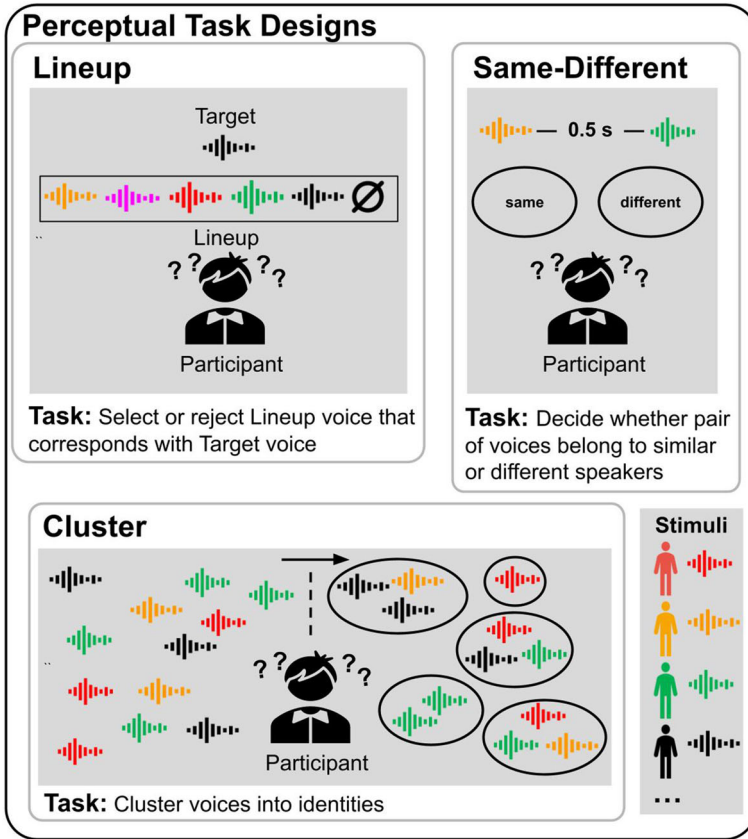


Fig. 2 Illustration of the perceptual task designs

Table 2 Overview of perceptual tasks

Task	Trials (#)	Stimuli (#)	Design	Duration (s)
Lineup	30	6	Different utterance for Target & Lineup speakers	27.7 ± 8.6
			Same utterance for Lineup speakers	
			Target speaker in Lineup 1/2-trials	
			Target speaker out of Lineup 1/2-trials	
Same-Different	100	2	Different utterance per trial	1.3 ± 0.3
			Same speaker for 1/2-trials	
Cluster	10	12	Different speaker for 1/2-trials	69.3 ± 12
			Different utterance per trial	
			4 different speakers per trial 2–5 different utterances per speaker	

35 people (27 F and 8 M; mean age 26.2 ± 8.0 years) evaluated Lineup and Same-Different task trials on desktop computers at CEP-LPL. Throughout the study participants wore Superlux HD 681B headphones. Prior to testing, participants listened to a speech recording and adjusted the volume to their comfort.

19 people (17 F and 2 M; mean age 24.8 ± 2.7 years) evaluated Cluster task trials online. This change in testing location was due to the 2020 pandemic. Participants were encouraged to use personal headphones and were provided detailed instructions on how to complete the task and use the interface.

3.4 Automatic speaker verification system

The state-of-the-art automatic speaker verification (ASV) model developed for the current study was trained on the *VoxCeleb-1,2* [9, 40] corpus, which contains around 2800 hours of multilingual speech from 7363 (2912 F and 4451 M) speakers. This corpus was extracted from videos uploaded to YouTube and designed for speaker verification research. Similar to recent work by Tomashenko et al. [68, 69], the Kaldi toolkit [52] was used to train the ASV model. As shown in Fig. 3, the ASV model relies on x-vector [63] speaker embeddings and probabilistic linear discriminant analysis (PLDA) [53]. The ASV model has a time delay neural network (TDNN) architecture with the following configuration. 30-dimensional Mel Frequency Cepstral Coefficients (MFCC) were used as input features. The model contains 7 hidden layers including a single (6th) statistics pooling layer. The statistics pooling layer aggregates all frame-level outputs from the previous (5th) layer and computes its mean and standard deviation. The dimension of the output layer is 7232 that corresponds to speaker ids. The neural network was trained to classify the speakers in the training data using cross entropy criteria. The 512-dimensional x-vectors were extracted after the statistics pooling layer. Additional details about model training can be found in Tomashenko et al. [69] and Snyder et al. [63].

The ASV model was used to obtain PLDA scores [18] for evaluated pairs of speech recordings, where s_a, s_b denote a pair of utterances. PLDA scores were computed as log-likelihood ratios (LLR) between corresponding x-vectors x_a, x_b as (1):

$$\text{PLDA}(s_a, s_b) = \log \frac{P(x_a, x_b | H_{\text{same}})}{P(x_a, x_b | H_{\text{different}})}, \quad (1)$$

where H_{same} and $H_{\text{different}}$ are the hypotheses *same speaker* and *different speakers*, respectively. Following the training with the *VoxCeleb* dataset, we reported an equal error rate of 11.55% and corresponding true positive (sensitivity) and true negative (specificity) response rates are provided in Table 3.

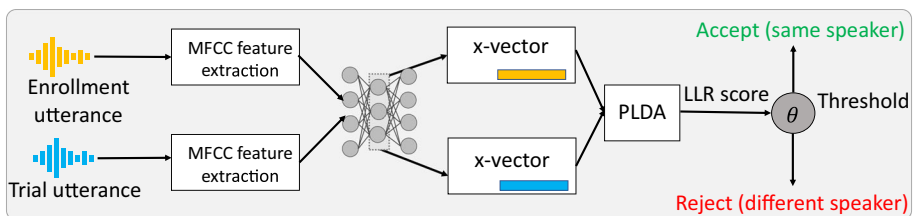


Fig. 3 Illustration of the ASV system

Table 3 Sensitivity, specificity, and temporal responses across tasks

Task	Metric	Result
Lineup	sensitivity	49.1%
	trial duration	25.9 ± 14.4 s
	specificity	34.9%
	trial duration	31.5 ± 16.4 s
Same-Different	sensitivity	88.4%
	reaction time	1.26 ± 0.5 s
	specificity	88.0%
	reaction time	1.51 ± 0.7 s
Cluster	sensitivity	83.0%
	listen count	3.8 ± 2.2
	specificity	84.1%
	listen count	5.1 ± 2.53
ASV (log-likelihood ratio)	sensitivity	88.5%
	specificity	88.5%

3.5 Data processing

To measure the effect of task on response performance, sensitivity, commonly known as “hit” rate, and specificity, or “correct reject” rate, were obtained from each participant per task. Equations (2) and (3) describe sensitivity and specificity metrics, where TP , TN , FP , and FN represent the number of true positive, true negative, false positive, and false negative responses, respectively.

$$sensitivity = \frac{TP}{TP + FP} \quad (2)$$

$$specificity = \frac{TN}{TN + FN} \quad (3)$$

For each Lineup task trial participants received a TP or TN for correctly identifying or rejecting the Target from the Lineup. Otherwise, they received a FP or FN for falsely identifying a Non-Target or incorrectly rejecting the Target from the Lineup, respectively. For each SD task trial participants received a TP or TN for correctly identifying the pair of voices as belonging to the same or different speakers, respectively. Otherwise, they received a FN or FP for the trial. For each Cluster task trial, mean specificity, i.e., the number of Target utterances in a cluster divided by the cluster size, and mean sensitivity, i.e., the number of the same Non-Target speaker utterances in a cluster divided by cluster size, were calculated.

To accurately reflect task design discrepancies, i.e., the different number of stimuli and outcomes per task, scores were adjusted by a task baseline coefficient. For each task accuracy ((4)) was simulated in MATLAB by randomly making responses after 10,000 trials.

$$accuracy = \frac{TP + TN}{TP + TN + TP + FN} \quad (4)$$

After 40 simulations, the mean accuracy for Lineup, Same-Different, and Cluster tasks were 16.5%, 49.8%, 45.6%, respectively. Equation (5) describes original S_i and adjusted A_i ,

sensitivity and specificity and task baseline coefficient e_t :

$$A_t = \frac{S_t - e_t}{100 - e_t} \quad (5)$$

Linear mixed models (*lmer* from the *lme4* R-package) were used to evaluate the effects of task design on perceptual performance. Task (Lineup, Same-Different, Cluster), Target speaker (10 total), and response type (sensitivity, specificity) were set to fixed factors with random participant factor. Chi-squared ($\chi_{d,N}^2$) tests were used to report p -values (*Anova* from the *car* R-Package) with d degrees of freedom and N samples. Main effects were reported for task, response, and their interactions with speaker. Estimated marginal means (*emmeans*) were used to conduct pairwise comparisons, where $X \pm Y$ represent mean and standard error, respectively.

Pearson correlation procedures were used to evaluate the effects of task design on efficiency of automatic scores to model perceptual responses. In addition to mean accuracy per trial, different task-dependent temporal metrics were measured: Lineup task trial duration (s); Same-Different task trial reaction time (s); and mean number of listens (“listen count”) in a cluster for the Cluster task. For automatic scores, log-likelihood ratios (LLR) were used differently across tasks. For each Lineup task trial, the LLR between the Target and the selected Lineup utterance was used, except when the Lineup was rejected, whereupon a mean value was calculated from Target and Lineup utterances. For each SD task trial the LLR for each pair were used. A mean LLR value was calculated from all cluster utterances in a Cluster task trial.

3.6 Preliminary analysis

To evaluate participant normalcy, normal distribution functions were fitted to the mean trial duration across tasks. All participant data were included, except responses collected from two SD task participants, i.e., their means were greater than three standard deviations from the group mean. Table 2 illustrates mean duration per task. Table 3 illustrates sensitivity, specificity, and temporal responses across tasks.

4 Results

Main effects of adjusted score were observed for task $\chi_{2,3718}^2 = 301.22$ and response type $\chi_{1,3718}^2 = 21.06$, $p < 0.001$, as well as interactions task x response x target speaker $\chi_{18,3718}^2 = 35.94$, $p < 0.01$. Participants were more accurate when evaluating SD ($79.2 \pm 2.3\%$) and Cluster ($83.2 \pm 3.6\%$) tasks in comparison to the Lineup task ($43.4 \pm 2.1\%$), $p < 0.001$. They were also more sensitive ($71.0 \pm 2.2\%$) rather than selective ($65.3 \pm 2.0\%$), $p < 0.01$.

Pairwise comparisons on task and response type interactions revealed participants performed better when evaluating SD (sensitivity: $82.6 \pm 3.5\%$; specificity: $75.8 \pm 2.0\%$) and Cluster (sensitivity: $82.2 \pm 4.6\%$; specificity: $84.1 \pm 4.6\%$) task trials in comparison to Lineup task trials (sensitivity: $50.9 \pm 2.5\%$; specificity: $35.9 \pm 2.8\%$), $p < 0.001$. The following describes pairwise comparisons for interactions between task x response x speaker: Table 4 compares task sensitivity and specificity across speakers; Table 5 compares task sensitivity across speakers; and Table 6 compares task specificity across speakers. Figure 4 illustrates response type (sensitivity, specificity) and task (Cluster, Lineup, Same-Different) interactions across speakers.

Table 4 Within-task performance across speakers. mean \pm se, t , and {*, **, ***} represent mean difference (true positive - true negative) and standard error, t-ratio, and $p < \{0.05, 0.01, 0.001\}$ significance, respectively

Speaker	Cluster			Lineup			Same-Different		
	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p
LG001	-19.1 \pm 19.5	-1.0		30.0 \pm 9.6	3.1	**	12.9 \pm 10.7	1.2	
LG002	-3.3 \pm 17.2	-0.2		25.5 \pm 9.4	2.7	**	38.8 \pm 11.1	3.5	***
LG003	1.0 \pm 20.8	0.0		2.9 \pm 10.1	0.3		12.9 \pm 10.8	1.2	
LG004	1.9 \pm 17.7	0.1		13.5 \pm 9.6	1.4		3.7 \pm 11.1	0.3	
LG005	10.5 \pm 16.8	0.6		0.2 \pm 10.4	0.0		41.3 \pm 11.1	3.7	***
LG006	3.3 \pm 6.8	0.2		4.4 \pm 10.3	0.4		10.0 \pm 11.1	0.9	
LG007	-21.6 \pm 20.1	-1.1		24.7 \pm 9.8	2.5	*	54.4 \pm 10.8	-5.1	***
LG008	6.1 \pm 20.1	0.3		11.9 \pm 10.1	1.2		-22.5 \pm 11.1	-2.0	*
LG009	43.6 \pm 20.3	0.3		22.8 \pm 10.2	2.3	*	28.2 \pm 10.7	2.6	***
LG010	-3.5 \pm 20.1	-0.2		19.0 \pm 10.4	1.8		3.5 \pm 10.8	0.3	

Table 7 shows the results of Pearson correlation procedures applied to log-likelihood ratios and perceptual responses (trial accuracy, temporal metrics) across tasks.

5 Discussion

The primary goal of the current study was to examine whether perceptual SID task design affected performance by unfamiliar listeners. Our findings revealed participants performed both Same-Different (SD) and Cluster tasks relatively similarly with sensitivity and specificity greater than 80%, however, performance dropped below 50% when evaluating Lineup task trials. In general, the task comparison results confirmed our hypothesis that the degree

Table 5 Task sensitivity across Target speakers. mean \pm se, t , and {*, **, ***} represent mean difference (between tasks) and standard error, t-ratio, and $p < \{0.05, 0.01, 0.001\}$ significance, respectively

Speaker	Cluster - Lineup			Same-Different - Lineup			Same-Different - Cluster		
	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p
LG001	24.1 \pm 15.5	1.6		52.8 \pm 11.8	4.5	***	28.7 \pm 17.3	1.7	
LG002	39.5 \pm 13.9	2.8	*	50.0 \pm 12.0	4.2	***	-10.5 \pm 16.1	-0.7	
LG003	45.8 \pm 16.6	2.8	*	24.8 \pm 12.2	2.0		-21.0 \pm 18.0	-1.2	
LG004	38.6 \pm 14.3	2.7	*	39.3 \pm 12.2	3.22	**	0.7 \pm 16.3	0.1	
LG005	30.9 \pm 14.1	2.2		-10.7 \pm 12.6	-0.9		-41.6 \pm 15.9	-2.6	*
LG006	25.4 \pm 14.1	1.8		28.0 \pm 12.5	2.3		2.7 \pm 15.9	0.2	
LG007	18.9 \pm 16.0	1.2		51.9 \pm 11.9	4.3	***	33.0 \pm 17.6	1.9	***
LG008	30.5 \pm 14.4	2.1		5.3 \pm 12.4	0.4		-25.2 \pm 16.3	-1.5	
LG009	39.0 \pm 16.0	2.4	*	44.6 \pm 12.0	3.7	***	5.7 \pm 17.6	0.3	
LG010	21.1 \pm 16.1	1.3		31.3 \pm 12.1	2.6	*	10.2 \pm 17.6	0.6	

Table 6 Task specificity across Target speakers. mean \pm se, t , and {*, **, ***} represent mean difference (between tasks) and standard error, t-ratio, and $p < \{0.05, 0.01, 0.001\}$ significance, respectively

Speaker	Cluster - Lineup			Same-Different - Lineup			Same-Different - Cluster		
	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p	Mean \pm se (%)	t	p
LG001	73.1 \pm 16.5	4.5	***	69.8 \pm 9.1	7.6	***	-3.3 \pm 14.1	-0.2	
LG002	68.4 \pm 14.5	4.7	***	36.8 \pm 9.0	4.1	***	-31.5 \pm 13.3	-2.4	*
LG003	42.0 \pm 16.8	2.5	*	34.8 \pm 9.0	3.9	***	-7.1 \pm 15.8	-0.5	
LG004	50.1 \pm 14.7	3.5	*	56.5 \pm 9.0	6.3	***	6.4 \pm 13.6	0.5	
LG005	20.7 \pm 14.5	1.4		30.8 \pm 9.3	3.3	**	10.2 \pm 13.1	0.7	
LG006	26.5 \pm 14.5	1.8		22.5 \pm 9.4	2.4	*	-4.0 \pm 13.1	-0.3	
LG007	65.2 \pm 16.5	4.0	***	22.1 \pm 9.1	2.4	*	-43.1 \pm 15.3	-2.8	*
LG008	36.5 \pm 14.9	2.4	*	39.7 \pm 9.3	4.3	***	3.2 \pm 13.6	0.2	
LG009	55.6 \pm 16.6	3.3	**	39.2 \pm 9.4	4.2	***	16.4 \pm 15.3	-1.1	
LG010	43.6 \pm 16.7	2.6	*	46.8 \pm 9.5	4.9	***	3.2 \pm 15.3	0.2	

of constraints designed into a perceptual SID task can influence performance. The target-absent feature distinguished the Lineup task from the other tasks, which had an adverse effect on performance. Participants appeared to be more inclined to find a target despite its absence from the lineup, which, in turn, decreased specificity in comparison to sensitivity (Table 3). Our findings were consistent with those reported in Smith et al. [61], which found participants were 39% accurate when identifying targets present in lineups, while

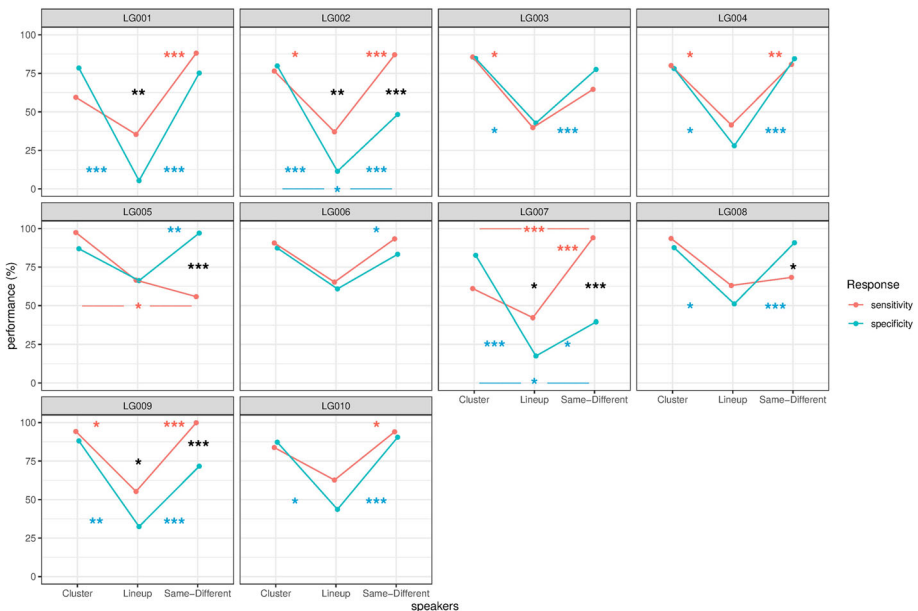


Fig. 4 Interactions between performance (sensitivity, specificity) and tasks (Cluster, Lineup, Same-Different) across speakers. {*, **, ***} represent $p < \{0.05, 0.01, 0.001\}$. Black represents within-task performance, while red and blue represent sensitivity and specificity, respectively, across tasks

Table 7 Pearson correlation procedures between log-likelihood ratio scores and mean accuracy and temporal-metrics across tasks. {*, ***} represent $p < \{0.05, 0.001\}$

Task	Metric	Log-likelihood ratio	
		ρ	p
Lineup	accuracy	0.49	***
	trial duration	-0.6	***
Same-Different	accuracy	-0.21	*
	response time	0.08	
Cluster	accuracy	0.58	***
	listen count	-0.26	***

only 6% accurate when judging their absence. These findings underscore the importance of minimizing artificial biases designed into perceptual SID tasks.

Although SD and Cluster task performance differed significantly from the Lineup task, no significant main effects were observed between them. Our observations are consistent with those reported in Johnson et al. [21], which found no significant correlations between voice discrimination and sorting tasks. These collective findings suggest each task requires unique processing of sensory information. On one hand, Jenson and Saltuklaroglu [19] showed same-different tasks affect working memory processing, where more recent items are processed more rapidly and efficiently. The authors found left hemisphere brain activations were stronger during different speaker trials and weaker activations were observed in the right hemisphere during same speaker trials. Their findings suggest that the mismatch between different speech materials leads to a shift in speech and language processing (left hemisphere), whereas the repetition of the same speaker leads predictive coding to repetition suppression. This distinction between auditory and decision-making processing was investigated by Venezia et al. [71], who used different signal-to-noise ratios to neutralise perceptual speech processing variability. The authors identified brain regions that were activated during the decision making process, i.e., the temporal lobe was involved in speech analysis processing, whereas motor-related regions were involved in task responses. These findings highlight that although same-different tasks are simple and efficient, they appear to divide processing and are sensitive to bias via stimuli sequencing. On the other hand, the more natural and open-ended Cluster task provided listeners with a platform to dynamically engage with the speech materials in highly personal ways. Interestingly, this increase in sensory information did not encumber performance. Lavan et al. [28] suggested that the voice sorting task provided familiar listeners with an advantage over unfamiliar listeners. Although the listeners in the current study were unfamiliar with speakers, the Cluster task results suggest they were able to take advantage of any accessible information, i.e., Target or Non-Target utterances alike, when grouping voices into perceived identities. This observation is consistent with O'Brien et al. [46], where 20 speech stimuli were provided. While all auditory perceptual tasks are constrained to sequentially processing, the Cluster task affords listeners with time to perceive and compare vocal qualities extracted from all available stimuli.

The stimuli used in the current study were consistent across tasks, which made it possible to observe any interactions with specific Target speakers on performance. In general some Target speakers were more difficult to discriminate from Non-Target speakers. For example, the lowest mean specificity across tasks was from speakers LG002 (46.5%) and LG007 (46.6%), who appeared to be quite similar to Non-Target speakers (see Table 1 and Fig. 1). It was possible that features, such as pitch [4], vowel quality [38, 42], and speech

tempo [11, 54] were difficult for unfamiliar listeners to process from non-regional speakers, as suggested by findings in Dufour, Nguyen, and Frauenfelder [12]. The authors reported that standard-French was perceived differently depending on a listener's regional accent. As a majority of the participants were associated with Aix-Marseille University and originated from the region, it is plausible that they perceived the speakers from various French regions differently. The effects of accent on listener perception were also studied in Floccia et al. [14], which reported that, in order to overcome regional accents, unfamiliar listeners required short-term speech processing adjustments. This finding suggests that, when considering the different perceptual SID tasks, a voice sorting or clustering task is optimal, as it provides unfamiliar listeners with time to familiarise themselves with the vocal characteristics of unfamiliar speakers, i.e., they can capitalise their judgements with a larger set of speech stimuli. Alternatively, these findings suggest reaction times during perceptual binary tasks trials are affected by the presence of unfamiliar accents.

The secondary goal of the study was to evaluate whether task influenced the effectiveness of using automatic judgements to model perceptual performance. First, the log-likelihood ratios (LLR) had comparable sensitivity and specificity (88.5%) to the SD task. This is a promising observation, but not entirely surprising, as ASV systems similarly evaluate and judge speech recording pairs. It is likely that performance could be improved by training with a different (French) dataset.

Pearson correlation procedures revealed significant correlations between automatic and perceptual accuracy across tasks. In general these findings suggest LLR provide a good measurement for estimating the unfamiliar listener performance, however, their precision depends on the design of the perceptual SID task. The findings support those reported by Gerlach et al. [15], who observed positive relationships between listener judgements and automatic speaker recognition scores for both English and German language speakers. Interestingly the positive relationships between trial accuracy and LLR were only observed for the Lineup and Cluster tasks. However, this trend was not observed for the SD task. This is an important observation, when considering the use of ASV systems to select non-target speakers used in perceptual SID task. This difference in trends between tasks can be considered alongside observations reported by Lindh and Eriksson [33] and Zetterholm, Blomberg, and Elenius [75], which found differences between human judgements and scores produced by automatic speaker verifications. Taken together, these findings continue to support the idea that there are important nuances between how human and machines model speakers [48]. Moreover, they underline a larger issue considering the use of ASV systems: despite providing information regarding their efficiency, i.e., equal error rates, they do not provide any additional information as to how speakers are modelled and how pairs are judged. Recent developments by Amor and Bonastre [1] aim to provide metrics that explain decisions made by ASV systems.

When considering the relationship between task-dependent temporal metrics and LLR, significant negative correlations were observed for Lineup and Cluster tasks. Interestingly no significant correlation was observed between reaction time and LLR for SD task trials. This finding suggests correlating unfamiliar listener reaction times with likelihood scores based on a pair of speech recordings is far too limited. Alternatively they suggest that perceptual SID tasks with more than two speech recordings or response types provide wider contexts in which to identify the capacities and limitations of individual listeners. Moreover, likelihood scores generated by ASV systems appear to be more suited to model perceptual SID tasks designed with multiple speech recordings, response types, and unlimited listens, as observed with the Lineup and Cluster tasks.

6 Conclusion

This paper detailed the development of three perceptual speaker identification tasks with a similar set of speech stimuli. The findings served as important benchmarks for the effects of task design on perceptual performance by unfamiliar French listeners. While optimising perceptual performance is important, its value depends on whether the task provides users with the means to make correct choices and avoids introducing artificial biases. Although both humans and machines complete pairwise comparisons of speech materials in order to evaluate their similarity, there are still differences in their approaches to modelling speakers. Our results revealed context affects the efficiency of using automatic scores to model perceptual performance. One approach to improving automatic speaker verification system performance is to consider how context shapes listeners responses. In comparison to more traditional tasks, the perceptual clustering method developed for the study highlighted how unfamiliar listeners performed at a high level, which correlated strongly to log-likelihood scores ($r^2 = 0.5$). Because of its design the cluster task produces a manifold of responses that can represent the perceptual profile of each listener. In comparison to more restrictive perceptual tasks, it is much more-detailed and sophisticated, capable of capturing nuances via speech material groupings. Future research in automatic speaker verification systems might aim to develop methods that, like cluster tasks, provide a context in which speaker model training takes advantage of all available materials.

AppendixA: Speech utterances

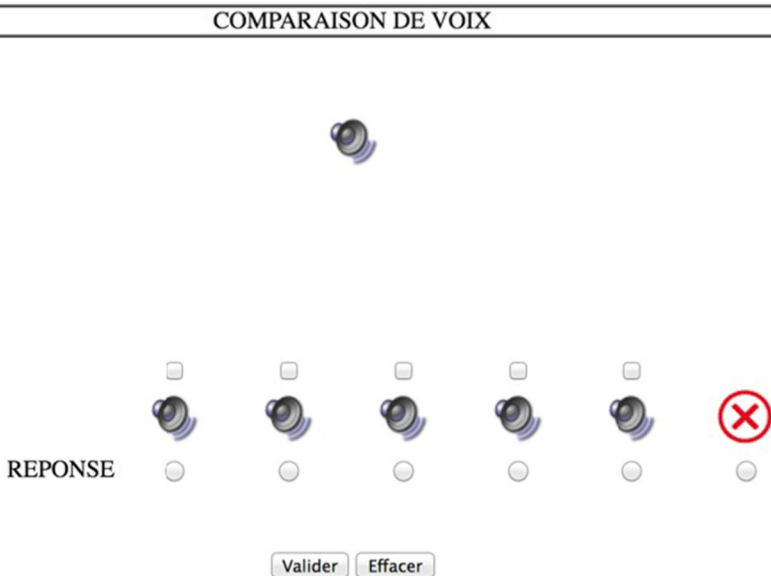
Description of each speech utterance group assignment, French text, and English translation.

Group	Text	English translation
Target	je m'approchais du bord de la fenêtre	I approached the edge of the window
Target	serrait son manteau autour de lui	[he] tightened his coat around him
Target	on trouve une espèce de chat	we found a species of cat
Target	la bise et le soleil se disputaient	the wind and the sun were fighting
Target	pour rencontrer ces deux espèces	to meet these two species
Target	le soleil a commencé à briller	the sun began to shine
Target	faire ôter son manteau au voyageur	to make the traveler take off his coat
Target	s'éloignant d'un nid perché sur un arbre	[it] moved away from a nest perched on a tree
Target	il avait dû faire fuir l'oiseau	he had to scare the bird away
Target	son plumage était beau et doux	its plumage was beautiful and soft
Target	ses deux ailes étaient blessées	his two wings were injured
Target	serait regardé comme le plus fort	[he] would be regarded as the strongest

Group	Text	English translation
Non-Target	vit une colonie d'oiseaux	lives a colony of birds
Non-Target	ma sœur n'a qu'à traverser la rue	my sister only has to cross the street
Non-Target	son cœur battait très vite	his heart was beating very fast
Non-Target	que le soleil était le plus fort des deux	that the sun was the strongest of the two
Non-Target	au cœur d'un parc naturel	in the heart of a natural park
Non-Target	sur le coup de midi	at the stroke of noon
Non-Target	pour regarder dans la rue	to look in the street
Non-Target	quand ils ont vu un voyageur qui s'avancait	when they saw a traveler coming forward
Non-Target	leur poil est beau et doux	his hair is beautiful and soft
Non-Target	ma sœur est venue chez moi hier	my sister came to my house yesterday
Non-Target	elle me parlait de ses vacances en mer du Nord	she spoke about her vacations at the North Sea
Non-Target	ils sont noirs avec deux tâches blanches sur le dos	they are black with two white spots on their backs

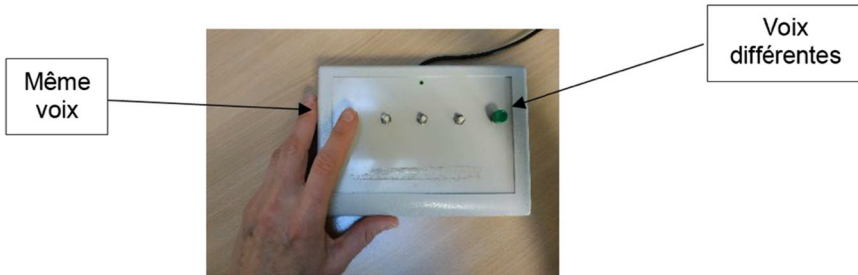
Appendix B: Lineup task interface

A screenshot of the Lineup task interface, where each speaker icon represents a speech recording. The top speaker icon is the Target, while the rest constitute the Lineup. The following are French to English translations: “Comparaison de voix” (“Voice comparison”); “Reponse” (“response”); “Valider” (“Confirm”); “Effacer” (“Reset”). The boxes above the Lineup were optional for marking progress.



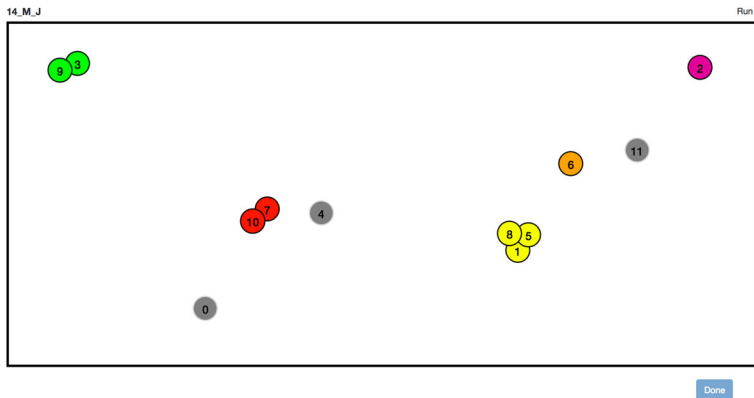
Appendix C: Same-Different task interface

A screenshot of the Same-Different interface. After listening to two speech recordings separated by 2 s, they decided whether the voices belonged to the same (“Même voix”) or different speakers (“Voix différentes”).



Appendix D: Cluster task interface

A screenshot of the Cluster task interface. Each numbered circle represents a speech recording and the color and grey assignments represent whether it has been assigned or not to a specific speaker, respectively.



Acknowledgments This work was funded by the French National Research Agency (ANR) under the VoxCrim project (ANR-17-CE39-0016).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Amor IB, Bonastre J-F (2022) Ba-lr: binary-attribute-based likelihood ratio estimation for forensic voice comparison. In: 2022 International workshop on biometrics and forensics (IWBF), pp 1–6. <https://doi.org/10.1109/IWBF55382.2022.9794542>
2. Andre C, Ghio A, Cavé C, Teston B (2007) Perceval: a computer-driven system for experimentation on auditory and visual perception. In: International congress of phonetic sciences, pp 1421–1424
3. Audibert N, Larcher A, Kahn J, Rossato S, Matrouf D, Bonastre J-F (2010) Lia human-based system description for nist hasr 2010. In: Proceedings NIST HASR 2010
4. Baumann O, Belin P (2008) Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol Res* 74:110–20. <https://doi.org/10.1007/s00426-008-0185-z>
5. Boersma P (2001) Praat, a system for doing phonetics by computer. 5(9/10):341–345
6. Bricker PD, Pruzanski S (1966) Effects of stimulus content and duration on talker identification. *J Acoust Soc Am* 40:1442–1449
7. Chanclu A, Georgeton L, Fredouille C, Bonastre J-F (2020) Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire. In: 6e Conférence conjointe journées d'études sur la parole, pp 73–81
8. Cheveigné A, Kawahara H (2002) Yin, a fundamental frequency estimator for speech and music. *J Acoust Soc Am* 111:1917–30. <https://doi.org/10.1121/1.1458024>
9. Chung JS, Nagrani A, Zisserman A (2018) VoxCeleb2: deep speaker recognition. In: Interspeech, pp 1086–1090
10. Clopper C, Pisoni D (2004) Some acoustic cues for the perceptual categorization of American English regional dialects. *J Phon* 32:111–140. [https://doi.org/10.1016/S0095-4470\(03\)00009-3](https://doi.org/10.1016/S0095-4470(03)00009-3)
11. Dellwo V, Ferragne E, Pellegrino F (2006) The perception of intended speech rate in English, French, and German by French speakers. In: Speech Prosody. Dresden, Germany, p 2006. <https://hal-univ-paris.archives-ouvertes.fr/hal-01240462>
12. Dufour S, Nguyen N, Frauenfelder UH (2007) The perception of phonemic contrasts in a non-native dialect. *J Acoust Soc Am* 121(4):131–136. <https://doi.org/10.1121/1.2710742>
13. Fleming D, Giordano B, Caldara R, Belin P (2014) A language-familiarity effect for speaker discrimination without comprehension. *Proc Natl Acad Sci* 111:0. <https://doi.org/10.1073/pnas.1401383111>
14. Floccia C, Goslin J, Girard F, Konopczynski G (2006) Does a regional accent perturb speech processing. *J Exp Psychol: Hum Percept Perform* 32:1276–93. <https://doi.org/10.1037/0096-1523.32.5.1276>
15. Gerlach L, McDougall K, Kelly F, Alexander A, Nolan F (2020) Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Comm*, 124. <https://doi.org/10.1016/j.specom.2020.08.003>
16. Greenberg C, Martin A, Brandschain L, Campbell J, Cieri C, Doddington G, Godfrey J (2010) Human assisted speaker recognition in NIST SRE: 10. In: Proceedings of the speaker and language recognition workshop (Odyssey 2010), p 32
17. Hautamäki V, Kinnunen T, Nosrati Ghods M, Lee KA, Ma B, Li H (2010) Approaching human listener accuracy with modern speaker verification. In: Proceedings of Interspeech 2010, pp 1473–1476. <https://doi.org/10.21437/Interspeech.2010-152>
18. Ioffe S (2006) Probabilistic linear discriminant analysis. In: European conference on computer vision. Springer, pp 531–542
19. Jensen D, Saltuklaroglu T (2021) Sensorimotor contributions to working memory differ between the discrimination of same and different syllable pairs. *Neuropsychologia* 159:107947. <https://doi.org/10.1016/j.neuropsychologia.2021.107947>
20. Jong GD, Nolan F, McDougall K, Hudson T (2015) Voice lineups: a practical guide. In: ICPhS
21. Johnson J, Mcgettigan C, Lavan N (2020) Comparing unfamiliar voice and face identity perception using identity sorting tasks. *Q J Exp Psychol* 73:174702182093865. <https://doi.org/10.1177/1747021820938659>
22. Kahn J, Audibert N, Rossato S, Bonastre J-F (2011) Speaker verification by inexperienced and experienced listeners vs. speaker verification system. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5912–5915. <https://doi.org/10.1109/ICASSP.2011.5947707>
23. Kelly F, Alexander A, Forth O, Kent S, Lindh J (2016) Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. In: INTERSPEECH
24. Kerstholt J, Jansen E, Amelvoort A, Broeders T (2006) Earwitnesses: effects of accent, retention and telephone. *Appl Cogn Psychol* 20:187–197. <https://doi.org/10.1002/acp.1175>
25. Kreiman J, Van Lancker Sidtis D (2011) Foundations of voice studies: an interdisciplinary approach to voice production and perception. <https://doi.org/10.1002/9781444395068>

26. LaRivière C (1971) Some acoustic and perceptual correlates of speaker identification. In: Proceedings of the seventh international congress of phonetic sciences, pp 558–564
27. Lavan N, Burston L, Garrido L (2018) How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *Brit J Psychol* 0:110. <https://doi.org/10.1111/bjop.12348>
28. Lavan N, Kreitewolf J, Obleser J, McGettigan C (2021) Familiarity and task context shape the use of acoustic information in voice identity perception. *Cognition* 215:104780. <https://doi.org/10.1016/j.cognition.2021.104780>
29. Lavner Y, Gath I, Rosenhouse J (2000) The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Comm* 30(1):9–26. [https://doi.org/10.1016/S0167-6393\(99\)00028-X](https://doi.org/10.1016/S0167-6393(99)00028-X)
30. Levi SV (2019) Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cogn Sci* 10(2):1483. <https://doi.org/10.1002/wcs.1483>. <https://arxiv.org/abs/https://wires.onlinelibrary.wiley.com/doi/pdf/>. <https://wires.onlinelibrary.wiley.com/doi/pdf/doi:10.1002/wcs.1483>
31. Levi S, Schwartz RG (2013) The development of language-specific and language-independent talker processing. *J Speech Lang Hear Res: JSLHR*. [https://doi.org/10.1044/1092-4388\(2012\)12-0095](https://doi.org/10.1044/1092-4388(2012)12-0095)
32. Lindh J (2009) Perception of voice similarity and the results of a voice line-up. In: Proceedings of FONETIK 2009, The XXII Swedish phonetics conference, Department of Linguistics, pp 186–1897
33. Lindh J, Eriksson A (2010) Voice similarity—a comparison between judgements by human listeners and automatic voice comparison. In: Proceedings of FONETIK 2010, pp 63–69
34. Mathias SR, von Kriegstein K (2014) How do we recognise who is speaking? *Front Biosci-Scholar* 6(1):92–109. <https://doi.org/10.2741/S417>
35. Mattys S, Davis M, Bradlow A, Scott S (2012) Speech recognition in adverse conditions: a review. *Lang Cogn Process* 27:953–978. <https://doi.org/10.1080/01690965.2012.705006>
36. McDougall K, Nolan F, Hudson T (2015) Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica* 72:257–272. <https://doi.org/10.1159/000439385>
37. Mendoza E, Valencia N, Muñoz López J, Trujillo Mendoza H (1996) Differences in voice quality between men and women: use of the long-term average spectrum (ltas). *J Voice: Official Journal of the Voice Foundation* 10:59–66. [https://doi.org/10.1016/S0892-1997\(96\)80019-1](https://doi.org/10.1016/S0892-1997(96)80019-1)
38. Meunier C, Ghio A (2018) Caractériser la distinctivité du système vocalique des locuteurs. In: Actes des XXXII Journées d'Etudes sur la Parole, pp 469–477. <https://doi.org/10.21437/JEP.2018-54>
39. Mühl C, Sheil O, Jarutyte L, Bestelmeyer P (2017) The bangor voice matching test: a standardized test for the assessment of voice perception ability. *Behav Res Methods* 50:1–9. <https://doi.org/10.3758/s13428-017-0985-4>
40. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. In: *Interspeech*, pp 2616–2620
41. Naika RBhalla S, Bhateja V, Chandavale AA, Hiwale AS, Satapathy SC (eds) (2018) An overview of automatic speaker verification system. Springer, Singapore
42. Nolan F, McDougall K, Hudson T (2011) Some acoustic correlates of perceived (dis)similarity between same-accent voices. In: Proceedings of the 17th international congress of phonetic sciences, Hong Kong, 17–21 August
43. Nolan F, McDougall K, Hudson T (2013) Effects of the telephone on perceived voice similarity: implications for voice line-ups. *Int J Speech Lang Law* 0:20. <https://doi.org/10.1558/ijssl.v20i2.229>
44. O'Brien B, Tomashenko N, Chanclu A, Bonastre J-F (2021) Anonymous speaker clusters: making distinctions between anonymised speech recordings with clustering interface. In: Proceedings of Interspeech 2021, pp 3580–3584. <https://doi.org/10.21437/Interspeech.2021-1588>
45. O'Brien B, Meunier C, Ghio A (2021) Presentation matters: evaluating speaker identification tasks. In: Proceedings of Interspeech 2021, pp 4623–4627. <https://doi.org/10.21437/Interspeech.2021-1211>
46. O'Brien B, Meunier C, Ghio A, Fredouille C, Bonastre J-F, Guarino C (2022) Discriminating speakers using perceptual clustering interface, pp 97–111. <https://doi.org/10.17469/O2108AISV000005>
47. Öhman L, Eriksson A, Granhag P (2010) Mobile phone quality vs. direct quality: how the presentation format affects earwitness identification accuracy. *Eur J Psychol Appl Legal Context* 2
48. Park SJ, Yeung G, Vesselinova N, Kreiman J, Keating P, Alwan A (2018) Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles. *J Acoust Soc Am* 144:375–386. <https://doi.org/10.1121/1.5045323>
49. Perrachione TK (2017) Speaker recognition across languages. OpenBU. <https://open.bu.edu/handle/2144/23877>
50. Poddar A, Sahidullah M, Saha G (2017) Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biomet* 7:0. <https://doi.org/10.1049/iet-bmt.2017.0065>

51. Pollack I, Pickett J (1954) On the identification of speakers by voice. *J Acoust Soc Am* 26. <https://doi.org/10.1121/1.1907349>
52. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N (2011) The Kaldi speech recognition toolkit. In: Proceedings of 2011 IEEE workshop on automatic speech recognition and understanding
53. Prince S, Elder J (2007) Probabilistic linear discriminant analysis for inferences about identity. In: IEEE 11th International conference on computer vision, pp 1–8. <https://doi.org/10.1109/ICCV.2007.4409052>
54. Quené H (2001) On the just noticeable difference for tempo in speech. *J Phon* 35:353–362. <https://doi.org/10.1016/j.wocn.2006.09.001>
55. Ramos D, Franco-Pedroso J, Gonzalez-Rodriguez J (2011) Calibration and weight of the evidence by human listeners. The atvs-uam submission to nist human-aided speaker recognition 2010. In: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 5908–5911. <https://doi.org/10.1109/ICASSP.2011.5947706>
56. Rietveld ACM, Broeders APA (1991) Testing the fairness of voice identity parades: the similarity criterion. In: Proceedings of the 12th international congress of phonetic sciences, pp 46–49
57. Roebuck R, Wilding J (1993) Effects of vowel variety and sample length on identification of a speaker in a line-up. *Appl Cogn Psychol* 7:475–481. <https://doi.org/10.1002/acp.2350070603>
58. Schwartz R, Campbell JP, Shen W, Sturim DE, Campbell WM, Richardson FS, Dunn RB, Granville R (2011) Ussss-mitl 2010 human assisted speaker recognition. In: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 5904–5907. <https://doi.org/10.1109/ICASSP.2011.5947705>
59. Schweinberger S, Kawahara H, Simpson A, Skuk V, Zaske R (2014) Speaker perception. *Wiley interdisciplinary reviews. Cogn Sci* 5. <https://doi.org/10.1002/wcs.1261>
60. Singh N, Agrawal A, Khan PR (2017) Automatic speaker recognition: current approaches and progress in last six decades. *Glob J Enterp Inf Syst* 9:38–45. <https://doi.org/10.18311/gjeis/2017/15973>
61. Smith HMJ, Bird K, Roeser J, Robson J, Braber N, Wright D, Stacey PC (2020) Voice parade procedures: optimising witness performance. *Memory* 28(1):2–17. <https://doi.org/10.1080/09658211.2019.1673427>
62. Smith HMJ, Roeser J, Pautz N, Davis JP, Robson J, Wright D, Braber N, Stacey PC (2023) Evaluating earwitness identification procedures: adapting pre-parade instructions and parade procedure. *Memory* 31(1):147–161. <https://doi.org/10.1080/09658211.2022.2129065>. <https://arxiv.org/abs/https://doi.org/10.1080/09658211.2022.2129065>. PMID: 36201314
63. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5329–5333
64. Stevenage SV (2018) Drawing a distinction between familiar and unfamiliar voice processing: a review of neuropsychological, clinical and empirical findings. *Neuropsychologia* 116:162–178. Special Issue: Familiar Voice Recognition
65. Stevenage S, Symons A, Fletcher A, Coen C (2019) Sorting through the impact of familiarity when processing vocal identity: results from a voice sorting task: familiarity and voice sorting. *Q J Exp Psychol* 73. <https://doi.org/10.1177/1747021819888064>
66. Sussman J (1991) Stimulus ratio effects on speech discrimination by children and adults. *J Speech Hearing Res* 34:671–8. <https://doi.org/10.1044/jshr.3403.671>
67. Titze I (2000) Principles of voice production, second printing. National Center for Voice and Speech, Iowa City, pp 245–251
68. Tomashenko N, Srivastava BML, Wang X, Vincent E, Nautsch A, Yamagishi J, Evans N et al The VoicePrivacy 2020 challenge evaluation plan. <https://doi.org/10.48550/ARXIV.2205.07123>
69. Tomashenko N, Wang X, Vincent E, Patino J, Srivastava B, Noe P-G, Nautsch A, Evans N, Yamagishi J, O’Brien B, Chanclu A, Bonastre J-F, Todisco M, Maoche M (2022) The voiceprivacy 2020 challenge: results and findings. *Comput Speech* 74:101362. <https://doi.org/10.1016/j.csl.2022.101362>
70. Van Lancker D, Kreiman J (1987) Voice discrimination and recognition are separate abilities. *Neuropsychologia* 25(5):829–834. [https://doi.org/10.1016/0028-3932\(87\)90120-5](https://doi.org/10.1016/0028-3932(87)90120-5)
71. Venezia J, Saberi K, Chubb C, Hickock G (2012) Response bias modulates the speech motor system during syllable discrimination. *Front Psychol* 3:157. <https://doi.org/10.3389/fpsyg.2012.00157>
72. Whiteside SP (1998) Identification of a speaker’s sex: a study of vowels. *Percept Mot Skills* 86:579–584
73. Wu K-C, Childers DG (1991) Gender recognition from speech. Part i: coarse analysis. *J Acoust Soc Am* 90(4 Pt 1):1828–1840
74. Yarmey A (2003) Earwitness identification over the telephone and in field settings, vol 10. <https://doi.org/10.1558/sll.2003.10.1.62>
75. Zetterholm E, Blomberg M, Elenius D (2004) A comparison between human perception and a speaker verification system score of a voice imitation

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.